GNA Journal of
Management & Technology

Vol. XV, No. I, June 2024

OPTIMIZING MODEL PARAMETERS FOR ENHANCED BREAST CANCER PREDICTION: A MANUAL APPROACH

[1]Ashima Aggarwal, Anurag Sharma[2]

[1]*Research Scholar, School of Engineering, Design & Automation, GNA University, Phagwara*

[2] *Professor School of Engineering, Design & Automation, GNA University, Phagwara*

Email:- ashimasagitarius@gmail.com

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 manual hyperparameter optimization 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: Machine Learning; Hyperparameter optimization; Prediction; Breast Cancer; OWBCD; WDBC; Coimbra; BRCA; Haberman and SEER.

Introduction

Cancer is a disease in which some of the body's cells grow uncontrollably and spread to other parts of the body. A cancer that has spread from the place where it first formed to another place in the body is called metastatic cancer. According to WHO, 10 million people worldwide are estimated to have died from cancer in 2020. According to GOLOBOCAN 2020 reports, number of new cases of cancer of females in the World and India are maximum. Machine Learning algorithms have been widely used for the prediction of Breast Cancer. Machine Learning can be used for improving breast cancer detection and diagnosis and to prevent from overtreatment. Therefore, the aim of this research is to review the role of machine learning techniques in breast cancer detection and diagnosis.

Cancer is the leading disease as compare to any other illness; it is accountable for the greater number of mortalities worldwide. Cancer patients and Cancer deaths are rising globally.

Breast cancer occurs in breast cells, the fatty tissue or the fibrous connective tissue within the breast. Breast cancer is malignant tumors tend to become progressively worse and grow fast leading to death.

Malignant: Refers to cancer cells that can invade and kill nearby tissue and spread to other parts of the body.

Machine learning allows building models to quickly analyze data and deliver results, leveraging historical and real-time data. In this study, six machine learning classifiers which are Logistic

Regression (LR), Support Vector Machine (SVM), Naïve Bayes (NB), K-Nearest Neighbors (KNN), Decision Tree (DT) and Random Forest Classifier (RFT). We use 6 different datasets such as OWBCD (Wisconsin Breast Cancer Dataset(Original)), WDBC(Wisconsin Diagnostic Breast Cancer), Coimbra, BRCA (BReast CAncer gene 1), Haberman and SEER(Surveillance, Epidemiology, and End Results) for classification of breast cancer.

Review of Literature

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 preprocessing 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. Figure 1 illustrates the sequential steps of the proposed method.

A. Data Analysis :This section provides the access of the six datasets that have been explored in this research for prediction of breast cancer which is shown in Table 1.

B. Data Pre-processing: Data preprocessing is a data mining technique which is used to ransform the raw data in a useful and efficient format.

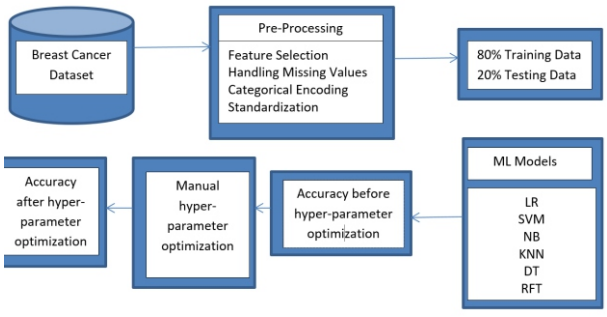


Figure 1. Flow diagram of the proposed manual hyper-parameter optimization.

:

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

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

I. To identify significant features: Here, after loading the each dataset, unnecessary features need to be removed. 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.

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)

II. Handling Missing Values: The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data [26]. In the datasets, tuples with the missing values are handled.

III. Categorical Encoding: Most Machine Learning algorithms cannot work with categorical data and needs to be converted into numerical data [27]. 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 step in machine learning. Algorithms that compute the distance between the features are biased towards numerically larger values if the data is not scaled. Tree-based algorithms are fairly insensitive to the scale of the features [28]. Here, we use the Standardization which is a feature scaling technique. **Standardization or Z-Score Normalization** is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score.

$$X\_new = (X - mean)/Std$$

C. Splitting the dataset: The dataset is splitted into 2 sets: 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 test set contains 20% data.

D. Classification: **Machine Learning** is the field of study that gives computers the capability to learn without being explicitly programmed [29]. Classification is a data analysis task, i.e. the process of finding a model that describes and distinguishes data classes and concepts. Classification is the problem of identifying to which of a set of categories (subpopulations), a new observation belongs to, on the basis of a training set of data containing observations and whose categories membership is known [30]. Here, we use these 6 classifiers of Machine Learning: Logistic Regression, Support Vector Machine, K-Nearest Neighbour, Bayesian Classifier, Decision Tree and Random Forest.

E. 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.

III. Experimental Results

A. Algorithm

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

1) Import all the modules for feature selection, normalization, data splitting, ML models, and accuracy score and for some other required modules.

2) Load the breast cancer dataset.

3) Divide the dataset as feature and class.

4) Check the significant features for prediction of class.

5) Rows with the missing values are replaced with mean values.

Dataset	Logistic Regression	Support Vector Machine	K-Nearest Neighbor with K=2	Naïve Bayes	Decision Tree with random state 0	Random Forest with 300 Trees and random state 0
OWBCD	95.62	95.62	94.16	94.89	94.16	97.08
WBCD	96.49	98.25	94.73	90.35	91.22	98.25
Coimbra	62.5	70.83	62.5	54.17	54.17	66.67
BRCA	81.54	81.54	81.54	78.46	75.38	81.54
Haberman	61.29	62.90	62.90	61.29	64.52	64.52
SEER	89.81	88.82	86.83	81.49	80.37	90.56

Table 3: Accuracy of ML models

In this approach, OWBCD has highest accuracy 97.08% in Random Forest Classifier and it is selected for hyper-parameter tuning. WBCD has highest accuracy 98.25% in case of Support Vector Machine and Random Forest Classifier. And accuracy with Random Forest classifier is selected for hyper-parameter tuning. Coimbra dataset has highest accuracy 70.83% in Support Vector Machine and it is selected for hyper-parameter tuning. BRCA has highest accuracy 81.54% in case of Logistic Regression, Support Vector Machine, K-Nearest Neighbor and Random Forest Classifier. And accuracy with Random Forest classifier is selected for hyper-parameter tuning. Haberman dataset has highest accuracy 64.52% in case of Decision Tree and Random Forest Classifier. And accuracy with Random Forest classifier is selected for hyper-parameter tuning. Random Forest Classifier is selected because it is the ensemble model. SEER has highest accuracy 90.56% in case of Random Forest Classifier and it is selected for hyper-parameter tuning.

C. 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. Manual hyper-parameter tuning is done to improve the accuracy.

a. OWBCD dataset: OWBCD has highest accuracy 97.08% in Random Forest Classifier and it is selected for hyper-parameter tuning. When n-estimator is 100 to 1900, there is change in accuracy as shown in figure 2. When the value of n-estimator is 500, 600 and 700, accuracy is 98.08%.

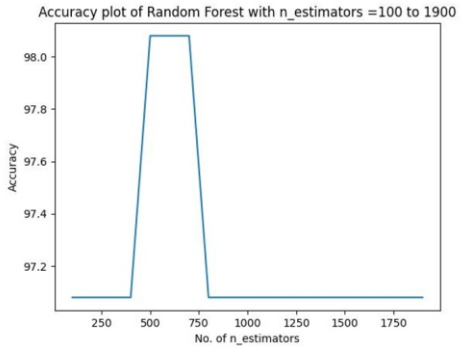


Figure 2: OWBCD dataset

b. WBCD: WBCD has highest accuracy 98.25% in case of Support Vector Machine and Random Forest Classifier. And accuracy with Random Forest classifier is selected for hyper-parameter tuning. When n-estimator is 100 to 1900, there is change in accuracy as shown in figure 3. When the value of n-estimator is 300 and 400, accuracy is 98.25%.

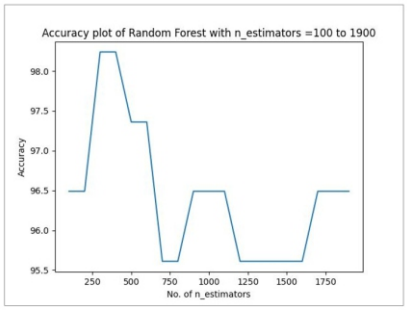


Figure 3: WBCD dataset

c. Coimbra: Coimbra dataset has highest accuracy 70.83% in Support Vector Machine and it is selected for hyper-parameter tuning. When value of C is 1 to 100, there is change in accuracy as shown in figure 4. When the value of C is 1 and 11, accuracy is 72.83%.

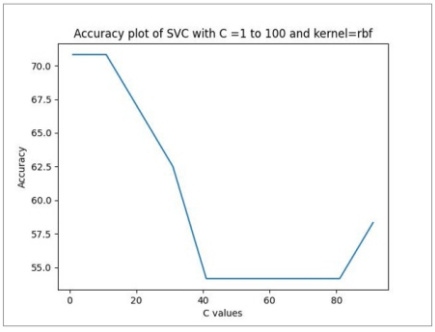


Figure 4: Coimbra dataset

d. BRCA: BRCA has highest accuracy 81.54% in case of Logistic Regression, Support Vector Machine, K-Nearest Neighbor and Random Forest Classifier. When n-estimator is 100 to 1900, there is change in accuracy as shown in figure 5. When the value of n-estimator is 900, 1200 and 1300, accuracy is 83.74%.

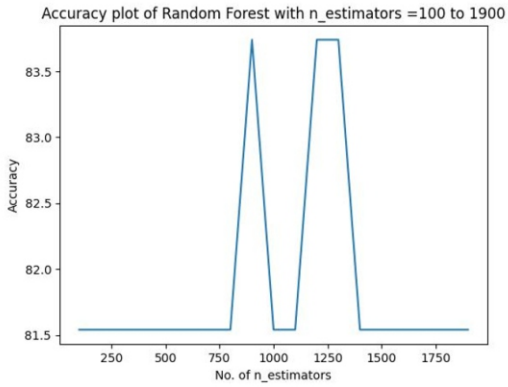


Figure 5: BRCA dataset

e. Haberman: Haberman dataset has highest accuracy 64.52% in case of Decision Tree and Random Forest Classifier. And accuracy with Random Forest classifier is selected for hyper-parameter tuning. When n-estimator is 100 to 1900, there is change in accuracy as shown in figure 6. When the value of n-estimator is 600, 700, 900, 1100, 1200, 1300, 1400, 1500, 1600, 1700, 1800 and 1900, accuracy is 66.13%.

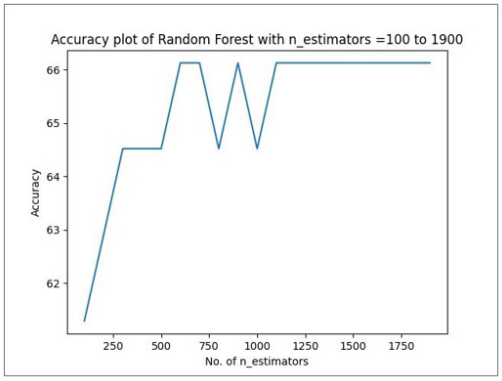
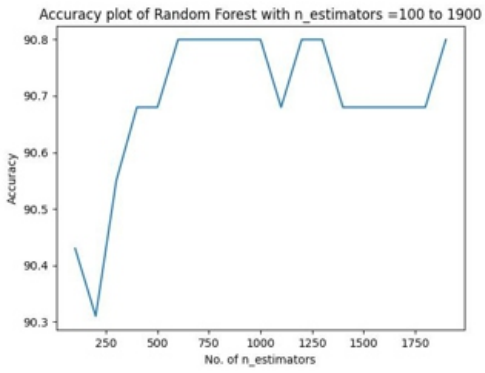


Figure 6: Haberman dataset

f. SEER: SEER has highest accuracy 90.56% in case of Random Forest Classifier and it is selected for hyper-parameter tuning. When n-estimator is 100 to 1900, there is change in accuracy as shown in figure 7. When the value of n-estimator is 600,700, 800,900, 1000, 1200, 1300 and 1900, accuracy is 90.81%.



Conclusion:

The aim of this research was identification of different datasets responsible for breast cancer prediction and to achieve higher accuracy. Accuracy of various datasets is improved by manual hyper-parameter tuning. Some other hyper-parameter tuning algorithms such as Grid Search, Random Search, Bayesian Algorithm and Genetic Algorithms can be used for improving the accuracy. Different features selection techniques can be used such as univariate, bi-variate tests and chi-square tests.

References:

National Cancer Institute (NIH) Available online: https://www.cancer.gov/about cancer/understanding/what-is-cancer (accessed on 25 Feb, 2021)

World Health Organization (WHO) Available online: https://www.who.int/news- room/fact sheets/detail/cancer (accessed on 27 Feb, 2021)

Global Cancer Society (GLOBOCAN 2020) Available online: https://www.uicc.org/news/globocan-2020-new-global-cancer-data (accessed on 2 Mar, 2021)

World Health Organization (WHO) Available online: https://gco.iarc.fr/today/data/factsheets/populations/356-india-fact-sheets.pdf (accessed on 4 Mar, 2021)

Centers for disease control and prevention (CDC) Available online: https://www.cdc.gov/cancer/breast/basic_info/what-is-breast cancer.htm#:~:text=The%20kind%20of%20breast%20cancer,the%20glands%20that%20produce %20milk (accessed on 10 March, 2021)

Cleveland Clinic Organization Available online: https://my.clevelandclinic.org/health/diseases/3986-breast-cancer (accessed on 14 March, 2021)

Cancer.Net Website Available Online: https://www.cancer.net/cancer-types/breast-cancer/types treatment (accessed on 16 March, 2021)

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/10.1007/978 981-1571060_56.

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

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.

E. A. Bayrak, P. Kirci, and T. Ensari, *f*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.

O. I. Obaid, M. A. Mohammed, M. K. Abd Ghani, S. A. Mostafa, 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.

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.

A. 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.

M. 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, *f*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.

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.

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.

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.

C.P. Utomo, A.Kardiana and R.Yuliwulandari, "Breast Cancer Diagnosis using Artificial Neural Networks with Extreme Learning Techniques," Internaltional Journal of Advanced Researc 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.

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.

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

https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/
https://www.geeksforgeeks.org/ml-one-hot-encoding-of-datasets-in-python/
https://www.geeksforgeeks.org/normalization-vs-standardization/
https://www.geeksforgeeks.org/machine-learning/
https://www.geeksforgeeks.org/basic-concept-classification-data-mining/