JOURNAL PUNJAB ACADEMY OF SCIENCES (Peer Reviewed Open Access Journal)
JPAS, 23: 183-193 (January-December, 2023)
Available online at www.jpas.in
ISSN 2229-7014 (P)

BREAST CANCER PREDICTION VIA GRID SEARCH HYPERPARAMETER OPTIMIZATION

Ashima Aggarwal[1], Anurag Sharma[2]

School of Engineering, Design & Automation, GNA University, Phagwara, India [1,2]
ashimasagitarius@gmail.com[1], anurag.sharma@gnauniversity.edu.in[2]

ABSTRACT

Breast Cancer is a serious health issue worldwide, and early detection is crucial in preventing deaths. Machine learning can help identify tumors efficiently, and this paper introduces the Grid Search Hyperparameter Optimization (GSHPO) method to optimize the parameters of six existing models, including Logistic Regression, Support Vector Machine, K-Nearest Neighbor, Naïve Bayes, Decision Tree, and Random Forest. The best parameters were applied to predict outcomes in six datasets, including OWBCD, WDBC, Coimbra, BRCA, Haberman, and SEER. The results show that tuning the hyperparameters of models has a significant positive impact on prediction accuracy.

Keywords: Breast Cancer, early detection, machine learning, hyperparameter optimization, logistic regression, support vector machine, k-nearest neighbor, naive bayes, decision tree, random forest, OWBCD, WDBC, Coimbra, BRCA, Haberman, SEER

INTRODUCTION

Cancer is a devastating disease that affects millions of people worldwide, with metastatic cancer being a particularly deadly form. Breast cancer is a type of cancer that begins with the abnormal growth of cells in various parts of the breast, and it is one of the most common cancers among females. To combat this disease, machine learning algorithms have been extensively used to predict breast cancer and prevent overtreatment. Machine learning enables healthcare providers to quickly analyze data and make informed decisions regarding patients' diagnoses and treatment options, thereby enhancing overall healthcare services. This paper proposes the Grid Search Hyperparameter Optimization (GSHPO) method to optimize the parameters of six existing models, namely Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Naïve Bayes (NB), Decision Tree (DT), and Random Forest (RF), and applies the best parameters to predict the outcomes of six datasets. The accuracy of these models is evaluated both before and after the application of the GSHPO method and it is observed that the accuracy of these models is improved significantly after optimization.

The remaining section is structured as follows: The related work is presented in Section II. Section III shows the detail of the proposed methods. Section IV presents experimental results of the research. The last Section presents the conclusion of the study.

RELATED WORK

Breast cancer is a significant health issue globally and has been widely studied with the various machine learning algorithms to analyze breast cancer datasets. The studies have used a range of programming languages and software, such as WEKA, Jupyter Notebook, Matlab, R, and SAS-EM, to implement and evaluate the algorithms. The highest accuracy achieved by the different algorithms ranges from 95.9% to 99.82%, depending on the dataset, algorithm, and other factors such as hyperparameters and feature selection techniques. Some of the commonly used algorithms across the studies include SVM, KNN, NB, RF, DT, MLP, and ANN, while some studies have also explored more advanced techniques such as Bayesian Networks and Gated Recurrent Units. The studies have primarily focused on analyzing the Wisconsin Breast Cancer dataset (WBCD) and its variations (OWBCD, WBCDD, WBCPD, WDBC, and WPBC) but have used different subsets of features and pre-processing techniques. Overall, the studies highlight the effectiveness of machine learning algorithms in breast cancer detection and diagnosis, with the potential to improve clinical decision-making and patient outcomes.

PROPOSED METHODOLOGY

This section introduced the proposed methodology for Breast Cancer prediction. The flow diagram of proposed method is shown in Fig.1.

Data Analysis

This section provides access to the six datasets that have been explored in this research for the prediction of breast cancer as shown in Table 1.

Table 1. The list of publicly available datasets of Breast Cancer

Sr No	Dataset	Access	URL
1	OWBCD	Online	UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set
2	WBCD	Online	UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set
3	Coimbra	Online	https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra
4	BRCA	Online	Breast_cancer_analysis | Kaggle
5	Haberman	Online	Haberman's Survival Data Set | Kaggle
6	SEER	Online	Breast Cancer | Kaggle

B. Data Pre-processing

Data preprocessing is a data mining technique that is used to transform the raw data in a useful and efficient format.

I. To identify significant features

Here, after loading each dataset, unnecessary features need to be removed as shown in Table 2. Here, no. of features is the number of attributes, and no. of instances is the number of rows. The features from each dataset which are not required are dropped and mentioned in the drop column. After dropping the selected features, the features are labeled x and y. All features are included in x except the class label(y) which is the target variable.

II. Handling Missing Values

The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling missing data. In the datasets, tuples with missing values are handled.

III. Categorical Encoding

Most Machine Learning algorithms cannot work with categorical data and needs to be converted into numerical data. For here, converting the categorical data into numerical data, one hot encoding is used.

IV. Standardization

Feature scaling is one of the most important data pre-processing steps in machine learning. Algorithms that compute the distance between the features are biased towards numerically larger values if the data is not scaled. Here, we use Standardization which is a feature scaling technique. Standardization or Z-Score Normalization is the transformation of features by subtracting from the mean and dividing by standard deviation. This is often called a Z-score.

X_new = (X - mean)/Std

C. Splitting the dataset

The dataset is split into 2 sets: A training set and testing set. The train set would contain the data which will be fed into the model. The test set contains the data on which we test the trained data. The train set contains 80% data and the test set contains 20% data.

D. Classification

Classification is the process of predicting a discrete class label for given observation based on its features. Here, we use these 6 classifiers of Machine Learning: Logistic Regression, Support Vector Machine, K-Nearest Neighbour, Bayesian Classifier, Decision Tree, and Random Forest.

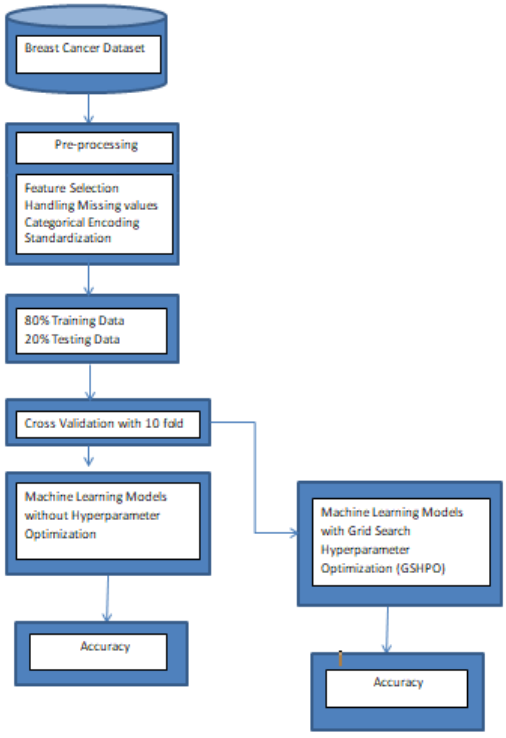

Fig.1: Flow diagram of the proposed grid search hyper-parameter optimization (GSHPO) method.

E. Hyper-parameter Tuning

It is the process of selecting best hyperparameters for a machine learning algorithm to improve its performance on a given task.

Table 2: Pre-processing of the dataset

Dataset	No. of Features	No. of Instances	Drop Features	Class Label (y)
OWBCD	11	683	Sample code number	Class: (2 for benign, 4 for malignant)
WBCD	32	569	Id	Diagnosis :

				(M = malignant, B = benign)
Coimbra	10	116	No feature drop	Classification: (1=Healthy controls, 2=Patients)
BRCA	16	341	Patient_ID, Date_of_ Surgery, Date_of_ Last_Visit	Patient_Status: (Alive/ Dead)
Haberman	4	306	Patient year of operation	Survival Status: (**1** = the patient survived 5 years or longer **2** = the patient died within 5 years).
SEER	16	4024	Unnamed: 3, Marital Status	Status: (Alive/ Dead)

Experimental Results

Algorithm

For this approach, we use Jupyter notebook for python programming. Following algorithmic steps, we follow:

I Import all the modules for feature selection, normalization, data splitting, ML models, accuracy score, and for some other required modules.
II Load the breast cancer dataset.
III Divide the dataset into features and class.
IV Check the significant features for the prediction of class.
V Rows with missing values are handled.
VI Categorical Encoding of features with one hot encoding.
VII Normalize the features to scale in one range with standard scaling.
VIII Split the dataset into two training and testing sets at 80:20 respectively.
IX Build the various machine learning models with 10 fold cross validation.
X Print accuracy of different models before hyperparameter tuning.
XI Compare the accuracy of different models after grid search hyperparameter optimization (GSHPO).

Accuracy

Accuracy is the common metric used to evaluate the performance of a machine learning model. It is the ratio of correct predictions to the total number of predictions made by the model. Here, the training and test dataset are 80:20. The accuracies of various machine learning models with the 10-fold cross-validation (cv) before hyperparameter optimization is shown in Table 3.

Table 3: The 10-fold cv score for all models before optimization

Dataset	LR	SVM	KNN	NB	DT	RF
OWBCD	96.79	97.08	97.08	96.05	95.76	97.08
WBCD	94.38	91.39	92.98	93.68	91.03	96.84
Coimbra	64.39	55.15	47.27	57.73	67.2	70.45
BRCA	78.83	79.44	76.0	64.61	65.13	78.82
Haberman	74.83	72.85	66.81	75.52	60.95	64.84
SEER	87.4	86.68	83.25	79.79	29.58	79.36

Hyper-parameter Tuning:

Hyper-parameter tuning is choosing a set of optimal hyper-parameters for a learning algorithm. A hyper-parameter is a parameter whose value is set before the learning process begins. All the machine learning models were

given different hyperparameters as shown in Table 4. Then, the Grid search tests all the combinations of hyperparameters given to the grid configuration. Grid Search Hyperparameter Optimization (GSHPO) optimizes the parameters of six existing models that are Logistic Regression(LR), Support Vector Machine(SVM), K-Nearest Neighbor(KNN), Naïve Bayes(NB), Decision Tree(DT) and Random Forest (RF) and applied the best parameters to predict the outcomes of six datasets like OWBCD (Wisconsin Breast Cancer Dataset(Original)) as shown in Table 5, WDBC (Wisconsin Diagnostic Breast Cancer) as shown in Table 6, Coimbra as shown in Table 7, BRCA(BReast CAncer gene 1) as shown in Table 8, Haberman as shown in Table 9 and SEER (Surveillance, Epidemiology, and End Results) dataset as shown in Table 10.

Table 4: The machine learning models hyperparameters configuration space

Model	Hyperparameter with search space
LR	solver' :['newton-cg', 'lbfgs', 'liblinear'], penalty' : ['l2'], C' : [100, 10, 1.0, 0.1, 0.01]
SVM	C': [0.1, 1, 10, 100, 1000], gamma': [1, 0.1, 0.01, 0.001, 0.0001], kernel': ['linear','poly','rbf','sigmoid']
KNN	n_neighbors' : range(1,21,2), weights':['uniform', 'distance'], metric' : ['minkowski','euclidean','manhattan']
NB	var_smoothing': np.logspace(0,-9, num=100)
DT	max_features': ['log2', 'sqrt'], min_samples_split':np.arange(2,16), min_samples_leaf':np.arange(1,12), criterion': ['entropy', 'gini'],
	random_state':[42]
RF	criterion':['gini','entropy'], n_estimators' : [10,100,300,600,1000], random_state':[1], max_features' : ['sqrt', 'log2']

Table 5: The 10-fold cv score of OWBCD model after applying GSHPO

Model	Best Hyperparameters	Highest Score
LR (max_iter=10000)	{'C': 100, 'penalty': 'l2', 'solver': 'newton-cg'}	96.79
SVM	{'C': 1, 'gamma': 0.01, 'kernel': 'rbf'}	97.23
KNN	{'metric': 'minkowski', 'n_neighbors': 7, 'weights': 'uniform'}	97.23
NB	{'var_smoothing': 0.1}	96.93
DT	{'criterion': 'gini', 'max_features': 'log2', 'min_samples_leaf': 2, 'min_samples_split': 13, 'random_state': 42}	95.91
RF	{'criterion': 'entropy', 'max_features': 'sqrt', 'n_estimators': 300, 'random_state': 1}	97.23

Table 6: The 10-fold cv score of WBCD model after applying GSHPO

Model	Best Hyperparameters	Highest Score
LR		

(max_iter=10000)	{'C': 100, 'penalty': 'l2', 'solver': 'newton-cg'}	96.49
SVM	{'C': 50, 'gamma': 'scale', 'kernel': 'linear'}	95.79
KNN	{'metric': 'manhattan', 'n_neighbors': 11, 'weights': 'distance'}	93.86
NB	{'var_smoothing': 2.3}	94.03
DT	{'criterion': 'entropy', 'max_features': 'log2', 'min_samples_leaf': 3, 'min_samples_split': 8, 'random_state': 42}	94.37
RF	{'criterion': 'entropy', 'max_features': 'sqrt', 'n_estimators': 600, 'random_state': 1}	97.19

Table 7: The 10-fold cv score of Coimbra model after applying GSHPO

Model	**Best Hyperparameters**	**Highest Score**
LR (max_iter=10000)	{'C': 1.0, 'penalty': 'l2', 'solver': 'newton-cg'}	70.45
SVM	{'C': 0.01, 'gamma': 'scale', 'kernel': 'linear'}	69.62
KNN	{'metric': 'manhattan', 'n_neighbors': 1, 'weights': 'uniform'}	65.38
NB	{'var_smoothing': 0.0006579332246575676}	59.39
DT	{'criterion': 'gini', 'max_features': 'log2', 'min_samples_leaf': 1, 'min_samples_split': 6, 'random_state': 42}	67.95
RF	{'criterion': 'gini', 'max_features': 'sqrt', 'n_estimators': 300, 'random_state': 1}	70.45

Table 8: The 10-fold cv score of BRCA model after applying GSHPO

Model	**Best Hyperparameters**	**Highest Score**
LR (max_iter=10000)	{'C': 1.0, 'penalty': 'l2', 'solver': 'newton-cg'}	79.44
SVM	{'C': 50, 'gamma': 'scale', 'kernel': 'linear'}	79.44
KNN	{'metric': 'minkowski', 'n_neighbors': 19, 'weights': 'uniform'}	79.44
NB	{'var_smoothing':1.0}	79.44
DT	{'criterion': 'gini', 'max_features': 'log2', 'min_samples_leaf': 8, 'min_samples_split': 2, 'random_state': 42}	78.50
RF	{'criterion': 'gini', 'max_features': 'sqrt', 'n_estimators': 100, 'random_state': 1}	79.13

Table 9: The 10-fold cv score of Haberman model after applying GSHPO

Model	Best Hyperparameters	Highest Score
LR (max_iter=10000)	{'C': 100, 'penalty': 'l2', 'solver': 'newton-cg'}	74.83
SVM	{'C': 10, 'gamma': 'scale', 'kernel': 'rbf'}	75.83
KNN	{'metric': 'minkowski', 'n_neighbors': 19, 'weights': 'uniform'}	75.49
NB	{'var_smoothing':0.001}	75.52
DT	{'criterion': 'entropy', 'max_features': 'log2', 'min_samples_leaf': 6, 'min_samples_split': 13, 'random_state': 42}	72.25
RF	{'criterion': 'gini', 'max_features': 'sqrt', 'n_estimators': 100, 'random_state': 1}	65.81

Table 10: The 10-fold cv score of SEER model after applying GSHPO

Model	Best Hyperparameters	Highest Score
LR (max_iter=10000)	{'C': 0.1, 'penalty': 'l2', 'solver': 'liblinear '}	87.85
SVM	{'C': 0.1, 'gamma': 'scale', 'kernel': 'linear'}	87.52
KNN	{'metric': 'manhattan', 'n_neighbors': 17, 'weights': 'distance'}	86.43
NB	{'var_smoothing': 0.1873817422860384}	86.93
DT	{'criterion':'entropy', 'max_features': 'log2', 'min_samples_leaf': 10, 'min_samples_split': 2, 'random_state': 42}	83.06
RF	{'criterion': 'gini', 'max_features': 'log2', 'n_estimators': 600, 'random_state': 1}	80.47

a. OWBCD dataset

The highest level of accuracy achieved was by Support Vector Machine (SVM), K-Nearest Neigbor(KNN), and Random Forest(RF) with a score of 97.08% before hyperparameter optimization.

After GSHPO, there is a change in accuracy as shown in Fig. 2. With the cross-validation and GSHPO, the highest level of accuracy achieved was by Support Vector Classifier (SVC), K-Nearest Neighbor(KNN), and Random Forest(RF) with a score of 97.3%.

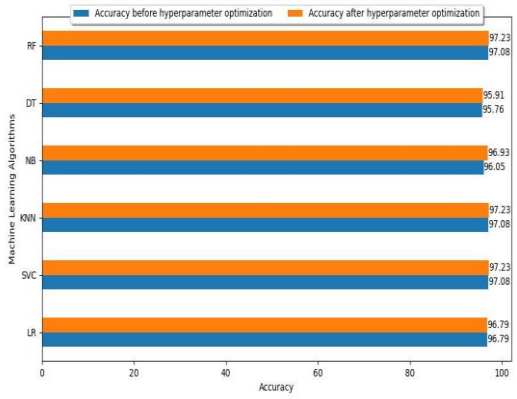

Fig. 2: Accuracy plot of OWBCD dataset before and after GSHPO

b. WBCD

The highest level of accuracy achieved was by Random Forest(RF) with a score of 96.84% before hyperparameter optimization as shown in Fig.3. After GSHPO, the highest level of accuracy achieved was by Random Forest(RF) with a 97.19%.

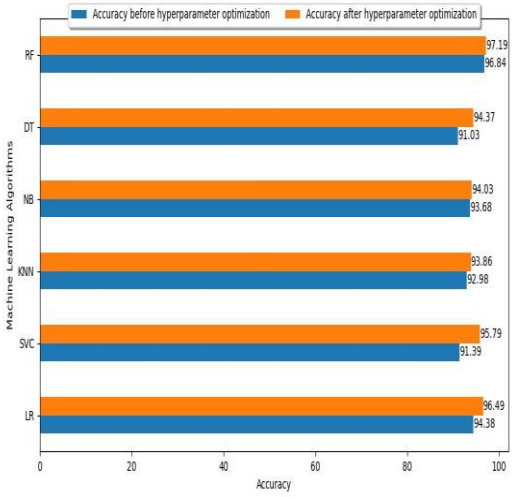

Fig. 3: Accuracy plot of WBCD dataset before and after GSHPO

c. Coimbra

The highest level of accuracy achieved was by Random Forest(RF) with a score of 70.45% before hyperparameter optimization. After Grid Search Hyperparameter Optimization (GSHPO), there is a change in accuracy as shown in Fig. 4. With the cross-validation and GSHPO, the highest level of accuracy achieved were by Logistic Regression(LR) and Random Forest(RF) with a score of 70.45%.

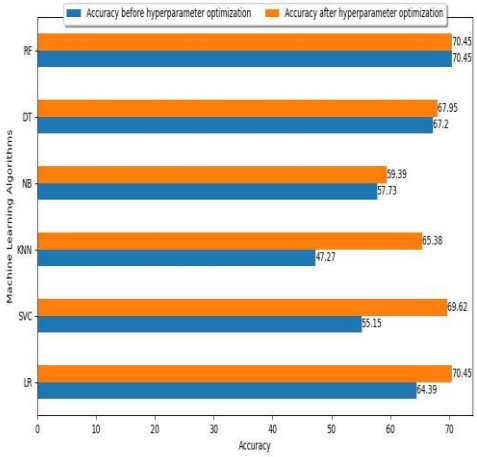

Fig. 4: Accuracy plot of Coimbra dataset before and after GSHPO

d. BRCA

The highest level of accuracy achieved was by Support Vector Classifier(SVC) with a score of 79.44% before GSHPO. After Grid Search Hyperparameter Optimization (GSHPO), there is a change in accuracy as shown in Fig. 5. With the cross-validation and GSHPO, the highest level of accuracy achieved was by Naïve Bayes(NB), K-Nearest Neigbor(KNN), Support Vector Classifier(SVC), and Logistic Regression(LR) with a score of 79.44%.

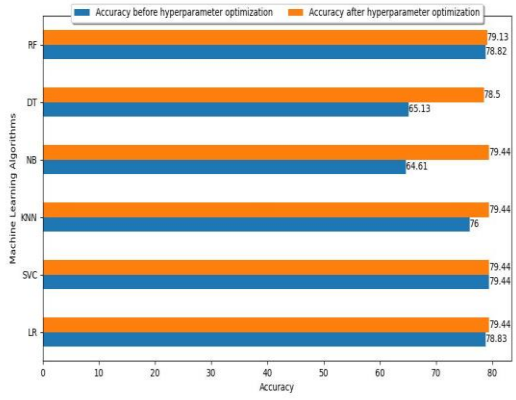

Fig. 5: Accuracy plot of BRCA dataset before and after GSHPO

e. Haberman

The highest level of accuracy achieved was by Naïve Bayes(NB) with a score of 75.52% before hyperparameter optimization. After (GSHPO), the highest level of accuracy achieved was by Support Vector Classifier(SVC) with a score of 75.83%.

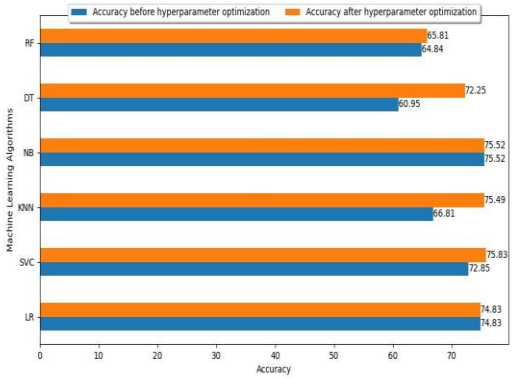

Fig. 6: Accuracy plot of Haberman dataset before and after GSHPO

f. SEER

The highest level of accuracy achieved was by Logistic Regression(LR) with a score of 87.4% before hyperparameter optimization. After Grid Search Hyperparameter Optimization (GSHPO), there is a change in accuracy as shown in Figure 7. With the cross-validation and GSHPO, the highest level of accuracy achieved was by Logistic Regression(LR) with a score of 87.85%.

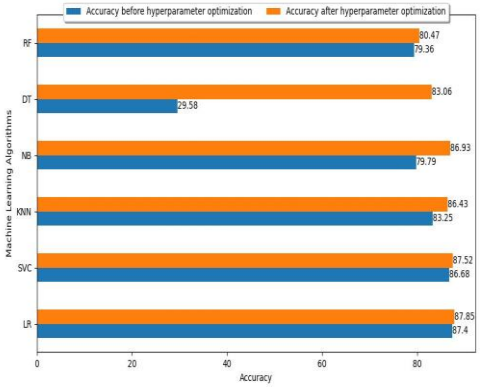

Fig. 7: Accuracy plot of SEER dataset before and after GSHPO

Conclusion

The aim of this research was the identification of different datasets and parameters responsible for breast cancer prediction and to achieve higher accuracy. The results obtained after applying 10-fold cross-validation techniques and the proposed GSHPO are promising. Our findings suggest that the hyper-parameters of tuning models have a statistically important positive impact on the model's prediction accuracy. The accuracy of various datasets can be improved by various hyper-parameter tuning algorithms such as Random Search, Bayesian Algorithm, and Genetic Algorithms. Here numerical data is taken and machine learning algorithms were applied. Deep learning techniques can be applied to these datasets. Various datasets of images which are important for breast cancer prediction need to be explored.

Acknowledgment

My sincere gratitude to my supervisor, Dr. Anurag Sharma, Professor, FEDA, GNA University, Phagwara for his invaluable guidance and support througout my research work. I am extremely grateful to Dr. Kamal Malik, Professor, CT University, Ludhiana and Dr. Prashant Singh Rana, Associate Professor, Thapar University, Patiala. I would like to acknowledge the unwavering support of my family througout my work.

REFERENCES

Amrane, S. Oukid, I. Gagaoua, and T. Ensari, “Breast cancer classification using machine learning,” 2018 Electr. Electron. Comput. Sci. Biomed. Eng. Meet. EBBT 2018, pp. 1–4, 2018, doi: 10.1109/EBBT.2018.8391453.

A. Bharat, N. Pooja, and R. A. Reddy, “Using Machine Learning algorithms for breast cancer risk prediction and diagnosis,” 2018 IEEE 3rd Int. Conf. Circuits, Control. Commun. Comput. I4C 2018, no. x, pp. 1–4, 2018, doi: 10.1109/CIMCA.2018.8739696

A. Bazila Banu and P. Thirumalaikolundusubramanian, “Comparison of bayes classifiers for breast cancer classification,” Asian Pacific J. Cancer Prev., vol. 19, no. 10, pp. 2917–2920, 2018, doi: 10.22034/APJCP.2018.19.10.2917.

A. Paulin, F, and Santhakumaran, "Classification of Breast cancer by comparing Backpropagation training algorithms,Int. J. Comput. Sci. Eng., vol. 3, no. 1, pp. 327–332, 2011.

C.P. Utomo, A.Kardiana, and R.Yuliwulandari, "Breast Cancer Diagnosis using Artificial Neural Networks with Extreme Learning Techniques," International Journal of Advanced Research in Artificial Intelligence, Vol.3, No.7, 2014.

C. Nguyen, Y. Wang, and H. N. Nguyen, “Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic,” J. Biomed. Sci. Eng., vol. 06, no. 05, pp. 551–560, 2013, doi: 10.4236/jbise.2013.65070.

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), 2016, pp. 1-4, doi: 10.1109/ICEDSA.2016.7818560.

E. A. Bayrak, P. Kirci, and T. Ensari, “Comparison of machine learning methods for breast cancer diagnosis,” 2019 Sci. Meet. Electr. Biomed. Eng. Comput. Sci. EBBT 2019, pp. 4–6, 2019, doi: 10.1109/EBBT.2019.8741990.

F. M. Agarap, “On breast cancer detection: An application of machine learning algorithms on the Wisconsin diagnostic dataset,” ACM Int. Conf. Proceeding Ser., no. 1, pp. 5–9, 2018, doi: 10.1145/3184066.3184080

G. I. Salama, M. B. Abdelhalim, and M. A. Zeid, “Experimental comparison of classifiers for breast cancer diagnosis Experimental Comparison of Classifiers for Breast Cancer Diagnosis,” no. November, 2012, doi: 10.1109/ICCES.2012.6408508.

H. Asri, H. Mousannif, H. A. Moatassime, T. Noel, “Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis,” Procedia Computer Science, Volume 83, 2016, Pages 1064-1069, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2016.04.224.

L. Rodrigues, “Analysis of the Wisconsin Breast Cancer Dataset and Machine Learning for Analysis of the Wisconsin Breast Cancer Dataset and Machine Learning for Breast Cancer Detection,” no. December, 2016.

M. M. Islam, H. Iqbal, M. R. Haque, and M. K. Hasan, “Prediction of breast cancer using support vector machine and K-Nearest neighbors,” 5th IEEE Reg. 10 Humanit. Technol. Conf. 2017, R10-HTC 2017, vol. 2018-January, pp. 226–229, 2018, doi: 10.1109/R10-HTC.2017.8288944.

O. I. Obaid, M. A. Mohammed, M. K. Abd Ghani, S. A. ostafa, and F. T. Al-Dhief, “Evaluating the performance of machine learning techniques in the classification of Wisconsin Breast Cancer,” Int. J. Eng. Technol., vol. 7, no. 4.36 Special Issue 36, pp. 160–166, 2018, doi:

10.14419/ijet.v7i4.36.23737.

P. Gupta and S. Garg, “Breast Cancer Prediction using varying Parameters of Machine Learning Models,” Procedia Comput. Sci., vol. 171, pp. 593–601, 2020, doi: 10.1016/j.procs.2020.04.064.

S. A. Mohammed, S. Darrab, S.A. Noaman and G.Saake, Data Mining and Big Data Book, 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

S. Sharma, A. Aggarwal and T.Choudary, "Breast Cancer Detection Using Machine Learning Algorithms," 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), 2018, pp. 114-118, doi: 10.1109/CTEMS.2018.8769187.

S.Sharma and S.Deshpande, “Breast Cancer Classification Using Machine Learning Algorithms,” In: Joshi A., Khosravy M., Gupta N. (eds), Machine Learning for Predictive Analysis. Lecture Notes in Networks and Systems, vol 141, Springer, Singapore. https://doi.org/10.1007/978-981-15-71060_56.