Detection of Breast Cancer Using Machine Learning and Deep Learning Methods

Sweta Bhise
Computer Engineering
SRIEIT,Goa University
Shiroda,India
swetabhise09@gmail.com

Shrutika Gadekar
ComputerEngineerig
SRIEIT,Goa
University
Shiroda,India
shrutikagadekar333@gmail.com

Aishwarya Singh Gaur
Computer Engineering
SRIEIT,Goa University
Shiroda,India
cbaishwaryasingh27@gmail. com

Simran Bepari
ComputerEngineering
SRIEIT,Goa
University
Shiroda,India
mustafabepri1976@gmail.com

Deepmala Kale
ComputerEngineeng
SRIEIT,Goa
University
Shiroda,India
deepmalakale@gmail.com

Shailendra Aswale *Computer Engineering SRIEIT,Goa University Shiroda,India*
aswale.shailendra@gmail. com

Abstract— ***The most common reason for women's deaths worldwide is Breast Cancer, which can be controlled if detected at an early stage. There are numerous techniques based on Machine Learning as well as Deep Learning can be used for the diagnoses of the cancer. This paper employs Convolutional Neural Network (CNN), which also performs feature extraction for the Machine Learning algorithms in comparison with Support Vector Machine (SVM), Naïve Bayes classifier (NB), K- Nearest Neighbor (KNN), Logistic Regression (LR) and Random Forest (RF). The system is experimented on BreaK-Hist Dataset obtained from Kaggle. The train_test_split method is used to bifurcate the raw data to distinct sets in order to achieve appropriate results. Softmax activation function predicts the outcomes based on probabilities. The main objective is to correctly classify the data and asses the performance of the models in comparison based on precision and recall rate. Further it also asses based on f1- score and as well as accuracy. When experimented the results show that CNN gives maximum accuracy with the lower most loss/error rate.***

Keywords— ***Breast Cancer, Dataset, CNN, KNN, NB,RF, SVM, LR***

I. INTRODUCTION

One of the most diagnosed disease among women worldwide is Breast cancer. It accounts for one in four cases. The survivability of the diagnosed patients on the whole depends on early detection of even the minute symptoms as early as possible so they have the best chance for successful treatment. Breast cancer is the leading cause of death from cancer in women. Around 170,000,000 cases of breast cancer were diagnosed worldwide in 2012 which accounts for 12% of total cancer cases [4].

Medical practitioners assert that the tumor is called malignant if the cells growing out of control and found abnormal which does not function as that of the normal cells. These have a higher chance of escalating to the other body parts, making it necessary to diagnose the disease early. In the past the cancer was detected widely by using the test called "Pap screening". After advancement in technology X- rays and mammograms were discovered. The issue with this disease is that till now there is no proper mechanism present to detect this before time.In such scenarios it is difficult for person to start the treatment early. It is found that breast cancer can be caused due to hormonal changes, shift in life style and environmental change [16]. It can also be associated with the age, family - history, genetics, over-weight, drinking alcohol and also lack of physical activity [22].

The absence of an efficient model for diagnosis of cancer at the primitive stage makes it difficult for medical experts to prepare a treatment plan for a patient to prolong the survival time. Hence, time stands in need of devising a proper method that gives the accurate result with least error. The methodology followed by the practitioners to detect the tumor are generally tedious and time taking which urges to come up with a unique diagnostic method. Currently mammograms are highly trusted for cancer detection, yet they have a high possibility of giving false results. This may lead to unnecessary surgeries and biopsies which may further lead to permanent damage [30].The process of classifying tumors can be done by machine learning and deep learning techniques, the latter can give better accuracy when data is complex and large compared to machine learning algorithms.

This paper presents comparison between different ML and DL algorithms to detect the type of breast cancer. The ML algorithms that are compared are SVM, KNN, RF, NB and LR along with a deep learning algorithm CNN. The comparison is done on the basis of f1-score, precision, accuracy and recall rate.

The paper organization is as follows: Section II briefs out the review on literature where multiple researches work on cancer detection techniques are considered briefly. Under Section III the recommended approach is CNN that is applied for breast cancer detection. Our CNN model is compared and analyzed in section IV. Section V contains conclusion of the paper.

II. RELATED WORK

A comparative analysis was done on distinct ML algorithms specifically; KNN, Decision Tree (C4.5) NB and SVM the using the Wisconsin Breast Cancer dataset. Also, efficacy and usability of the algorithms had been examined with respect to accuracy, specificity along with precision and sensitivity to discover the fine classification accuracy of which SVM attained the highest accuracy of 97.13%. The experiments had been performed using WEKA data mining tool and in the bounds of simulated situations [2].

In this paper a have a look at become conducted on three famous ML techniques particularly; RF, SVM, and Bayesian Networks (BN) using Wisconsin Breast Cancer set as the instruction set. The consequences acquired showed that the category of completion varies relying on the method selected. Out of the three techniques SVM outperformed well towards precision, accuracy and specificity but it changed into also seen that RF's had the pinnacle most possibility of correctly figuring out the tumor [3].

A performance assessment of numerous ML algorithms were performed including SVM, Random Forest Decision Tree and KNN by maintaining the objective of finding the satisfactory classifier. These techniques had been carried out on Wisconsin Dataset and performed in R tool. The consequences obtained exhibited that KNN attained the maximum accuracy with recognize to accuracy, specificity, sensitivity and precision accompanied by using RF, Decision Tree and SVM [4].

A study was carried out on distinct techniques of machine learning such as CNN, RF algorithm, SVM and Bayesian methods. It became visible that CNN supplied reliable outcome in comparison to other models that is because of the advancement in image under biomedical category due to the Global Feature Extraction present into it. This feature allows the CNN model to extract hidden structures from the supplied pictures [5].

The paper offers a detailed study on different ML techniques that may be used to predict Breast Cancer using RF, KNN and NB. These strategies were applied on Wisconsin Breast Cancer dataset as schooling dataset to examine the performance of Different ML algorithms with respect to the key parameters inclusive of accuracy and precision. It was found that KNN was the most effective among all the other techniques and also it was observed that each algorithm attained an accuracy above 94% [10].

The recognition of this paper become to investigate one of a kind ML strategy inclusive of SVM, ANN and Naïve Bayes using the WDBC and examine their performances to identify the maximum appropriate technique. The experimental outcomes acquired showed that SVM could acquire more potent outcome than other algorithms in terms of accuracy, specificity and sensitivity [11].

This paper performs an evaluation between many ML techniques which includes Artificial Neural Network (ANN), CNN, KNN, SVM and Inception V3. The foremost idea is to compare the results of SVM and ANN. The overall analysis depicted that the results were more accurate and efficient after applying ANN on SVM in Training and testing phase [12].

An analogy between different machine learning techniques such as ANN, Back Propagation Network, CNN and SVM using the Wisconsin Breast Cancer datasets was performed. The simulation outcome culminates that SVM is finest procedure and had acquires better outcome (94%) [14].

A comparative analysis is performed between SVM, CNN and RF. The simulation outcome concluded that CNN is the first-rate classifier because it gives genuine classification of virtual mammograms using the appropriate filtering as well as morphological operations [15].

According to these studies paper several ML algorithms were carried out to predict the Breast Cancer namely; SVM, RF, NB and Logistic Regression. The primary objective was to determine which algorithm gave the highest accuracy and efficiency with less time consumption. The results obtained proved that RF obtained the highest accuracy with least error rate [17].

A narrative method to detect breast cancer with the aid of taking on strategies of Machine Learning inclusive of NB classifier, SVM classifier, Recursive Convolutional Neural Network (RCNN) classifier, Bi-clustering Ada Boost methods and Bidirectional Recurrent Neural Networks (HA-BiRNN) were presented. It was concluded that the Deep Neural Network (DNN) algorithm was advantageous in terms of performance and efficiency [21].

The dataset used were from Dr. William H. Walberg of the University of Wisconsin Hospital. The different Data visualization and ML techniques such as SVM, LR, KNN, Decision tree, RF, Rotation Forest and NB have been applied to this dataset. An analogy was performed amongst all of the different methods. Outcome received with the LR model performed better by giving maximum accuracy (98.1%) under classification. This outcome has encouraged enhancement in accuracy performances [22].

The overall performance assessment among three algorithms specifically LR, KNN and Ensemble Learning using Wisconsin breast cancer diagnosis (WDBC) information set retrieved from UCI machine learning repository was performed. Ensemble Learning approach comprising of 5 machine learning algorithms - LR, KNN, RF classifier, Linear Discriminant Analysis (LDA) and SVM classifier gave the accuracy of 99.30% [23].

A novel method to predict breast cancer by means of using strategies of Machine Learning that is LR, RF, KNN, Decision tree, SVM, NB classifier technique and strategies of Deep Learning that is Recurrent Neural Network (RNN), ANN and CNN were presented. The comparative evaluation among the Machine Learning strategies and Deep Learning techniques concluded that the accuracy acquired using CNN model (97.3%) and that of ANN model (99.3%) became the most efficient than the Machine Learning models [29].

An analogy of the machine learning algorithms namely, SVM, Gaussian Mixture Modeling (GMM), KNN, RSDA, LRC, ANN, LR and BN was conducted. These different algorithms were either used alone or in combination with other algorithms. The machine learning algorithm that achieved maximum accuracy after this study was SVM [30].

III. PROPOSED METHODLOGY

A. Dataset

A well-arranged dataset is an indispensable requirement to produce a functional and sturdy method

for the identification of breast cancer. Dataset in use for this paper was acquired from Kaggle.

Data set was acquired from a trusted medical source with little to no bias. It was assumed that the data set was normalized. Break Hist_Dataset containing four directories storing the amplified images is being used. All the directories contain a sum of 4,906 instances. We divide the dataset in the ratio of 4:1 with respect to training data and testing data respectively.

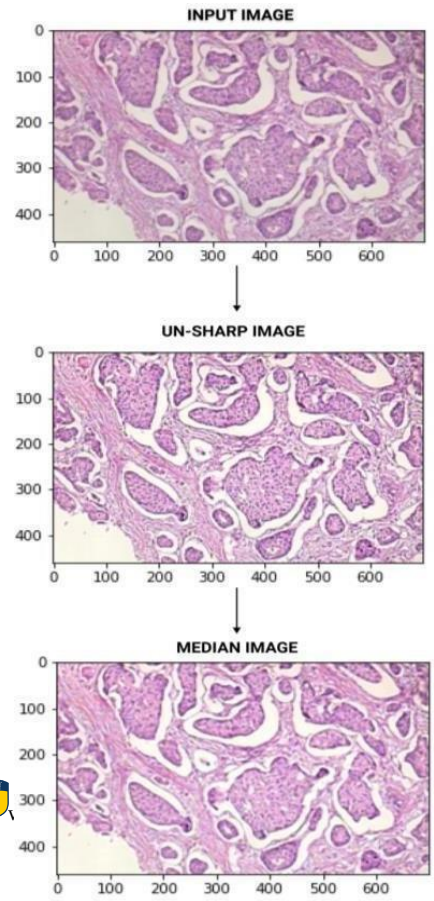
TABLE I. DESCRIPTIONOF BREAKHIST DATASET

Classes	Training	Testing	Total
Benign	2000	415	2415
Malignant	2015	476	2491
Total	4015	891	4906

B. Image Preprocessing Normalization

Usually, the dataset contains huge number of instances and each instance may vary from the other in shape, size, contrast or even magnification. Hence this data needs to be preprocessed before being given as an input to any model. In our experiment two filters were used: Un-sharp and Median filter. Un-sharp filter enhances the edges of an image and the Median filter is used for noise reduction. Fig 1. shows the process of image filtering. After the process of filtering, the data is normalized. The purpose of normalization is simply to transform or scale data between 0 and 1 so that it is either dimensionless or/and have similar distributions.

Fig. 1. Image Filtering Process



C. Feature Extraction

The input data supported by the machine learning algorithms is either in one dimensional or two-dimensional format. In order to obtain this, we have used CNN model for feature extraction, as the Flatten layer available in the model reduces the dimension of the data after extracting the features, thereby keeping the procedure free of complexities. The features are then fed in to the machine learning models/algorithms i.e., KNN, SVM, NB, RF and LR.

D. Methodology for Machine Learning Algorithms

To implement the ML algorithms an inbuilt machine learning library sklearn has been used. Sklearn provides different useful tools for machine learning and statistical modeling which includes classification, regression and clustering. The main motive of this project is to classify breast cells either into benign or malignant, hence sklearn is most suitable.

The various inbuilt functions that were imported from sklearn library were SVC for the execution of SVM classifier, KNeighborsClassifier function used for the execution of KNN, RandomForestClassifier function used for the execution of Random Forest algorithm, LogisticRegression function used for the execution of LR algorithm and GaussianNB function used for the execution of Naïve Bayes classifier.

E. Methodology of Proposed System Model

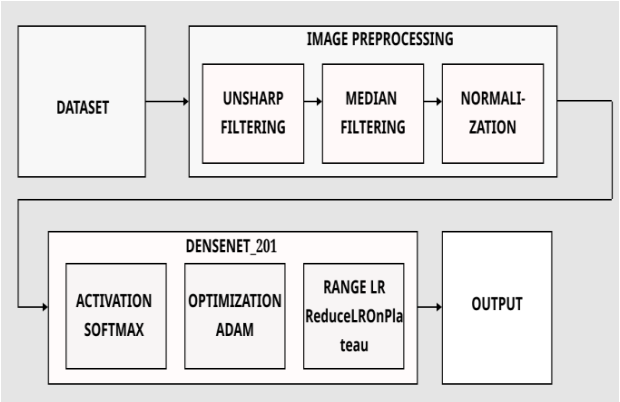
CNN will ease the process of differentiating the malignant and benign tumor rapidly. This specialized neural network extract all the necessary features automatically. It is better than any other feed forward neural network because of its nature of sharing parameters due to which the number of parameters is potentially reduced which in turn eases the overall process of computation without losing on the quality of model.

For the training of the CNN model the framework that has been used is, DenseNet201. DenseNet is capable of training deeper CNN models and hence resulting in greater accuracy.

The activation function that has been used is ‘softmax’. Softmax is used to scale numbers into probabilities, the output is a vector with probabilities with each possible outcome. The optimizer that has been used is ‘Adam’. Adam works efficiently when it comes to large datasets or if a dataset has large number of parameters (‘18,333,506’ parameters in current scenario). To set a range for the learning rate, we have used ReduceLROnPlateau function. This function helps whenever a specific metric stop showing better results.

Models often avail from minimizing the learning rate by a point of 2–10 once learning gets saturated. This monitors a quantity and if there is no improvement seen for a 'patience' number of epochs, the learning rate is decreased. In our model we have set the LR range between 1e-4 and 1e-7. We have also used Model Checkpoint function to restore a copy of the efficiently performing model which enables to retrain the model for higher number of epochs.

Fig. 2. Workflow of CNN



IV. EXPERIMENTATION AND ANALYSIS

In this section we intend to estimate the performance of various machine learning algorithms in comparison and the proposed system model.

A. Performance of Machine Learning Algorithms

The hyper-parameter matching is performed by SVM using kernels. The performance of the kernels may vary from one dataset to the other. In order to identify which kernel suits the data set used in this project, a comparison was conducted between the four SVM kernels- sigmoid, polynomial, linear and rbf, which were then assessed on the basis of precision, recall rate, f1- score and f1-score. Experimental results show that rbf kernel outperformed the kernels in comparison by attaining the maximum accuracy (75%), precision (77%), recall-rate (76%) and f1-score (76%). Fig 3. displays the experimental results of SVM kernels.

The LR classifier managed to achieve 75% accuracy, 77% precision,76% recall rate and 76% f1-score. RF produced approximately similar results to LR but suffered a slight loss in all aspects with a margin of one-two percent. Fig 4. displays the experimental results of RF and LR.

The NB classifier attained an accuracy of 76%, precision of 78%, recall rate of 77% and f1- score of 77%, thereby outperforming all the models in comparison while the KNN classifier performed critically in comparison to other classifiers in all aspects. The performance obtained for the metrics accuracy was 66%, precision was 68%, recall rate was 64% and f1-score was 64%.

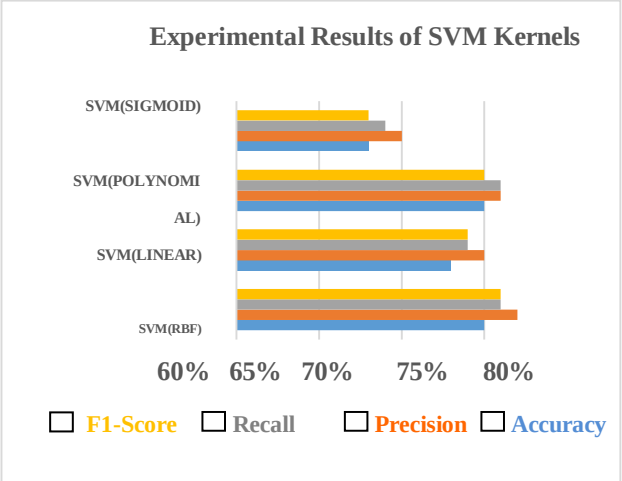


Fig. 3. Test Results of SVM Kernels

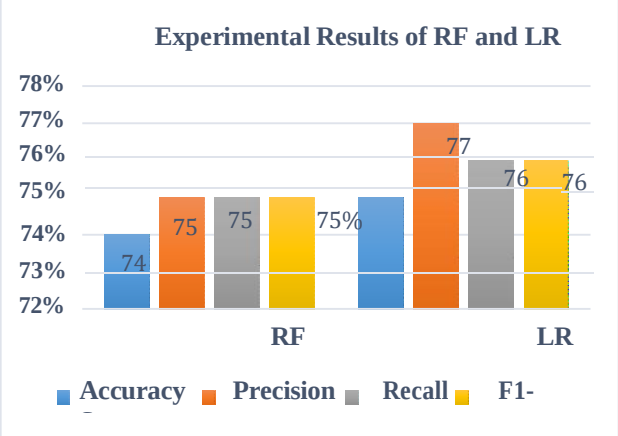


Fig. 4. Test Results of RF and LR

B. Performance of Proposed System Model

To obtain better performance of CNN, we have emphasized on F1-score along with accuracy. The reason to not wholly rely on accuracy being, accuracy is the measure of four variables – True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). Even if TP and FP variable is zero there is a possibility of gaining a higher accuracy, on the contrary F1-score represents the model score as a function of precision and recall.

The model was trained for 100 epochs and it was observed that the accuracy was saturated after the 60th epoch. Examining all the performance metrics throughout 100 epochs it was concluded that the best results with respect to precision, recall rate, f1-score, accuracy and confusion matrix were obtained on the 60th epoch. Fig. 5 ,Fig. 6 and Fig. 7 shows the loss and accuracy for 50th, 60th and 100th epochs respectively.

The training and the validation accuracy obtained at 60th epoch as 97% and 99% respectively. The training and the validation loss attained were 0.07 and 0.03 respectively. The following graphs represent the proposed system model-CNN with respect to the accuracy and the loss analyzed for the through hundred epochs. Table II depicts the comparison between all machine learning and deep learning models.

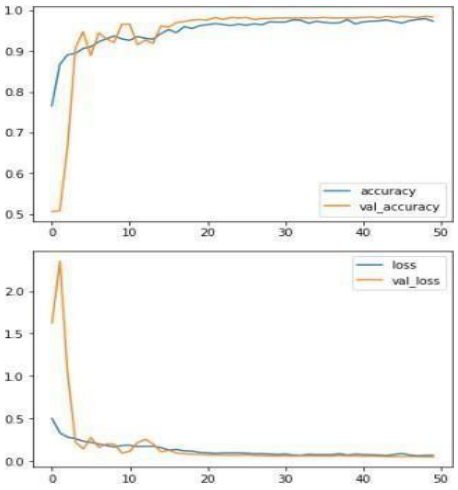


Fig. 5. Accuracy and Loss at 50th Epochs

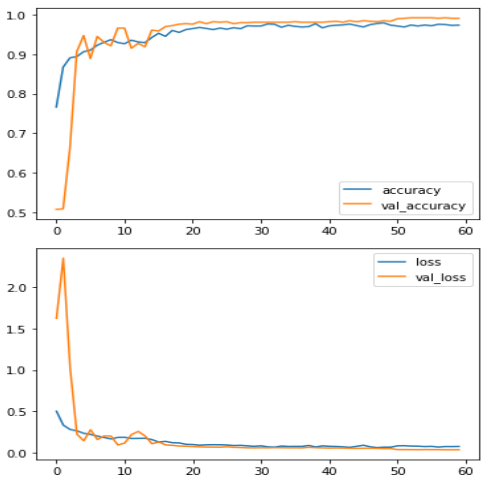


Fig. 6. Accuracy and Loss at 60th Epochs

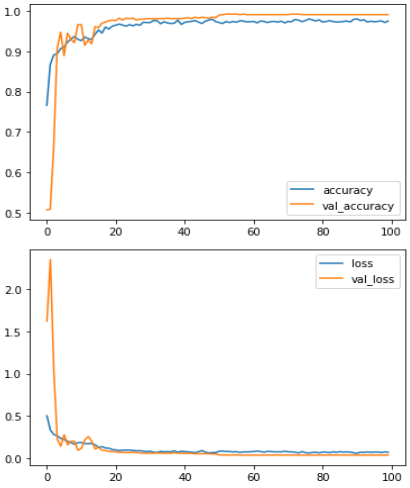


Fig. 7. Accuracy and Loss at 100th Epochs

TABLE II. COMPARISION OF ACCURACY MEASURESFOR CNN, NB, SVM, LR, KNN AND RF

Model Name	Class	Precision (%)	Recall (%)	F1-Score (%)
CNN	Benign	87	97	92
	Malignant	97	88	92
NB	Benign	69	89	78
	Malignant	87	65	75
SVM	Benign	68	87	77
	Malignant	86	65	74
LR	Benign	68	88	77
	Malignant	86	64	74
KNN	Benign	72	43	54
	Malignant	63	85	73
RF	Benign	69	82	75
	Malignant	81	67	74

I. CONCLUSION

This paper analyzed BreakHist dataset with feature extraction using CNN and distinct deep learning and machine learning techniques for breast cancer detection. The comparison was done by means of precision, recall rate, f1-score and accuracy. It was observed that amongst the machine learning algorithms the NB classifier attained an accuracy of 76%, precision of 78%, recall rate of 77% and f1-score of 77%, thereby outperforming all the models in comparison. It was also concluded that deep learning models perform more efficiently than machine learning algorithms. From the results obtained it was seen that CNN performed exceptionally well giving precision of 92%, recall rate of 93% and f1-score of 99% and an accuracy of 99%.

I. REFERENCES

[1] M. Rana, P. Chandorkar, A. Dsouza and N. Kazi " Breast cancer diagnosis and recurrence prediction using machine learning techniques" in IJRET, vol 4, April 2015. [Online].

[2] H. Asri, H. Mousannif, H. A. Moatassime, and T. Noel " Using machine learning algorithms for breast cancer risk prediction and diagnosis," in Elsevier B.V. 2016, pp. 1066-1069. [Online]. doi: 10.1016/j.procs.2016.04.224

[3] D. Bazazeh and R. Shubair "Comparative study of machine learning algorithms for breast cancer detection and diagnosis," in IEEE2016,p.4. [Online].

[4] A. Joshi and A. Mehta " Comparative analysis of various machine learning techniques for diagnosis of breast cancer," in International Journal on Emerging Technologies, India 2017, p. 522. [Online].

[5] A. A. Nahid and Y. Kong," Involvement of machine learning for breast cancer image classification: a survey," in Hindawi Computational and Mathematical Methods in Medicine, Article ID 3781951, 31 December 2017. [Online]. doi:https://doi.org/10.1155/2017/3781951

[6] A. A. Nahid, M. A. Mehrabi, and Y. Kong " Histopathological breast bancer image classification by deep neural network techniques guidedby local clustering" in Hindawi Journal of Healthcare Engineering 2018,p.

[7] 19. [Online]. Available: https://doi.org/10.1155/2018/2362108

[8] A. A. Nahid, A. Mikaelian, and Y. Kong " Histopathological breast- image classification with restricted Boltzmann machine along with backpropagation" in Biomedical Research 2018, p. 2076. [Online].

[9] S. J. Mambou, P. Maresova, O. Krejcar, A. Selamat and K. Kuca," Breast cancer detection using infrared thermal imaging and a deep learning model," in Sensors ,2018, p. 17. [Online]. doi:10.3390/s18092799

[10] H. Le "Using machine learning models for breast cancer detection," 2018.

[11] S. Sharma, A. Aggarwal and T. Choudhury " Breast cancer detection using machine learning algorithms," in IEEE, 2018, pp. 114-117.

[12] D. A. Omodiagbe, S. Veeramani and A. S. Sidhu" Machine learning classification techniquesfor breast cancer diagnosis," in IOPConf. Series: Materials Science and Engineering , vol 495, 2019. [Online]. doi:10.1088/1757-899X/495/1/012033

[13] K. Wadkar, P. Pathak and N. Wagh " Breast cancer detection using ANN network and performance analysis with SVM," in IJCET,2019,pp. 75-85. [Online]. Available: https://ssrn.com/abstract=3555041

[14] K. Sekaran, S. P. Ramalingam and C. Mouli P.V.S.S.R, " Breast cancer classification using deep neural networks," in Research Gate, February 2018. [Online]. doi: https://doi.org/10.1007/978-981-10-6680- 1_12

[15] S.Vasundhara , B.V. Kiranmayee and C. Suresh,"Machine learning approach for breast cancer prediction," in IJRTE, vol 8, 2019,pp.2619- 2625.[Online].

[16] S. J. S. Gardezi, A. Elazab, B. Lei and T. Wang,"Breast cancer detection anddiagnosisusingmammographicdata: systematicreview,"in Journal of Medical Internet Research,2019. [Online]. doi: 10.2196/14464

[17] K.Anastraj, Dr.T.Chakravarthy,and K.Sriram, " Breast cancer detection either benign or malignant tumor using deep convulionalneural network with machine learning techniques," in Adalya Journal, vol 8, 2019, pp. 77-83.[Online].

[18] S. J, A. K. V, S. S. S, and S. S " Breast cancer prediction using machine learning," in IJRTE, vol 8, 2019, pp. 4879-4881. [Online]. doi:10.35940/ijrte.D8292.118419

[19] S. H. Nallamala, P. Mishra and S. V. Koneru, " Breast cancer detection usingmachinelearningapproaches," in IJRTE, vol7, Issue-5S4, February 2019. [Online].

[20] S. Z. Ramandan " Methods used in computer-aided diagnosis for breast cancer detection using mammograms: a review," in Hindawi Journal of Healthcare Engineering 2020, p. 21. [Online]. Available: https://doi.org/10.1155/2020/9162464

[21] R. Rawal " Breast cancer prediction using machine learning" in

[22] JETIR, vol 7, May 2020.

[23] A. Reddy V., B. Soni and S. Reddy K.," Breast cancer detection by leveraging machine learning," in ICT Express,22 April 2020. [Online]. doi: https://doi.org/10.1016/j.icte.2020.04.009

[24] M.F.Ak, " A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications,"in Healthcare 2020, pp. 1-23.[Online]. doi: 10.3390/healthcare8020111

[25] R. M. Rawat, S. Panchal, V. K. Singh and Y. Panchal " Breast cancer detection using K-nearest neighbors, logistic regression and ensemble learning," in ICESC 2020, pp. 534-540.

[26] N. Rane, J. Sunny, R. Kanade and Prof. S. Devi,"Breast cancer classification and prediction using machine learning," in IJERT, vol. 9, February 2020. [Online].

[27] S. A. Mohammed, S. Darrab, S. A. Noaman, G. Saake " Analysisof breast cancer detection using different machine learning techniques" in DMBD, 2020. Available: https://doi.org/10.1007/978-981-15-7205-0_10

[28] S. A. Alanazi, M. M. Kamruzzaman, Md N. I. Sarker, M. Alruwaili,

[29] Y. Alhwaiti, N. Alshammari and M. H. Siddiqi, " Boosting breast cancer detection using convolutional neural network," in Hindawi Journal of Healthcare Engineering, Article ID 5528622, 5 April 2021.[Online].doi: https://doi.org/10.1155/2021/5528622

[30] Priyankaand K. Sanjeev, "Areviewpaper on breast cancerdetection usingdeep learning," in IOP Conf. Ser.: Mater. Sci. Eng., 2021.[Online]. doi: 10.1088/1757-899X/1022/1/012071

[31] H. Masood, " Breast Cancer Detection Using Machine Learning Algorithm," in IRJET, vol 8 Issue: 02, February 2021. [Online].

[32] M. Tiwari, R. Bharuka, P. Shah and R. Lokare, " Breast cancer

[33] prediction using deep learning and machine learning techniques".

[34] M. Tahmooresi, A. Afshar, B. B. Rad, K. B. Nowshath and M. A. Bamiah," Early detection of breast cancer using machine learning techniques," in Journal of Telecommunication, ElectronicandComputer Engineering, vol 10, pp. 21-27.