Date of publication JUN-30, 2023, date of current version MAR-15, 2023.
www.computingonline.net / computing@computingonline.net

Print ISSN 1727-6209
Online ISSN 2312-5381
DOI 10.47839/ijc.22.2.3093

Early Detection of Breast Cancer Using Machine Learning and Ensemble Techniques

DISHA H. PAREKH[1], VISHAL DAHIYA[2]
[1]Research Scholar and Assistant Professor, Department of Computer Science, IIICT, Indus University, Ahmedabad, 382115, India
[2] Supervisor and Professor Head, Department of Computer Science, IIICT, Indus University, Ahmedabad, 382115, India

Corresponding author: Disha H. Parekh (e-mail: disha.hparekh213@gmail.com, dishaparekh.19.rs@indusuni.ac.in).

ABSTRACT Breast Cancer is found as the most dangerous and most commonly affecting diseases in the world by WHO. The severity of breast cancer and early diagnosis of it has gained the attention of researchers to save humankind from such devastating disease. Early prediction of breast cancer has geared up its journey after the introduction to machine learning supervised algorithms. In the paper, the use of various machine learning algorithms along with the ensemble algorithms is shown. The results obtained are highly accurate to help one correctly predict cancer. The paper aims at early diagnosis of breast cancer with a humble motto of saving patients suffering from the disease by allowing them to know whether the diagnosed tumor is cancerous or non-cancerous, being Malignant and Benign respectively. This paper would be useful and aiding for those who are novel researchers in prediction and diagnosis of breast cancer using machine learning.

KEYWORDS Breast cancer prediction; Ensemble Machine Learning algorithms; AdaBoost; XGBoost; F1-Score.

I. INTRODUCTION

ACCORDING to the statistics of World Health Organization during 2020, breast cancer has been the most predominant disease of the world. It stated that during 2020, 2.3 million of women across world had been identified with breast cancer and by the end of the month almost 7.8 million of women had been surviving in the world with the record of past five years (Breast Cancer, 2021). Breast cancer has been an invasive disease since 1930 and is right now an area of attraction for researchers to infringe this invasive disease and bring awareness amongst the population with early detection and diagnosis of the disease. Breast cancer disease can be treated effectively when detected in its early stages. This early detection is an area where many researchers are working today. There are several researchers working on medicine development and its discovery for eliminating the meddlesome disease. Thus, cancer biology is found to be gearing up the interest of researchers across the world.

Breast cancer is not a contiguous or transferrable disease. It is a disease spreading widely due to mutations in cells. It is not a viral or bacterial infectious disease but is mutant to changes in gene material, particularly protein sequencing. Though breast cancer is mostly observed in females, some 1% of males are also victims to this disease. The disease is caused by a lump in the breast. This lump is painless but it is abnormal and hence should be treated urgently by consulting surgeons. There are basic two genes called Breast Cancer Gene 1 and Gene 2, usually referred as BRCA1 and BRCA2 which produce proteins to remove ruptured DNA. These genes help in suppressing tumors in the body. But any pathogenic disorders or any mutations in the gene sequence of any of these genes, leads to breast cancer. About 13% of females in general develop breast cancer during their lifetimes (N et al., 2020). If compare, 55%–72% of females who receive a harmful BRCA1 variant and 45%–69% of females who get a harmful BRCA2 variant will progress breast cancer by 70–80 years of age (Kuchenbaecker KB).

Breast cancer is represented in two different ways. The lumps in breast cancer can be either cancerous or non-cancerous. All those lumps which are non-cancerous are usually called as Benign type which means there exists no cancer. While all those lumps which are cancerous in nature are termed as Malignant tumors. These malignant tumors need diagnosis using biopsy of the lump mass or can be diagnosed using breast imaging.

Breast cancer can be treated with 90% of survival ratio when diagnosed in early stages (Caplan L.). The treatment consists of oral chemotherapy, intravenous chemotherapy,

radiation or surgery in usual cases to control the disease in the breast, its surrounding areas and the lymph nodes. There are anti-cancer medicines available to control cancerous cells which includes endocrine therapy which is a hormonal control therapy, chemotherapy and sometimes even antibody therapies which is a targeted biological therapy.

The aim of the WHO Global Breast Cancer Initiative (GBCI) is to decrease worldwide breast cancer transience by 2.5% each year, thereby preventing 2.5 million breast cancer deaths worldwide between 2020 and 2040. Dropping universal breast cancer death ratio by 2.5% each year would prevent 25% of breast cancer deceases by 2030 and 40% by 2040 amongst females under 70 years of age. The three supports towards attaining these aims basically intend to be 1) fitness elevation for early detection; 2) well-timed diagnosis; and 3) complete breast cancer management (DeSantis, et al.).

There is an organization called National Breast Cancer Coalition (NBCC) which works dedicatedly towards the end of breast cancer through action and advocacy. Accor ding to their study carried out in 2022, breast cancer is found to be the most common disease where there are estimated to be 2,87,850 new cases of invasive breast cancer in women and 2710 new cases in men. They have even shown that there will be additional 51,400 cases of ductal carcinoma in situ diagnosis in women. The claim made by the NBCC is depicted in the below figure which even justifies the rise in mortality rate as age increases.

Figure 1. An image showing the statistics of breast cancer from 1975 to 2022 (Breast Cancer Statistics | Facts & Figures | NBCC, n.d.)

Breast cancer diagnosis has been a major concern for decades and thus, there are several research communities working on this area for finding solutions to the cancer, its treatments, or maybe even drug discovery. The diagnosis and treatment of breast cancer has been an area of interest for researchers of the computer science community today. Those interested in biomedical research or life sciences bioengineering are focusing today on healthcare industry and its solvability implementing computer fundamentals like Artificial Intelligence (AI), Machine Learning, Blockchain Technology, and Deep Learning.

Machine Learning is a recent technology that is used to train machines with various algorithms in order to improve automatically through learning. The diagnosis of tumors in benign or malignant conditions can prevent a human from unnecessary treatments if found benign. The bifurcation of breast cancer into benign or malignant, just in order to avoid unnecessary surgery and treatment if not cancerous, is the subject of many investigations today. Due to its uniqueness in feature categorization, breast cancer diagnosis of complex datasets uses Machine Learning (ML) as one of the most prevailing methods.

Today, due to the advent of computers and technology, it is very easy to store huge amounts of datasets and train computers using ML and its implied algorithms to process the data and assess the result of the analyzed work. When we implement ML concepts in healthcare industries for easy diagnosis of such invasive diseases, we usually refer to them as intelligent healthcare systems. Such an intelligent healthcare system assists physicians in diagnosing patients with higher accuracy and also aids the humans to understand their physical conditions and uproot them towards planning of physical conditions in future. ML techniques can take several complex tasks from physicians like understanding the mammographic images, spotting of tumors, or diagnosing the size of tumors while even concluding on the medical treatment necessary to operate on cancerous cells. The advancement in the world of computers, especially the next gen technologies like ML and AI have turned out to be a boon for such complex analysis.

There are ML algorithms classified into supervised and unsupervised learning algorithms and their enhancements as ensemble learning algorithms. Researchers opt for various sets of algorithms and even sometimes use more than one learning algorithm to help in analysis of datasets. In the paper (Parekh & Dahiya, 2021), use of Machine Learning supervised algorithms on Wisconsin Diagnostic Breast Cancer Dataset (WDBC) was carried out showing the accuracy and F1 score for several algorithms. The code was done in the R environment using RStudio Framework. The WDBC dataset contains around 569 observations only which turned out to be very few for training a machine learning model. Secondly, the diagnosis was carried out using Statistical Programming Language R which was found to be visually a bit less in exploration of data analysis as compared to Python.

Moving an extra mile ahead, this paper focuses on Python technology and Supervised Machine Learning Algorithms. This paper also shows the result and analysis between different classifiers. Secondly, the code of implementation done is shown in the appendix for reference to the research community. Here, as an extra step to showcase Neural Networks, a dense layer neural network is also built and the result is analyzed using ensemble algorithm integration with NNET.

II. LITERATURE REVIEW

In paper (Hiba et al., 2016), the use of different supervised machine learning algorithms was shown. They mentioned accuracy, sensitivity and specificity on the Wisconsin Breast Cancer Dataset with 580 observations. According to the experiment carried out, the best accuracy obtained was through Support Vector Machine which was found to be 97.13%. Although the paper did not show the experiments with Random Forest which is one of the known supervised learning algorithms.

Researchers also took the Wisconsin Breast Cancer Dataset and applied three major Machine Learning approaches, namely Support Vector Machine, K-Nearest Neighbor and Decision Tree in paper (Obaid et al., 2018). They even showed the accuracy levels between each of the stated algorithms and evaluated Receiver Operating Characteristic Curve (ROC curve). According to their experimentation the best classifier found was quadratic SVM. The paper did not show any analysis of the dataset using Random Forest or any ensemble algorithms.

In paper (Jiande & Chindo, n.d.,et.al.), researchers showed the classification of two types of breast cancer using four different algorithms, mainly, Support Vector Machines, K-

nearest neighbor, Naïve Bayes and Decision tree using features selected at different threshold levels to train the models. The RNA-Seq dataset hosted on Genomics Data Commons was used. The best algorithm concluded was the Support Vector Machine.

Researchers showed the feature selection and feature extraction methods performed on the data to reduce the dimension of features in paper (David A. et al., 2019), thereby producing reduced versions of the original dataset. The methods considered were Support Vector Machine, Artificial Neural Network, and Naive Bayes Classifier, which were employed to train the Wisconsin Diagnosis Breast Cancer dataset.

In paper (Ovelade et al.), a very unique approach to oncology diagnostic with select and test oncology system was presented in order to diagnose the breast cancer occurrence. The model first reads and then filters the inputted dataset. The dataset used here is Wisconsin Breast Cancer Dataset and the sensitivity achieved in the model is o.81 and the specificity of 0.89 is noticed. The model does focus on the use of various supervised algorithms.

Several approaches for early detection of time series classification and associated diseases have been a focused study in recent times. In paper (Wu W. and Zhou H.), early prediction and classification of cervical cancer which is a kind of a gynecological disease was proposed and shown using the machine learning algorithm, Support Vector Machine. They determined four target variables and the experimental study attempted to reduce the processing time by selecting only few relevant features using Principal Component Analysis and Recursive Feature Selection.

In paper (Dundar et al.), an automatic and efficient grouping system for classifying breast microscopic tissues into different actionable sub categories, namely atypical ductal hyperplasia and ductal carcinoma in situ was identified. In the paper, they focused on the statistical features, like the perimeter, mean gray-level intensity and the ratio of major to minor axis of the best-fitting ellipse. The selected features were exploited to model the cell shape, cell size and the nucleoli appearance which led to an accurate result analysis. Also Niwas et al. in their paper proposed a method where they extracted first-order statistical and second-order statistical features respectively to gain a precise outcome for detecting the cancerous cells.

After the literature review, a novel approach to the proportional analysis of several such ML algorithms is targeted in this paper. This paper majorly addresses the fallback in the accuracy of diagnosis carried out by the supervised algorithms over the ensemble techniques of ML.

III. MATERIALS AND METHODS

To carry out the prediction of breast cancer occurrence in terms of malignant “M” and benign “B”, a mammography dataset of breast masses known as CBIS which is Curated Breast Imaging Subset of Digital Database for Screening Mammography, usually known as CBIS - DDSM, was used. DDSM is a database containing 2620 scanned film mammography studies. The images here are decompressed and converted in DICOM format, which was then used to get access to the .csv file. Here the dataset used consists of records of “B” and “M” kinds of breast cancers for exactly 1319 patients. The study involves the use of Google Colab for computational code. It also involves the use of MATLAB for conversion of images dataset to .csv file.

Then the analysis of breast cancer with various machine learning supervised algorithms is carried out. The results obtained will be discussed in chapter 4 of the paper. Though the algorithms show better accuracy, the use of ensemble algorithms is recommended. Hence, two ensemble mechanisms, AdaBoost and Xtreme Gradient (XG) Boost are used. The model constructed is depicted in the figure below:

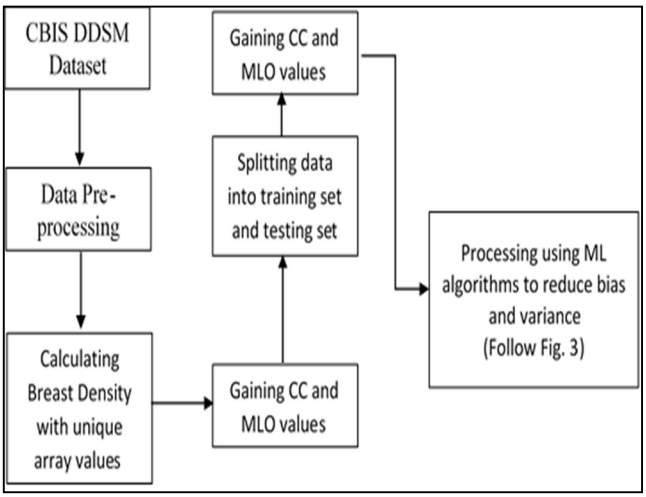

Figure 2. Processing of CBIS-DDSM dataset

A. Support Vector Classifier

It is one of the most popular and high performing supervised learning algorithms. It usually gives the most accurate results for datasets. SVM would choose the extreme points which helps in creating the hyperplane. These extreme cases are called support vectors, and hence the algorithm is termed as Support Vector Machine. It is used both, either as a classification technique or as a regression technique, though the use of the algorithm as the former is more notable.

```
# Support vector classifier
from sklearn.svm import SVC
svc_classifier = SVC()
svc_classifier.fit(X_train, y_train)
y_pred_svc = svc_classifier.predict(X_test)
#y_pred_svc_prob = svc_classifier.predict_proba(X_test)
svm = accuracy_score(y_test, y_pred_svc)
```

Figure 3. SVC Code Snippet

B. Logistic Regression

Logistic regression predicts the output of a categorical dependent variable using a set of independent variables. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, True or False and thus usually in boolean logic values. Here the major point is that instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1 (Rymarczyk et.al.).

```
# Logistic Regression
from sklearn.linear_model import LogisticRegression
lr_classifier = LogisticRegression(random_state = 51, penalty = 'l2')
lr_classifier.fit(X_train, y_train)
y_pred_lr = lr_classifier.predict(X_test)
lr = accuracy_score(y_test, y_pred_lr)
```

Figure 4. LR Code Snippet

C. K-Nearest Neighbor Classification

It is one of the simplest classification supervised algorithms. It assumes the similarity between the new case and available cases and puts the new case into the category that is most similar to the available categories. It can be sometimes even used as a regression technique (Alarabeyyat A. & Alhanahnah).

```
from sklearn.neighbors import KNeighborsClassifier
knn_classifier = KNeighborsClassifier(n_neighbors = 5)
knn_classifier.fit(X_train, y_train)
y_pred_knn = knn_classifier.predict(X_test)
knn = accuracy_score(y_test, y_pred_knn)
```

Figure 5. KNN Code Snippet

D. Naive Bayes Classifier

It is used as a classification supervised algorithm and generally never preferred for regression analysis. It constructs fast machine learning models which give very quick predictions, usually being probabilistic in nature, which means it predicts on the basis of the probability of an object (Berrar).

```
# Naive Bayes Classifier
from sklearn.naive_bayes import GaussianNB
nb_classifier = GaussianNB()
nb_classifier.fit(X_train, y_train)
y_pred_nb = nb_classifier.predict(X_test)
nbc = accuracy_score(y_test, y_pred_nb)
```

Figure 6. NB Code Snippet

E. Decision Tree Classifier

Decision Tree is a classification learning algorithm which is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. It can be sometimes used as a regression model as well (Blanco-Justicia et.al.).

```
# Decision Tree Classifier
from sklearn.tree import DecisionTreeClassifier
dt_classifier = DecisionTreeClassifier(criterion = 'entropy', random_state = 51)
dt_classifier.fit(X_train, y_train)
y_pred_dt = dt_classifier.predict(X_test)
dt = accuracy_score(y_test, y_pred_dt)
```

Figure 7. DT Code Snippet

F. Ensemble Learning: Random Forest Classifier

It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. It is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset (Prastyo et.al.).

```
# Random Forest Classifier
from sklearn.ensemble import RandomForestClassifier
rf_classifier = RandomForestClassifier(n_estimators = 100, criterion = 'gini')
rf_classifier.fit(X_train, y_train)
y_pred_rf = rf_classifier.predict(X_test)
randf = accuracy_score(y_test, y_pred_rf)
print(randf)
```

Figure 8. RF Code Snippet

G. Ensemble Algorithm: AdaBoost Classifier

Boosting is an ensemble method of converting weak learners into strong learners. Weak and strong refer to a measure of how correlated are the learners to the actual target variable. In boosting, each training sample is used to train one unit of the decision tree and picked with replacement over-weighted data. The trees will learn from predecessors and update the residuals error. Adaptive boosting was formulated by Yoav Freund and Robet Schapire. AdaBoost is the first practical boosting algorithm, and remains one of the most widely used and studied, with applications in numerous fields (Prastyo et.al.). AdaBoost algorithm works on changing the sample distribution by modifying weight data points for each iteration.

```
from sklearn.ensemble import AdaBoostClassifier
adb_classifier = AdaBoostClassifier(DecisionTreeClassifier(criterion = 'entropy', random_stat
                                    n_estimators=100,
                                    learning_rate=0.1,
                                    algorithm='SAMME.R',
                                    random_state=0,)
adb_classifier.fit(X_train, y_train)
y_pred_adb = adb_classifier.predict(X_test)
```

Figure 9. AdaBoost Code Snippet

H. Ensemble Algorithm: XGBoost Classifier

Extreme Gradient Boosting, commonly known as XGB, as stated in paper (Chen et. al.), is a new algorithm that utilizes the gradient tree boosting concept. XGBoost is a specific implementation of the Gradient Boosting Model which uses more accurate approximations to find the best tree model (Prastyo et.al.). XGBoost has specifically used a more regularized model formalization to control overfitting, which gives it better performance.

```
# XGBoost Classifier
from xgboost import XGBClassifier
xgb_classifier = XGBClassifier()
xgb_classifier.fit(X_train, y_train)
y_pred_xgb = xgb_classifier.predict(X_test)
xgb = accuracy_score(y_test, y_pred_xgb)
print(xgb)
```

Figure 10. XGBoost Code Snippet

The data preprocessing is a very important step which is then followed by implementing different ML algorithms and then a desired model generates an accuracy output which will be further analyzed.

Figure 11 represents the data pre-processing steps and the outcomes targeted after using different ML algorithms.

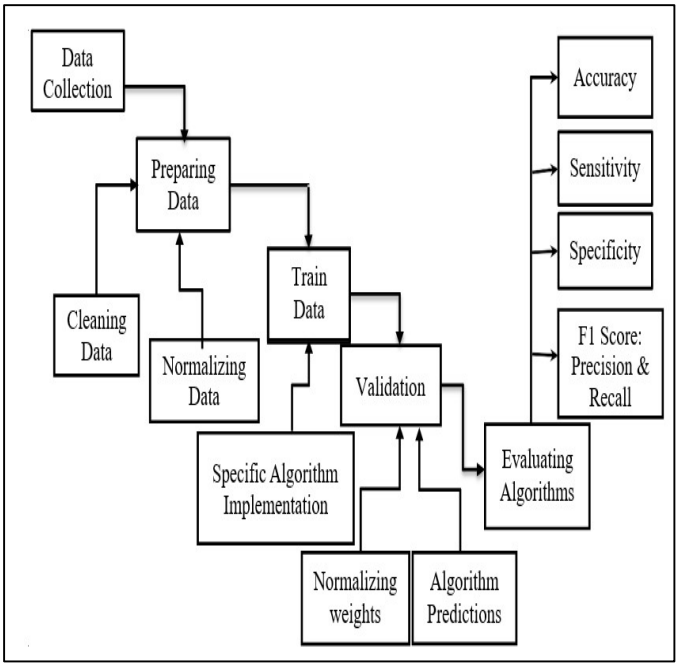

Figure 11. Data Preprocessing Stages in the Model

IV. EXPERIMENTAL RESULTS

Ensemble learning is a Machine Learning paradigm where more than one model, colloquially known as "weak learners", are trained to solve a unique problem and combining it further to obtain better results. The primary thought behind this is that, when weak learners are combined with each other correctly, we can obtain more accurate and robust models as discussed in paper of Schwenker F. et. al. Ensemble methods usually combine stacking, blending, and lastly boosting and bagging, as notified in paper of Bellmann et. al. These techniques and combination had been applied successfully in various regression and classification tasks as stated in paper (Kachele et. al.), as well as in clustering, as shown in paper (Boongoen et. al.) and reinforcement learning, stated in paper (Fauber et. al.).

The experimental setup has used several ensemble methods namely random forest, AdaBoost and XGBoost with a comparative study of other supervised machine learning algorithms too.

Data analysis for predicting breast cancer is carried out on unscaled data followed by feature scaling. Feature scaling of data is important in Machine Learning as it helps in standardizing all the independent variables or features available in the data for a fixed range. Though, sometimes the problem of outliers may pop up due to this scaling, which may lead to unwanted results. The results are obtained in terms of Accuracy, Precision, Recall and lastly the F1-Score. Table 1 depicts the results of each algorithm.

Table 1. Result Analysis of Algorithms

Algorithm	Accuracy	Precision	Recall	F1-Score
Support Vector Classifier	74.60	83.33	76.92	0.8000
Logistic Regression	73.56	75.78	86.66	0.8086
K-Nearest Neighbor	82.98	86.66	86.66	0.8666
Naive Bayes Classifier	62.04	64.70	84.61	0.7333
Decision Tree Classifier	83.24	89.36	87.50	0.8842
Random Forest Classifier	85.05	91.06	84.02	0.8843
AdaBoost Classifier	83.24	87.76	85.49	0.8661
XGBoost Classifier	85.86	92.30	86.15	0.8912

After fetching the results, the study concludes with comparative analysis of each of the algorithmic values with scaled / unscaled except the two ensemble algorithms which are boosting algorithms and hence does not require scaling the values. Figure 12 shows the bar chart of the comparative analysis of all the algorithms.

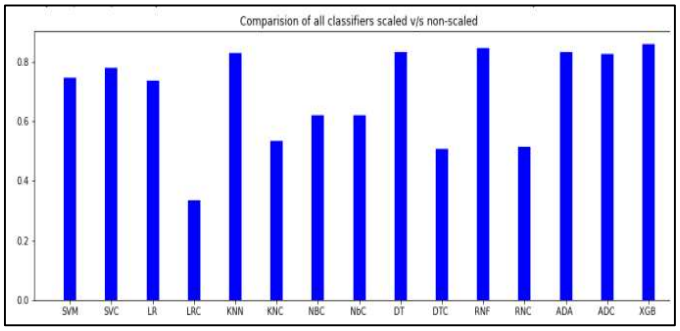

Figure 12 Bar Chart representing accuracy of each algorithm

Lastly, a scatter plot, representing the F1-scores of each algorithm is plotted, which is shown in Figure 13.

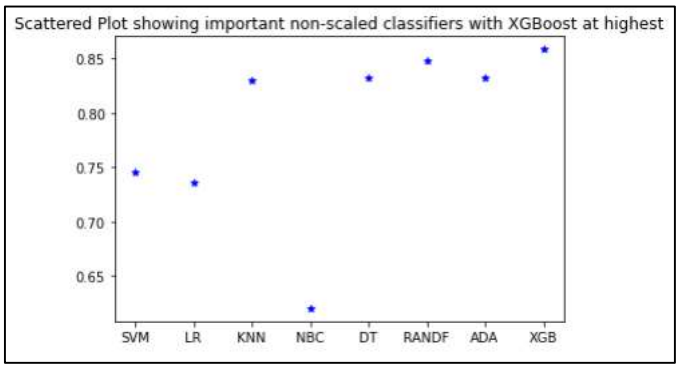

Figure 13. Scatter Plot representing the precisions of each algorithm

V. CONCLUSION AND FUTURE ENHANCEMENT

Breast Cancer prediction is a very novel topic for data analysts as it helps diagnosing breast cancer and today many researchers are also working on drug discovery on the basis of the prediction. This paper aims at predicting breast cancer with accuracy ratio of supervised algorithms in machine learning. While carrying out the experiment on the CBIS-DDSM dataset, it is found that supervised learning algorithms show pretty good results but using ensemble algorithms will enhance the accuracy and finally the F1-score. Hence, Random Forest ensemble, AdaBoost ensemble and XGBoost ensemble are used and to our hypothesis, *they have* proved to be better.

As results show in Table 1, F1-score for Ensemble methods is found to be comparatively higher than the normal supervised algorithms. From the analysis done, it can be concluded that using XGBoost Ensemble technique would enhance the performance of the model and will lead to better F1-score. F1-score is basically a harmonic mean between precision and recall and is primarily used to compare the performance of classifiers. Better the F1-Score, better is the classification of observations into perfect classes. F1-score lies between 0 to 1, and the score obtained in XGBoost is 0.8912 which is better than every other algorithm used, which justifies optimal and better classification of the observations in the dataset.

This paper can aid researchers in carrying out their studies on breast cancer prediction. It may further help in diagnosing the mammography directly. Here the mammographic dataset is converted into a csv file and then the prediction is carried out, but any researcher can directly diagnose mammography also without converting it into a csv. The limitation of the approach proposed here is that it is not multimodal, where if multiple algorithms are combined, the results yielded will be more accurate. The paper can be enhanced in future to show the AUC scores of predicted probabilities with a mammographic dataset.

References

[1] Breast cancer. (March 262021,). WHO | World Health Organization. [Online]. Available at: https://www.who.int/news-room/fact-sheets/detail/breast-cancer

[2] Breast Cancer Statistics | Facts & Figures | NBCC. (n.d.). National Breast Cancer Coalition. [Online]. Available at: https://www.stopbreastcancer.org/information-center/facts-figures/

[3] P. Bellmann, P. Thiam, F. Schwenker, "Multi-classifier-systems: Architectures, algorithms and applications," In *Computational Intelligence for Pattern Recognition*, Pedrycz, W., Chen, S.M., Eds., Springer International Publishing: Cham, Switzerland, 2018, pp. 83–113. https://doi.org/10.1007/978-3-319-89629-8_4.

[4] T. Boongoen, N. Iam-On, "Cluster ensembles: A survey of approaches with recent extensions and applications," *Comput. Sci. Rev.*, no. 28, pp. 1–25, 2018. https://doi.org/10.1016/j.cosrev.2018.01.003.

[5] L. Caplan, "Delay in breast cancer: implications for stage at diagnosis and survival," *Front Public Health*, no. 2, article no. 87, 2014. https://doi.org/10.3389/fpubh.2014.00087. PMID: 25121080; PMCID: PMC4114209.

[6] T. Chen, C. Guestrin, "XGBoost: A scalable tree boosting system," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 13–17 August 2016, pp. 785–794. https://doi.org/10.1145/2939672.2939785.

[7] D. A. Omondiagbe, S. Veeramani, A. S. Sidhu, "Machine learning classification techniques for breast cancer diagnosis," *IOP Conf. Series: Materials Science and Engineering*, vol. 495, issue 1, pp. 1-19, 2019. https://doi.org/10.1088/1757-899X/495/1/012033.

[8] C. DeSantis, F. Bray, J. Ferlay, J. Lortet-Tieulent, B. Anderson, & A. Jemal, (n.d.), "International variation in female breast cancer incidence and mortality rates," *Cancer Epidemiol Biomarkers Prev.*, vol. 24, issue 10, pp. 1495-1506, 2015. https://doi.org/10.1158/1055-9965.EPI-15-0535.

[9] M. M. Dundar, S. Badve, G. Bilgin, V. Raykar, R. Jain, O. Sertel, M. N. Gurcan, "Computerized classification of intraductal breast lesions using histopathological images," *IEEE Trans. Biomed. Eng.*, vol. 58, pp. 1977–1984, 2011. https://doi.org/10.1109/TBME.2011.2110648.

[10] S. Fauber, F. Schwenker, "Neural network ensembles in reinforcement learning," *Neural Process. Lett.*, vol. 41, pp. 55–69, 2015. https://doi.org/10.1007/s11063-013-9334-5.

[11] A. Hiba, M. Hajar, M. Hassan Al, & N. Thomas, "Using machine learning algorithms for breast cancer risk prediction and diagnosis," *Procedia Computer Science*, Elsevier, vol. 83, issue 1, pp.1064–1069, 2016. https://doi.org/10.1016/j.procs.2016.04.224.

[12] J. Wu, & C. Hicks, (n.d.). "Breast cancer type classification using machine learning," *Journal of Personalized Medicine*, vol. 11, issue 2, article no. 6), pp. 1-12, 2021. https://doi.org/10.3390/jpm11020061.

[13] M. Kächele, P. Thiam, G. Palm, F. Schwenker, M. Schels, "Ensemble methods for continuous affect recognition: Multi-modality, temporality, and challenges," *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, Brisbane Australia, 26 October 2015, pp. 9–16. https://doi.org/10.1145/2808196.2811637.

[14] K. B. Kuchenbaecker, J. L. Hopper, D. R. Barnes et al, "Risks of breast, ovarian, and contralateral breast cancer for BRCA1 and BRCA2 mutation carriers," *JAMA*, vol. 317, issue 23, pp. 2402–2416, 2017.

[15] Howlader N, Noone AM, Krapcho M, Miller D, Brest A, Yu M, Ruhl J, Tatalovich Z, Mariotto A, Lewis DR, Chen HS, Feuer EJ, Cronin KA (eds). SEER Cancer Statistics Review, 1975-2017, National Cancer Institute. Bethesda, MD, https://seer.cancer.gov/csr/1975_2017/, based on November 2019 SEER data submission, posted to the SEER web site, April 2020.

[16] S. I. Niwas, P. Palanisamy, W. Zhang, N. A. M. Isa, R. Chibbar, "Log-gabor wavelets based breast carcinoma classification using least square support vector machine," *Proceedings of the 2011 IEEE International Conference on Imaging Systems and Techniques*, Batu Ferringhi, Malaysia, 17–18 May 2011, pp. 219-223, https://doi.org/10.1109/IST.2011.5962184.

[17] O. N. Oyelade, A. A. Obiniyi, S. B. Junaidu, and S. A. Adewuyi, "ST-ONCODIAG: A semantic rule-base approach to diagnosing breast cancer base on Wisconsin datasets," *Informat. Med. Unlocked*, vol. 10, pp. 117–125, 2018, https://doi.org/10.1016/j.imu.2017.12.008.

[18] O. I. Obaid, M. Mazin Abed, M. K. Abd Ghani, S. A. Mostafa, & F. T. AL-Dhief, "Evaluating the performance of machine learning techniques in the classification of Wisconsin breast cancer," *International Journal of Engineering and Technology*, vol. 7, issue 4, pp. 160-166, 2018.

[19] D. H. Parekh, & V. Dahiya, "Predicting breast cancer using machine learning classifiers and enhancing the output by combining the predictions to generate optimal F1-score," *Biomedical and Biotechnology Research Journal (BBRJ)*, vol. 5, issue 3, pp. 331–334, 2021. https://doi.org/10.4103/bbrj.bbrj_131_21.

[20] K. Polyak, "Heterogeneity in breast cancer," *The Journal of Clinical Investigation*, vol. 121, issue 10, pp. 3786–3788, 2011. https://doi.org/10.1172/JCI60534. pmid:21965334.

[21] F. Schwenker, F. Roli, J. Kittler, (Eds.), "Multiple classifier systems," *In Proceedings of the 12th International Workshop*, Günzburg, Germany, 29 June–1 July 2015, Lecture Notes in Computer Science, Springer, Berlin, Germany, 2015, vol. 9132. https://doi.org/10.1007/978-3-319-20248-8.

[22] W. Wu, H. Zhou, "Data-driven diagnosis of cervical cancer with support vector machine-based approaches," *IEEE Access*, vol. 5, pp. 25189–25195, 2017. https://doi.org/10.1109/ACCESS.2017.2763984.

[23] N. I. R. Yassin, S. Omran, E. M. F. El Houby, H. Allam, "Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review," *Computer Methods and Programs in Biomedicine*, vol. 156, pp. 25–45, 2018. https://doi.org/10.1016/j.cmpb.2017.12.012. pmid:29428074

[24] L. Juwara, N. Arora, M. Gornitsky, P. Saha-Chaudhuri, A. M. Velly, "Identifying predictive factors for neuropathic pain after breast cancer surgery using machine learning," *International Journal of Medical Informatics*, vol. 141, 104170, 2020. https://doi.org/10.1016/j.ijmedinf.2020.104170. pmid:32544823.

[25] T. Rymarczyk, E. Kozłowski, G. Kłosowski, & K. Niderla, "Logistic regression for machine learning in process tomography," *Sensors*, vol. 19, issue 15, 3400, 2019. https://doi.org/10.3390/s19153400.

[26] A. Alarabeyyat, & M. Alhanahnah, "Breast cancer detection using k-nearest neighbor machine learning algorithm," *Proceedings of the 2016 9th IEEE International Conference on Developments in eSystems Engineering (DeSE)*, 2016, pp. 35-39.

[27] D. Berrar, "Bayes' theorem and naive Bayes classifier," Encyclopedia of Bioinformatics and Computational Biology, vol. 1, pp. 403-412, 2019. https://doi.org/10.1016/B978-0-12-809633-8.20473-1.

[28] A. Blanco-Justicia, & J. Domingo-Ferrer, "Machine learning explainability through comprehensible decision trees," *Proceedings of*

the International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Springer, Cham, 2019, pp. 15-26. https://doi.org/10.1007/978-3-030-29726-8_2.

[29] P. H. Prastyo, I. G. Y. Paramartha, M. S. M. Pakpahan, & I. Ardiyanto, "Predicting breast cancer: A comparative analysis of machine learning algorithms," *Proceeding of the International Conference on Science and Engineering*, 2020, vol. 3, pp. 455-459. https://doi.org/10.14421/icse.v3.545.

***DISHA H. PAREKH**, is working as an Assistant Professor with Indus University, Ahmedabad. She is passionate for research and carries an extensive ability to perform extraordinarily in research aspects. She is currently motivating her students to write papers during their UG level and also helping several PG students for writing quality papers in the field of computer science. She has completed her MCA and M.Phil. in Computer Science. Her area of interest lies with data science, bioinformatics and NLP approaches. She is in the education field since 2009 and she has a good databank of research papers under several headings, which can be fetched from scholar and can be contacted on disha.hparekh213@gmail.com or dishadoshi.mca@indusuni.ac.in*

***Dr. Vishal Dahiya** is working as a Professor and Head of Computer Science Department with Indus University. She has wide experience of 20+ Years. She has completed her Ph.D. in computer science from Sardar Patel University and currently is guiding more than 10 research scholars. She is acting as a chair person of research committee of Indus University for Computer Science and Engineering Departments. She is a motivator and a mentor for students and faculties interested in research. Her research area focuses on Image Processing and Big Data. She has been widely renowned for her literally work on image processing. She can be approached at cs.hod@indusuni.ac.in.*