KOKAND UNIVERSITY
HERALD
SCIENTIFIC E-JOURNAL
ISSN: 2181-1695

QO‘QON UNIVERSITETI XABARNOMASI
KOKAND UNIVERSITY HERALD
ВЕСТНИК КОКАНДСКОГО УНИВЕРСИТЕТА
OAK: 01-08/1819/6

HERALD
www.herald.kokanduni.uz

MACHINE LEARNING METHODS FOR BREAST CANCER CLASSIFICATION BY USING DATA SCIENCE TECHNIQUES

Meliboev Azizjon Ikromjon o‘g‘li
Faculty of Digital Technologies and Mathematics,
Kokand University,

MAQOLA HAQIDA

Qabul qilindi: 24-iyun 2024-yil
Tasdiqlandi: 26-iyun 2024-yil
Jurnal soni: 11
Maqola raqami: 31
DOI: https://doi.org/10.54613/ku.v11i11.969

KALIT SO‘ZLAR/ Ключевые слова/ keywords

Machine learning, Breast cancer, Data science, data analysis, technique, machine, model

ANNOTATION

This study explores the sensitivity analysis of various machine learning methods applied to the probl of breast cancer classification. By examining the robustness and performance of different algorithms, aim to identify the most reliable techniques for accurate diagnosis. We assess the impact of key paramet and data variations on model outcomes to provide a comprehensive understanding of each methc strengths and limitations. Our findings offer valuable insights into the selection and optimization machine learning models for breast cancer detection, ultimately contributing to improved diagno accuracy and patient care

Introduction: Breast cancer remains one of the most prevalent and life-threatening diseases affecting women worldwide. Early and accurate detection is crucial for improving patient outcomes and survival rates. In recent years, machine learning (ML) techniques have shown great promise in enhancing the diagnostic process for breast cancer by automating the analysis of medical data and identifying patterns that may not be readily apparent to human experts. However, the effectiveness of these techniques can vary significantly based on the chosen algorithm, parameter settings, and the nature of the data used.

Sensitivity analysis in machine learning involves systematically varying input parameters to evaluate their impact on the model’s performance. This approach is essential for understanding the robustness and reliability of different ML methods in real-world applications, such as breast cancer classification. By identifying how sensitive each method is to changes in data and parameters, researchers and practitioners can make more informed decisions about which algorithms to deploy in clinical settings.

This study aims to conduct a comprehensive sensitivity analysis of various machine learning methods applied to the breast cancer classification problem. Using advanced data science techniques, we will explore the performance of different algorithms under varying conditions to determine their strengths, weaknesses, and suitability for medical diagnostics. By doing so, we hope to provide valuable insights that can guide the selection and optimization of machine learning models, ultimately enhancing the accuracy and reliability of breast cancer detection.

In this paper, we will first review the existing literature on machine learning applications in breast cancer classification and sensitivity analysis techniques. Next, we will describe the data sets and methodologies used in our experiments, followed by a detailed analysis of our findings. Finally, we will discuss the implications of our results for future research and clinical practice, highlighting the potential benefits and challenges of integrating machine learning into breast cancer diagnostics.

Related works: In the following we highlight previous studies which are closely related. Khourdifi et al. proposed an overview of the evolution of large data in the health system, and apply four ML classification algorithms which are Random Forest, Naive Bayes, Support Vector Machines SVM, and K-Nearest Neighbors to a breast cancer data set which is Wisconsin Hospitals Madison Breast Cancer Patríciol et al. (2013). They used an effective way to predict breast cancer based on patients’ clinical records. According to the performance of models, SVM achived the highest accuracy score which is 97.9%.

Yue and Wang provide an overview of machine learning algorithms that Artificial neural networks, SVM, Decision trees, and k-nearest neighbors. Their primary data is drawn from the Wisconsin breast cancer database (WBCD) which is the benchmark database for comparing the results through different algorithms. Machine learning algorithms that have been used on the WBCD database in diagnosis and prognosis show different levels of accuracy that ranged between 94.36% and 99.90%. In the same way, Assiri AS. et al. used simple Logistic regression, SVM with stochastic gradient descent optimization and Multilayer perceptron network and ensemble of classification used for a voting mechanism. They also evaluated the performance of hard and soft voting mechanism. For 10, 5 and 2 fold cross validation, the proposed algorithm achieved accuracies of 98.77%, 98.43%, and 99.23% respectively Alghunaim et al. address the problem of breast cancer prediction in the big data context. They considered two varieties of data that gene expression (GE) and DNA methylation (DM) and chosed Apache Spark as a platform. To create models that help in predicting breast cancer, they selected three different classification algorithms which are support vector machine, decision tree, and random forest. They conducted a comprehensive comparative study using three scenarios with the GE, DM, and GE and DM combined. These authors used two datasets from The Cancer Genome Atlas dataset. The results showed that SVM classifier in the Spark environment outperforms other classifiers by accuracy score of 99.68%.

Bharat et al. used SVM algorithm end evaluated their approach on the Wisconsin Breast Cancer dataset. For comparative study, they trained with the other algorithms such as KNN, Naives Bayes and decision tree variant of CART. Their accuracy of prediction for each algorithm is compared and results are analyzed. Bayrak et al. also used same dataset and applied SVM and ANN algorithms for prediction of the classification of breast cancer. SVM has showed the best performance in the accuracy of 96,9% for the diagnosis and prediction Rehman et al. proposed the implementation of ML models using Logistic Regression, SVM and KNN is done on the dataset taken from the UCI repository. Experimental results showed that SVM is the best for predictive analysis with an accuracy score of 92.7% and next KNN and Logistic regression with accuracy score respectively with 92.23%, 92.10%. In contrast, Sharma et al. focus integrate that ML algo rithms with feature selection methods and compare their performances to identify the most suitable approach. They investigate SVM, ANN and Naïve Bayes by using the Wisconsin Diagnostic Breast Cancer dataset and SVM achieved remarkable accuracy score which is 98.82%. In our work, we investigate the effect of feature selection techniques on improving the performance of a given machine learning-based models. Hence, we focus on this scope.

2.1 Dataset description.

We used one of the widely used Breast Cancer Coimbra dataset created by the Faculty of Medicine of the University of Coimbra and the University Hospital Centre of Coimbra. The dataset gathered 9 independent variables of health information from 64 breast cancer patients and 52 healthy people. The dataset contains following features that Age(years) which is age of patient, BMI(g/m2) is for body mass index of patient, Glucose(mg/dL) is for blood sugar of patient, Insulin(μ U/mL) is a hormone made by the pancreas, HOMA(homeostatic model assessment) is a method used to quantify insulin resistance and beta-cell function, Leptin(ng/mL) is a hormone produced mainly by adipocytes (fat cells) that is involved in the regulation of body fat, Adiponectin(μ g/mL) calculates a protein produced and secreted by fat cells that is normally abundant in the blood plasma, Resistant(ng/mL) means the bacteria can grow even if the drug is present, MCP-1(pg/dL) is Monocyte chemotactic protein-1 that is the most important chemokine that regulates migration and infiltration of monocytes/macrophages. Description of features are shown in Table 1. Also target variable that is encoded as 1 for healthy controls, 2 for patient with breast cancer. The gathered dataset is stored in a CSV file with 10 columns with first 9 are features and last column indicates whether breast cancer is malignant. The total number of rows is 116 which consists of the patient and healthy control groups. You can see the count and distribution of target variable.

Table 1. Statistical data description

	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1
mean	57.3	27.6	97.8	10.0	2.7	26.6	10.2	14.7	534.6
std	16.1	5.0	22.5	10.1	3.6	19.2	6.8	12.4	345.9
min	24.0	18.4	60.0	2.4	0.5	4.3	1.7	3.2	45.8
25% percentile	45.0	23.0	85.8	4.4	0.9	12.3	5.5	6.9	270.0
50% percentile	56.0	27.7	92.0	5.9	1.4	20.3	8.4	10.8	471.3
75% percentile	71.0	31.2	102.0	11.2	2.9	37.4	11.8	17.8	700.1
max	89.0	38.6	201.0	58.5	25.1	90.3	38.0	82.1	1698.4

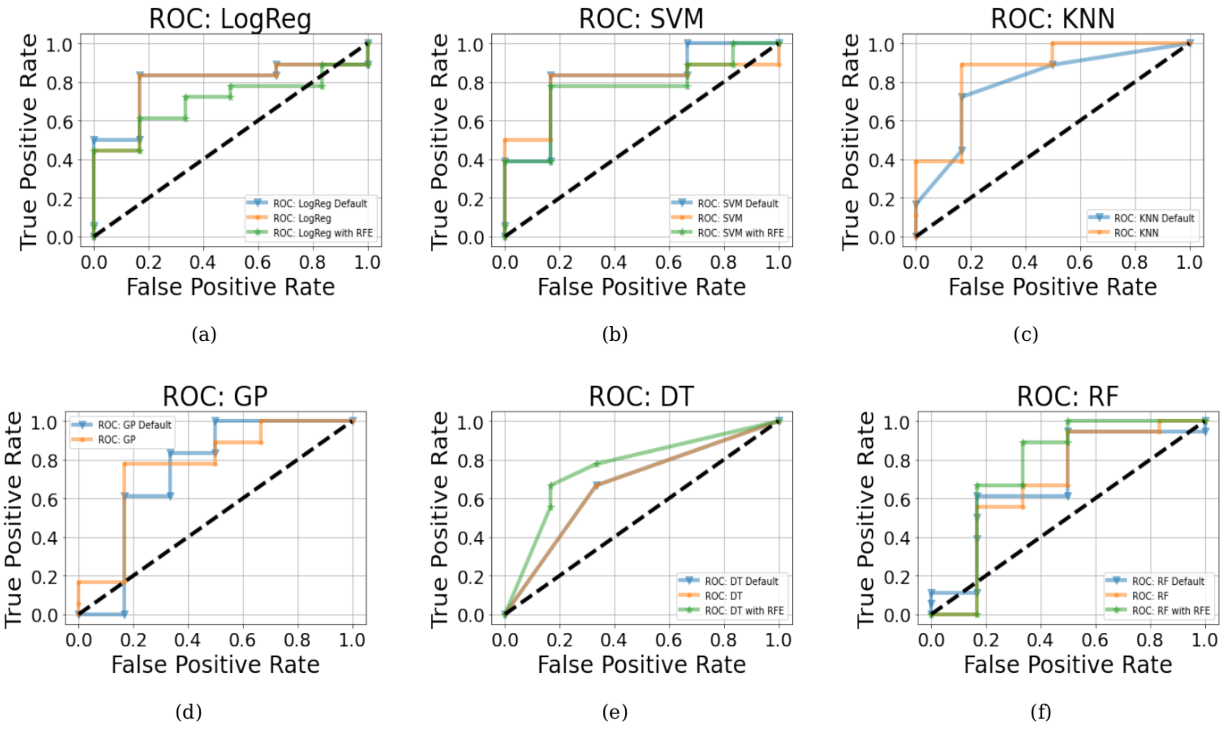
Methodologies and Results: Machine learning classifiers are algorithms that automatically categorize or sort data into one or more "classes" and this task is a type of supervised learning problem. In general, machine learning classifiers have two types of parameters: those that are learned from the training data such as the weights in logistic regression, and the parameters of a learning algorithm that are optimized separately. The latter are the tuning hyper-parameters of a model. For example, the strength of a regularization parameter in logistic regression or the parameter that defines the maximum depth of a decision tree. We applied one of popular hyperparameter optimization technique called Grid search, which can further help to improve the performance of a model by finding the optimal combination of hyper-parameter values. The Grid search approach is quite simple that can search for optimal parameters from a from a list of potential values for different hyperparameters and the grid search algorithm evaluates the model performance for each com bination to obtain the classifier with optimal combination of hyper-parameter values.

Cross validation is when part of the data is reserved (holdout) for evaluating the model. For small datasets, it is common to use K-Fold Cross Validation where data is split into train/test set K times and results are aggregated. Evaluating the model on each part of data allows to take advantage of all the data for evaluation. Additionally, repeating the split K-times ensures to mitigate the bias of particular split. Another approach, more commonly used on larger datasets is simple train/test set split where model development (training, validating) is done on train set and evaluation is performed on test set. Due to its simplicity we resort to using this approach in our experiments.

We speculate this is effected by the small data size where training set tuned hyper parameters are not the optimal ones for test set. Another interesting observation is RFE improved performance for LogReg, DT and RF while it insignificant effect on SVM. Useful insight from the results is that SVM with default parameters are well suited for our task. This could be perhaps explained by decision boundary of SVM where only specific examples are needed. In other words since SVM needs data points only in the decison boundary, it is well suited for small scale datasets. Another, aspect of comparision is ROC Curve which provides evaluation of classifier on multiple thresholds. As can be seen from Fig. 8, RFE based feature selection is improving perfromance only for DT and RF. RFE negatively affects LogReg and SVM.

Figure 1.

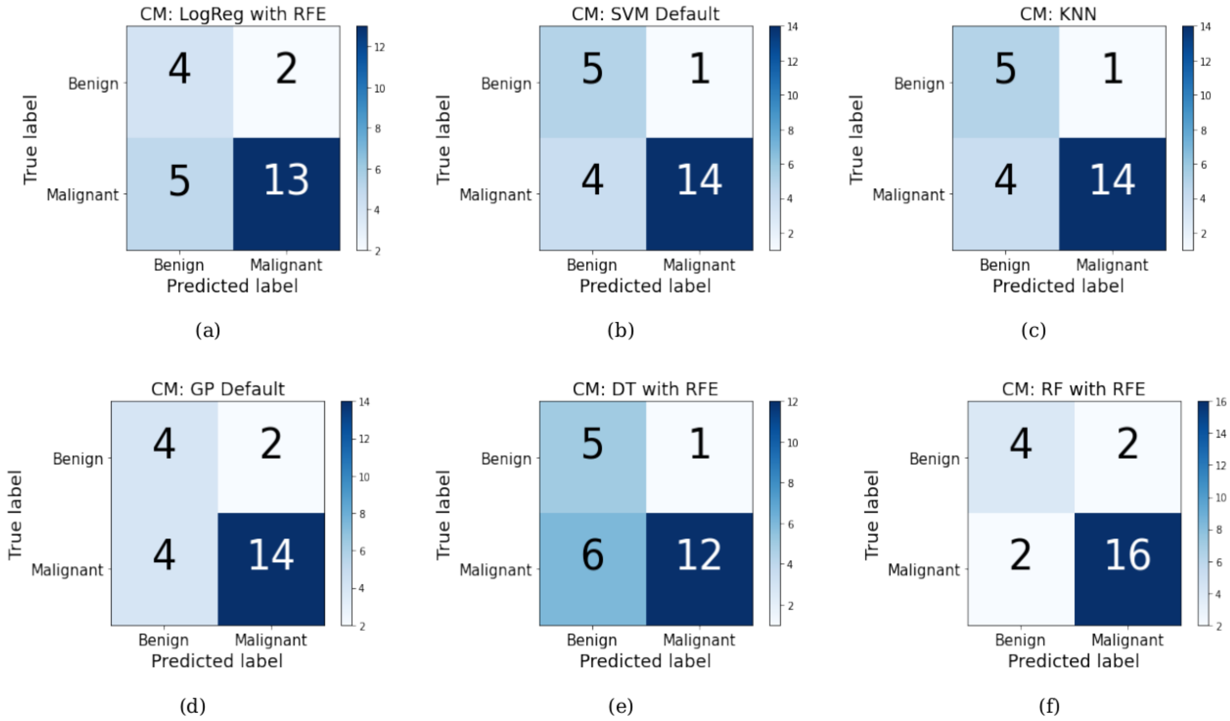
Comparison of Classifiers with ROC Curve



(a) (b) (c)

(d) (e) (f)

In this section we describe the confusion matrix as another aspect of classifier performance. For each classifier, their best setting from F1-score is used as our primary metric. Description of confusion matrix is explained in Experimental Setting section. Default threshold of 0.5 for binary classification setting is used for computing CM. It is noteworthy to mention that CM is depicted only for test set which is not seen during model development. Main takeaway message from this section is that, feature selection and hyper parameter tuning does not always improve performance. For some classifiers it improved but for some default setting resulted in best empirical performance in Figure 2.



(a) (b) (c)

(d) (e) (f)

Results and Discussion: Data preprocessing is a critical step in the data analysis and machine learning pipeline. It involves transforming raw data into a clean and usable format. This step ensures that the data is consistent, accurate, and suitable for analysis or model training. We used various techniques for analising the data.

- Check Missing values
- Check Duplicates
- Check data type
- Check the number of unique values of each column
- Check statistics of the dataset
- Check various categories present in the different categorical columns

Handling missing values is an essential part of data cleaning during EDA. Here is bar plot of missing values in Figure 1. Figure 2. Confusion matrix of Models

Conclusion: In this study, we conducted a comprehensive sensitivity analysis of various machine learning methods applied to breast cancer classification using advanced data science techniques. Our analysis revealed significant insights into the robustness and reliability of different machine learning algorithms under varying conditions, providing a clearer understanding of their performance in the context of breast cancer diagnostics.

The results indicate that while certain machine learning models demonstrate high accuracy and robustness, their performance can be significantly influenced by specific parameters and data variations. This underscores the importance of carefully selecting and tuning machine learning models to ensure optimal performance in clinical settings.

Moreover, our study highlights the necessity of incorporating sensitivity analysis in the development and evaluation of machine learning models for medical applications. By systematically assessing how changes in input data and model parameters affect outcomes, we can better understand each algorithm's strengths and limitations, leading to more informed decisions in their deployment for breast cancer detection.

Future research should focus on exploring more advanced and hybrid machine learning techniques, as well as validating findings on larger and more diverse datasets. Additionally, integrating domain expertise from oncologists and radiologists can further enhance the practical applicability and accuracy of these models.

In conclusion, the insights gained from this sensitivity analysis not only contribute to the field of machine learning in breast cancer classification but also pave the way for more reliable and accurate diagnostic tools, ultimately improving patient outcomes and advancing personalized medicine.

References:

1. Meliboev, A., Alikhanov, J., & Kim, W. (2022). Performance evaluation of deep learning based network intrusion detection system across multiple balanced and imbalanced datasets. *Electronics*, *11*(4), 515.

2. Azizjon, M., Jumabek, A., & Kim, W. (2020, February). 1D CNN based network intrusion detection with normalization on imbalanced data. In *2020 international conference on artificial intelligence in information and communication (ICAIIC)* (pp. 218-224). IEEE.

3. Armane, M., Oukid, S., & Ensari, T. (2018). Breast cancer classification using machine learning. *IEEE*, 1-4.

4. Fatima, N., Liu, L., Hong, S., & Ahmed, H. (2020). Prediction of breast cancer, comparative review of machine learning techniques, and their analysis. *IEEE Access*, 8, 150360-150376.

5. Khourdifi, Y., & Bahaj, M. (2018). Applying best machine learning algorithms for breast cancer prediction and classification. *IEEE*, 1-5.

6. Patrício, M., Pereira, J., Crisóstomo, J., Matafome, P., Gomes, M., Seiça, R., & Caramelo, F. (2018). Using Resistin, glucose, age and BMI to predict the presence of breast cancer. *BMC Cancer*, 1-8.

7. Yue, W., Wang, Z., Chen, H., Payne, A., & Liu, X. (2018). Machine learning with applications in breast cancer diagnosis and prognosis. *Designsb 2*, 13, 39.

8. Assiri, A. S., Nazir, S., & Velastin, S. A. (2020). Breast tumor classification using an ensemble machine learning method. *Journal of Imaging*, 6(39), 39.

9. Alghunaim, S., & Al-Baity, H. H. (2019). On the scalability of machine-learning algorithms for breast cancer prediction in big data context. *IEEE Access*, 91535-91546.

10. The Cancer Genome Atlas—Data Portal. (2018). Retrieved from https://portal.gdc.cancer.gov/.

11. Bharat, A., Pooja, N., & Reddy, R. A. (2018). Using machine learning algorithms for breast cancer risk prediction and diagnosis. *IEEE I4C*, 1-4.

12. Bayrak, E. A., Kırcı, P., & Ensari, T. (2019). Comparison of machine learning methods for breast cancer diagnosis. *IEEE EBBT*, 1-3.