Comparative Analysis of Breast and Prostate Cancer Prediction Using Machine Learning Techniques

Samta Rani, Tanvir Ahmad, and Sarfaraz Masood

Abstract Around the whole world, cancer is the most life-threatening disease. Basically, cancer can arise in any tissue of the body, and while each variety of cancer has unique characteristics, the fundamental processes that might cause cancer are highly common in all disease types. Breast cancer is one of the most ubiquitous types of cancer in females. In males, prostate cancer is the most dangerous during recent years. This study focuses on breast cancer as well as on prostate cancer in the direction of their early predictions. For early prediction, eight classification models had been used such as logistic regression (LR), Naïve Bayes (NB), decision tree (DT), stochastic gradient descent (SGD), K-nearest neighbors (KNN), decision tree (DT), random forest (RF), support vector machine (SVM), and artificial neural network (ANN). This work includes three different datasets for research analysis of breast and prostate cancer predictions. Two datasets for breast cancer (Coimbra and Wisconsin) and one for prostate cancer are taken from UCI and Kaggle repository, respectively. For improving the results of prediction, the normalization technique and feature selection method had been used in this paper. Performance in terms of accuracy, precision, recall, F1-score, and curves of each classifier are analyzed in this study. Most of the classifiers did well after using the feature selection method (ANOVA). In the case of Breast Cancer Coimbra, KNN give good results with 80% accuracy in both the cases with or without using feature selection. Logistic regression with feature selection doing the best work on Wisconsin Breast Cancer with 99% accuracy. There are four classifiers (SVM, RF, DT, and SGD) which gives highest accuracy (97%) on prostate cancer.

S. Rani (✉) · T. Ahmad · S. Masood
CSE, Jamia Millia Islamia University, Delhi, India
e-mail: rani.samta@gmail.com

T. Ahmad
e-mail: tahmad2@jmi.ac.in

S. Masood
e-mail: smasood@jmi.ac.in

D. Gupta et al. (eds.), *International Conference on Innovative Computing and Communications*, Lecture Notes in Networks and Systems 473,
https://doi.org/10.1007/978-981-19-2821-5_54

Keywords Breast cancer · Prostate cancer · Feature selection · Normalization · Classifications

1 Introduction

Cancer is one of the leading causes of mortality worldwide, based on WHO statistics. Breast cancer is the second most common cancer, after lung cancer, with 2.09 million cases among the predicted 9.6 million cancer fatalities. It is also the fifth most prevalent cause of cancer death, accounting for over 627,000 fatalities, or 15% of all cancer deaths among women. And breast cancer alone accounts for 30% of all new cancer diagnoses in women [1]. This work examined the breast cancer issue using publicly available data from the Portuguese city of Coimbra and Wisconsin. There were ten quantitative predictor factors in this dataset, which were anthropometric in nature and captured through standard blood tests used to determine the presence or absence of breast cancer. Breast cancer is the most frequent type of cancer in women, affecting about 2.1 million women each year and contributing to female cancer deaths being the leading cause of death. Breast cancer claimed the lives of over 627,000 women in 2018. Early detection is crucial for improving breast cancer and survival chances [2]. Prostate cancer is one of the most frequent malignancies in American males, and it has the second highest fatality rate after lung cancer. Now a days, one in every seven men would be diagnosed with prostate cancer. According to recent figures, the number of new patients diagnosed with prostate cancer in 2017 was approximately 161,360, with approximately 26,730 deaths [3]. Fortunately, if prostate cancer is detected early, the mortality rate can be reduced. This paper also includes the study on prostate cancer whose dataset is taken from Kaggle and analyzes all classification models on parameters of prostate cancer. This paper is organized as follows: Sect. 1 presents the introduction to the different types of cancer disease. Section 2 presents the review of various recent literatures for cancer detection. Section 3 describes each component of the methodology used in this work, which is followed by description of the datasets. The results obtained after various experiments are presented and discussed in Sect. 5 followed by the conclusion.

2 Related Work

Rahman et al. [4], the purpose of this research is twofold. The first is to identify the most relevant breast cancer biomarkers, and second is to improve the current computer-aided diagnostic (CAD) system for detecting early breast cancer. This work made use of a dataset that included nine anthropometrical and clinical variables. From all the techniques used by author, SVM model with radial basis function (RBF) kernel gives best results with 93.9% accuracy, 95% sensitivity, and 94% specificity.

Ray et al. [5], in this study, researchers worked on two different datasets. One dataset is based on diabetic, and another is based on breast cancer. Feature selection techniques also applied before applying the machine learning models for getting the reduced feature set to classify between healthy and non-healthy subjects. Feature set includes the features having majority that is generated by routine pathology examinations. Author focused on identifying biomarkers that entail pathological testing and those that do not.

Mushtaq et al. [6], in this research, breast cancer (Wisconsin) dataset was used for study. Different classification models are applied along with PCA reduction approach. Performance of different classifiers with variants of PCAs based on linear, sigmoid, cosine, poly, and radial basis functions is analyzed. Highest 99.20% accuracy got from sigmoid-based Naive Bayes. Using KNN, with all different kernels got accuracies within the range 96.4–97.8%.

Shakeel et al. [7] works on prostate cancer for which author initially collects information related to prostate cancer from DBCR dataset. After that, using mean mode process, irrelevant record was removed and collect other important elements using ant rough set hypothesis. Result is evaluated in the terms of mean square error rate, hit rate, and accuracy.

3 Proposed Methodology

Figure 1 depicts the workflow of proposed work, highlighting the overall steps taken in this work, which includes data preprocessing with normalization, feature selection techniques, training and testing with specified models, evaluation of results, and

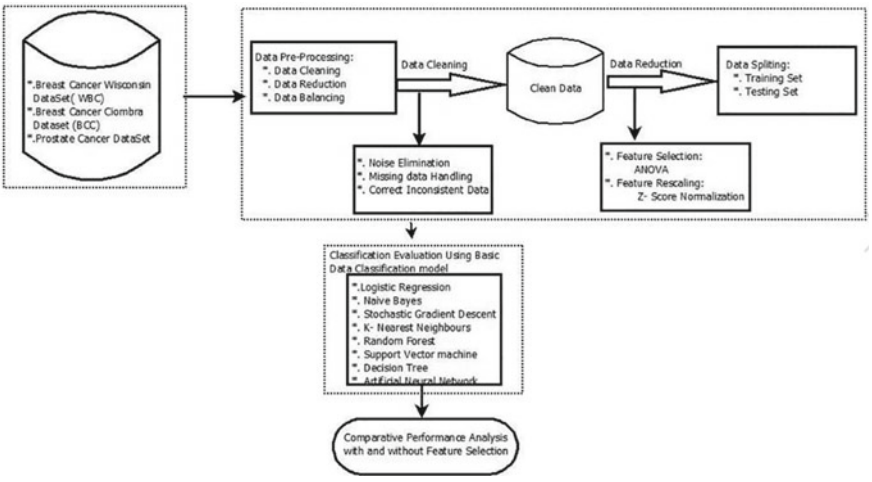


Fig. 1 Model for predicting cancer disease

Table 1 Description of breast and prostate cancer datasets

Dataset name	Total number of Patients	Number of parameters	Number of non-cancer patients	Number of cancer patients
Breast cancer coimbra data set (BCC)	116	10	52 (45%)	64 (55%)
Breast cancer wisconsin (diagnostic) data set (WBC)	569	32	357 (63%)	212 (37%)
Prostate cancer	100	10	38 (38%)	62 (62%)

prediction of breast cancer and prostate cancer. Python 3 was used to carry out this task.

Dataset

In this paper, three datasets had been used or analyzed for covering the famous cancer types in both males and females. Two datasets are based on breast cancer named as Breast Cancer Coimbra dataset and Breast Cancer Wisconsin, both had been collected from UCI repository. Third dataset had been collected from Kaggle named as prostate cancer. Table 1 shows the number of records under cancerous and non-cancerous cases in each dataset.

Coimbra Breast Cancer dataset has clinical parameters like body mass, hormone, leptin, glucosamine, etc. But another dataset which is also a breast cancer dataset WBC includes the real-valued parameters for each cell nucleus like texture, radius, compactness, etc. In this dataset, for each image, mean, standard error, and worst values were computed. Prostate cancer dataset having ten features like area, perimeter, radius, identification number, etc. In this paper, label 0 is used for non-cancer patients and label 1 for cancer patients.

4 Result Analysis

The proposed work considers eight classifiers for the analysis of performance comparison. Two normalization methods Z-score and min–max are used for data transformation. But, in this paper, only best results are discussed. Out of Z-score and min–max, Z-score gives good results. Tables 2, 3, and 4 show the results of BCC dataset, WBC dataset and prostate cancer dataset, respectively, using all the machine learning techniques. Every table divided into two parts having results based on without using ANOVA and with ANOVA.

Table 2 shows the comparison of results using the eight classifiers without feature selection and with feature selection on Breast Cancer Coimbra dataset. All classifiers

Table 2 Performance analysis of BCC

Models		Without feature selection				With feature selection			
		Precision	Recall	F1-score	Acc (%)	Precision	Recall	F1-score	Acc (%)
LR	0	0.64	0.41	0.50	60	0.67	0.71	0.69	69
	1	0.58	78	0.67		0.71	0.67	0.69	
KNN	0	0.86	0.71	0.77	80	0.75	0.88	0.81	80
	1	0.76	0.89	0.82		0.87	0.72	0.79	
NB	0	0.55	0.65	0.59	57	0.56	0.82	0.67	60
	1	0.60	0.50	0.55		0.70	0.39	0.50	
ANN	0	0.55	0.65	0.59	57	0.56	0.82	0.67	60
	1	0.60	0.50	0.55		0.70	0.39	0.50	
SVM	0	0.75	0.53	0.62	69	0.73	0.65	0.69	71
	1	0.65	0.83	0.73		0.70	0.78	0.74	
RF	0	0.75	0.53	0.62	69	0.73	0.65	0.69	71
	1	0.65	0.83	0.73		0.70	0.78	0.74	
DT	0	0.75	0.53	0.62	69	0.73	0.65	0.69	71
	1	0.65	0.83	0.73		0.70	0.78	0.74	
SGD	0	0.75	0.53	0.62	69	0.73	0.65	0.69	71
	1	0.65	0.83	0.73		0.70	0.78	0.74	

except the KNN give better results after using ANOVA. KNN classifier gives highest accuracy which is 80% and it remain same in both cases with or without feature selection.

Table 3 shows the performance of Wisconsin Breast Cancer dataset using all models. Logistic regression gives best result with 99% accuracy using ANOVA feature selection method. Here, only Naïve Bayes, logistic regression, and ANN classifiers improve their accuracies after using feature selection. Table4 showing the results of applied classifiers on prostate cancer dataset. Highest accuracy 97% is computed by five classifiers (NB, SVM, RF, DT, SGD). But the only difference is that Naïve Bayes gives best result without using feature selection and remaining classifiers gives their best accuracies after using ANOVA feature selection technique.

Figure 2 showing the learning curves of classifiers who gives highest accuracy in each dataset. In Fig. 2, curve (a) is showing the performance of KNN on Breast Cancer Coimbra dataset, curve (b) is showing the learning curve of logistic regression on Wisconsin Breast Cancer dataset, and curve (c) showing the results of support vector machine model on prostate cancer.

Table 3 Performance analysis of WBC

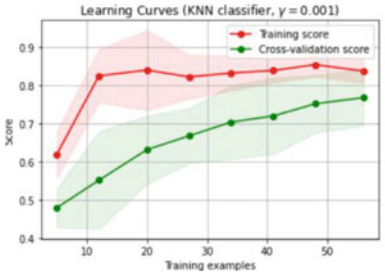
Models		Without feature selection				With feature selection			
		Precision	Recall	F1-score	Acc (%)	Precision	Recall	F1-score	Acc (%)
LR	0	0.97	0.99	0.98	98	0.99	0.99	0.99	99
	1	0.98	0.95	0.97		0.98	0.98	0.98	
KNN	0	0.95	0.99	0.97	96	0.96	0.96	0.96	0.95
	1	0.98	0.90	0.94		0.94	0.94	0.94	
NB	0	0.94	0.92	0.93	91	0.94	0.93	0.93	92
	1	0.86	0.90	0.88		0.88	0.90	0.89	
ANN	0	0.94	0.92	0.93	91	0.94	0.93	0.93	92
	1	0.86	0.90	0.88		0.88	0.90	0.89	
SVM	0	0.97	0.99	0.98	98	0.98	0.97	0.98	97
	1	0.98	0.95	0.97		0.95	0.97	0.96	
RF	0	0.97	0.99	0.98	98	0.98	0.97	0.98	97
	1	0.98	0.95	0.97		0.95	0.97	0.96	
DT	0	0.97	0.99	0.98	98	0.98	0.97	0.98	97
	1	0.98	0.95	0.97		0.95	0.97	0.96	
SGD	0	0.97	0.99	0.98	98	0.98	0.97	0.98	97
	1	0.98	0.95	0.97		0.95	0.97	0.96	

5 Conclusion

This work covers two main cancer types breast cancer (in females) and prostate cancer (in males) which are most dangerous and increase the mortality rate in whole world. It is very necessary to predict these diseases in their early stage for better treatment of patient. For early and correct predictions, all classification models are analyzed on each dataset. For improving the performance of models, firstly Z-score normalization method is used and analyze all the measuring parameters such as precision, recall, F1-score, and accuracy with or without using feature selection technique. The future anticipates the use of the aforementioned strategies to eliminate existing shortcomings and improve prediction rates, so giving a way to improve the survival rate for the well-being of mankind.

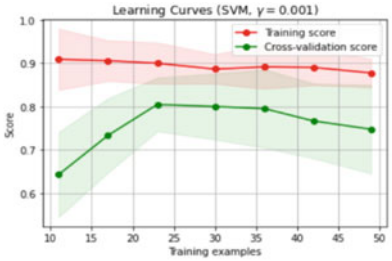
Table 4 Performance analysis of prostate cancer

Models		Without feature selection				With feature selection			
		Precision	Recall	F1-score	Acc (%)	Precision	Recall	F1-score	Acc (%)
LR	0	0.86	1.00	0.92	97	0.67	1.00	0.80	90
	1	1.00	0.96	0.98		1.00	0.89	0.93	
KNN	0	0.50	0.67	0.57	80	0.71	0.83	0.77	90
	1	0.91	0.83	0.87		0.96	0.92	0.94	
NB	0	0.45	0.83	0.59	77	0.45	0.83	0.59	77
	1	0.95	0.75	0.84		0.95	0.75	0.84	
ANN	0	0.45	0.83	0.59	77	0.45	0.83	0.59	77
	1	0.95	0.75	0.84		0.95	0.75	0.84	
SVM	0	0.57	0.67	0.62	83	1.00	0.83	0.91	97
	1	0.91	0.88	0.89		0.96	1.00	0.98	
RF	0	0.57	0.67	0.62	83	1.00	0.83	0.91	97
	1	0.91	0.88	0.89		0.96	1.00	0.98	
DT	0	0.57	0.67	0.62	83	1.00	0.83	0.91	97
	1	0.91	0.88	0.89		0.96	1.00	0.98	
SGD	0	0.57	0.67	0.62	83	1.00	0.83	0.91	97
	1	0.91	0.88	0.89		0.96	1.00	0.98	



(a) Model:KNN,Dataset:BCC

(b) Model:LR,Dataset:WBC



(c) Model: SVM, Dataset: Prostate Cancer

Fig. 2 Learning curves of models having best accuracy in every dataset

References

1. Prabadevi B, Deepa N, Krithika LB, Vani V (2020) Analysis of machine learning algorithms on cancer dataset. In: 2020 international conference on emerging trends in information technology and engineering (ic-ETITE), IEEE, pp 1–10
2. Asri H, Mousannif H, Al Moatassime H, Noel T (2016) Using machine learning algorithms for breast cancer risk prediction and diagnosis. Proc Comput Sci 83:1064–1069
3. Reda I, Ayinde BO, Elmogy M, Shalaby A, El-Melegy M, Abou El-Ghar M, Abou El-fetouh A, Ghazal M, El-Baz A (2018) A new CNN-based system for early diagnosis of prostate cancer. In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018), IEEE, pp 207–210
4. Rahman MM, Ghasemi Y, Suley E, Zhou Y, Wang S, Rogers J (2021) Machine learning based computer aided diagnosis of breast cancer utilizing anthropometric and clinical features. Irbm 42(4):215–226
5. Ray A, Ray H (2021) Performance analysis of machine learning classifiers on different healthcare datasets. In: Emerging technologies in data mining and information security, Springer, Singapore, pp 99–111
6. Mushtaq Z, Yaqub A, Hassan A, Feng Su S (2019) Performance analysis of supervised classifiers using PCA based techniques on breast cancer. In: 2019 International conference on engineering and emerging technologies (ICEET), IEEE, pp 1–6
7. Shakeel PM, Manogaran G (2020) Prostate cancer classification from prostate biomedical data using ant rough set algorithm with radial trained extreme learning neural network. Health Technol 10(1):157–165
8. Smita EK (2021) Probabilistic decision support system using machine learning techniques : a case study of Cardiovascular diseases. J Disc Math Sci Cryptogr (JDMC) 1487–1496
9. Doja MN, Kaur I, Ahmad T (2020) Age-specific survival in prostate cancer using machine learning. Data Technol Appl
10. Masood S, Luthra T, Sundriyal H, Ahmed M (2017) Identification of diabetic retinopathy in eye images using transfer learning. In: 2017 International conference on computing, communication and automation (ICCCA), IEEE, pp 1183–1187
11. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI (2015) Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 13:8–17
12. Turgut S, Dağtekin M, Ensari T (2018) Microarray breast cancer data classification using machine learning methods. In: 2018 electric electronics, computer science, biomedical engineerings' meeting (EBBT), IEEE, pp 1–3
13. Gao K, Wang D, Huang Y (2018) Cross-cancer prediction: a novel machine learning approach to discover molecular targets for development of treatments for multiple cancers. Cancer Informat 17:1176935118805398
14. Polat K, Sentürk U (2018) A novel ML approach to prediction of breast cancer: combining of mad normalization, KMC based feature weighting and AdaBoostM1 classifier. In: 2018 2nd International symposium on multidisciplinary studies and innovative technologies (ISMSIT), IEEE, pp 1–4
15. Mehdi M, Pahwa K, Sharma B (2019) Comparison of data mining algorithms for predicting the cancer disease using python. In: 2019 8th International conference system modeling and advancement in research trends (SMART), IEEE, pp 155–160