Performance Analysis of Breast Cancer Data Using Mann–Whitney U Test and Machine Learning

Priyanka Khanna and Mridu Sahu

Abstract Breast cancer is the most prevalent cause of mortality among women worldwide. Early detection and treatment can reduce the rate of mortality. Late prognosis and treatment of breast cancer (BC) patients lead to irreparable diseases and even death. As a result, in recent years, early BC diagnosis methods based on pathological breast images have been in high demand. In recent years, various models have been put up by the researcher for the early diagnosis of breast cancer. In this article, Wisconsin breast cancer (diagnostic) dataset (WDBC) is employed to categorize tumors into benign or malignant. Statistical-based Mann–Whitney U Test is applied for feature selection, followed by machine learning models for the classification of tumors. We compare two methods: a machine learning method with feature selection and one without. Finally, the results demonstrate that on selecting pertinent features, enhances the overall performance when tested on the WDBC dataset. The classification accuracy, sensitivity, and specificity obtained were 97.2%, 98.8%, and 94.5% using Random Forest with feature selection.

Keywords Breast cancer · Mann–Whitney U test · Machine learning · Feature selection · Accuracy

1 Introduction and Related Work

Cancer refers to the uncontrolled growth of certain cells in the human body [1]. These cells can spread into the surrounding tissue forming a lump known as tumor or malignancy [2]. After lung cancer, the second most common malignancies and reason of mortality for women worldwide are breast cancer [3]. Breast cancer (BC) is a frequently observed cancer in females of childbearing age. Breast cancer is the

P. Khanna (✉) · M. Sahu
National Institute of Technology, Raipur, India
e-mail: pkhanna.phd2019.it@nitrr.ac.in

M. Sahu
e-mail: mrisahu.it@nitrr.ac.in

V. V. S. S. S. Chakravarthy et al. (eds.), *Advances in Signal Processing, Embedded Systems and IoT*, Lecture Notes in Electrical Engineering 992,
https://doi.org/10.1007/978-981-19-8865-3_26

prevalent diagnosed cancer and is increasing every year very rapidly [4, 5]. According to the changes in the environment, the nature of the breast cancer is also changing day by day [6]. As a result, raising awareness of the benefits of screening and early detection is desirable. Ultrasound (US), mammography, contrast-enhanced (CE), breast tomosynthesis (3D mammography), magnetic resonance imaging (MRI), computed tomography (CT), and positron emission tomography (PET) are the currently used clinical practices for the early diagnosis of BC. These methods are used to examine significant parameters such as the size, shape, location, type of cancer, stage of cancer, or how quickly it is growing. These methods are sometimes combined for a more accurate prognosis.

The most crucial and significant task is classification. Multiple classifiers and feature selection strategies are used in numerous research on datasets related to breast cancer [7]. Different machine learning (ML) classifiers have been developed for classification and employed on medical datasets. Machine learning is a subset of artificial intelligence within the realm of computing. ML is not only confined to computer science, but also extended to many other branches.

1.1 Based on Breast Cancer

Sengar [8] compared machine learning algorithms like Logistic Regression, Decision Tree on taken dataset. Decision Tree reported maximum classification accuracy of 95%. The main limitation of this work is that only two classifiers are evaluated. Anji Reddy Vaka [9] used deep neural networks on collected dataset. They collected data from Mahatma Gandhi Cancer Hospital and Research Institute, Visakhapatnam, India. As the dataset is limited, data augmentation is done to enlarge the dataset. Gaussian filtering is used for removal of noise as preprocessing step, and neglected values are removed using entropy followed by different ML algorithms for classification. Deep neural network reported highest accuracy of 97.01%.

Moh'd Rasoul Al-hadidi [10] used radiography images, and all the images are of equal size, thus making processing easier. Weiner filter is used to remove the image blurriness followed by Logistic Regression and back-propagation neural network. Back-propagation network attained maximum accuracy of 93%. Bazazeh [11] used WBCD dataset to train the model. Different machine learning algorithms like support vector machine, Random Forest, and Bayesian networks are used for evaluation.

Sadhukhan [12] converted images into fine-needle aspiration images which are further converted into grayscale images by removing hue from the images. For segmentation, thresholding is used and radii, smoothness, compactness, texture are calculated. Adel [13] cropped images to separate B-mode images amid elastography images. Different features like signal-to-noise ratio, width-to-height ratio, area, difference, perimeter difference, solidity, contrast-to-noise ratio, and compactness were extracted. Further, dimensionality reduction is done and input is fed to support

vector machine achieving an accuracy of 94.12%. Kaklamanis [14] applied correlation matrix for feature selection. Further, CART, KNN, Naive Bayes, and SVM are used for classification reporting accuracy of 93%, 96%, 89%, and 96%, respectively.

1.2 Based on Feature Selection

Perez [15] used two datasets of breast mammography images. Features were selected using Mann–Whitney U Test and selected feature subset is fed as an input to feedforward back-propagation network. MacFarland [16] emphasized on Mann–Whitney U Test, and it is generally conducted on non-parametric and independent values. It was first started by testing on goats, and two groups of goats in a total of 30 were taken, in which one group received mineral supplement included in the diet, whereas the other group is supplied with normal meal. At the end of the treatment, mineral supplement supplied goats shown to be healthier than the other group. Some facts like details about mineral supplement, how it is added to the meal, cost, and treatment regulation are not disclosed.

1.3 Based on Machine Learning

Bhavsar [17] evaluated different machine learning classifiers, namely support vector machine, Decision Tree, Supervised Learning, and Nearest Neighbor Neural Network. Performance metrics' accuracy, specificity, sensitivity were evaluated. Morgan [18] evaluated the performance using Gaussian process and Gaussian kernel ridge regression. For selecting pertinent features, Leave-Group-Out cross-validation root mean squared error is used. Fatima [19] provides comparative analysis of different machine learning algorithms for prognosis of different diseases. It emphasizes the use of machine learning algorithms for the analysis of disease and its decision-making.

2 Material and Method

2.1 Material

In this article, Wisconsin breast cancer (diagnostic) dataset (WDBC) [20] collected from UCI repository is used to differentiate benign from malignant sample. WDBC has 32 attributes and 569 instances, 357 of which are benign and 212 are malignant. Fine-needle aspirate (FNA) digitized picture was used to calculate features. These features exhibit ten characteristics of each cell nucleus. Excluding ID and diagnosis,

Table 1 Description of the Wisconsin breast cancer dataset (WDBC)

S. No.	Attribute	Description
1	ID	Id number
2	Diagnosis	Diagnosis (b = benign, m = malignant)
3	Radius	The average distance separation between the center and edge points
4	Texture	Standard deviation of values in gray scale
5	Perimeter	Tumor mean size
6	Area	Tumor mean area
7	Smoothness	Mean of local length variation
8	Compactness	Mean of perimeter$^{2/\text{area}} - 1$
9	Concavity	Severity of a contour's mean concave portions
10	Concave points	Mean of the concave points on the contour
11	Fractal	Mean of coastline approximation -1

for each attribute, mean, standard error, and "worst" or largest (mean of the three largest values) are computed. There are no missing data in the dataset. Table 1 shows the description of the WDBC dataset features.

2.2 Method

In this article, we describe a feature selection method for WDBC dataset diagnostic that uses the Mann–Whitney U Test followed by different machine learning classifiers for classification. Firstly, WDBC dataset is taken and unwanted columns are removed as preprocessing step. Secondly, to improve the classification accuracy, feature selection using Mann–Whitney U Test is performed to choose relevant features. To categorize tumor as benign or malignant, selected features are finally fed via machine learning classifiers. Figure 1 demonstrates a proposed method for classifying breast tumor.

The assessment is conducted on the above datasets with and without feature selection method. And, the results are compared and analyzed. Evaluation metrics' sensitivity, specificity, and accuracy are calculated to access different machine learning classifiers. Experimental simulations were conducted using Jupyter Notebook.

2.2.1 Preprocessing

The first and most significant step is preprocessing, which enhances image quality while retaining key elements. Incorrect conclusions can be drawn from radiological images due to artifacts, noise, and other factors. The dataset consists of some unwanted columns which need to be removed for better result. No missing values are

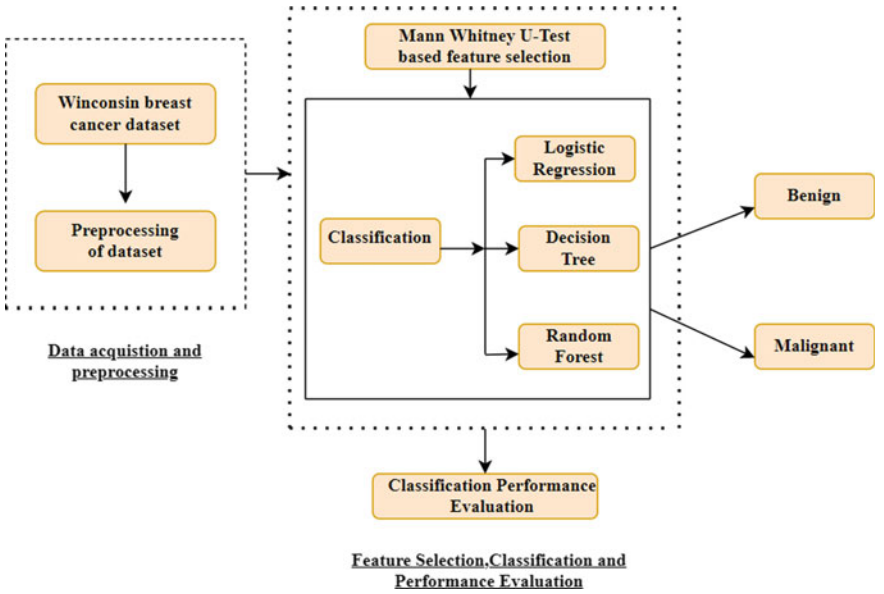

Fig. 1 Proposed methodology for classifying breast tumor

found in this dataset values. The categorical data diagnosis is changed to numerical data for compactness with the Mann–Whitney U Test [21].

2.2.2 Feature Selection

Following preprocessing, we carried out feature selection to select relevant features because they have a direct impact on classifier performance. The size of the feature space and computation time are reduced by removing redundant features. Gain ratio, recursive feature removal, Random Forest, Chi-square test, and searching algorithms are a few techniques frequently used for feature selection. In order to have effective prediction and computationally less costly models, the number of input classifier is limited. Mann–Whitney U Test is used as a feature selection technique in this paper. Mann–Whitney U Test is a statistical method used for non-uniformly distributed data.

In this test, calculation of U is done whose distribution under the null hypothesis is known. The normality of data was verified by U test, results obtained have a significant value less than 0.001 ($p < 0.001$), and 95% confidence interval (CI) marked those features was not normally distributed (non-parametric) [21]. A feature vector with 32 features (F [1], F [2], …., F [32]) is provided as input, and further, 26 features are chosen using the Mann–Whitney U Test.

2.2.3 Classification

Following feature selection, the classifier uses the pertinent features to categorize breast tumor as benign or malignant. Any automated system's classifier plots feature space as input to produce class labels [22]. The Naive Bayesian (NB), Decision Tree (DT), K-Nearest Neighbor (KNN), support vector machine (SVM), Random Forest, and Logistic Regression are examples of commonly used machine learning classifiers. Random Forest, Logistic Regression, and Decision Tree are evaluated in this study.

Logistic Regression transforms the linear regression model to allow us to probabilistically model the binary variables in consequence. A supervised procedure called Logistic Regression is used to predict the likelihood of a target variable. There are only two useful classes because the goal's or established variable's personality is binary. The established variable is binary in nature, with records encoded as 1 or 0. $P(Y = 1)$ is predicted by the Logistic Regression version as a function of X [8]. Decision Tree is a popular and unsupervised approach used for classification and prediction [23]. It is represented as a recursive partition of the instance, where leaves represent the class labels and branches refer to outcome in the form of features. It is a top-down approach which divides each result of the data into subsets. This predictive paradigm acts as a mapping between the item's qualities and values. Random Forest (RF) algorithm is based on multiple Decision Trees which is merged to produce an accurate and stable prediction [24]. RF is an ensemble of classifiers grown from a certain amount of randomness. RF stands for randomized ensembles of Decision Trees and is defined as a generic principle. Every observation is input into every Decision Tree. The final result is the most common outcome for each observation.

3 Experimental Results and Discussion

This section discusses the findings of the proposed method and compares them with the other related work. Experimental simulations were conducted using Jupyter Notebook. On the WDBC dataset, simulations were used to categorize the breast tumor as benign or malignant. The proposed method employed Mann–Whitney U Test for feature selection, using Statistical Package for Social Sciences (SPSS) software with 95% confidence interval, and the significance level was chosen to be less than 0.001. The values shown in Table 2 are the asymptotic significance values obtained on conducting Mann–Whitney U Test (non-parametric test). If asymptotic significance is greater than 0.001, then the features will be eliminated. The benign and malignant values, which are in categorical form, are converted into ordinal form. Out of 30 features, 26 features are selected based on U test and four features are eliminated. Selected features are further passed through classifier, and result is evaluated with and without feature selection.

Further features selected are fed as an input to Random Forest, Logistic Regression, and Decision Tree classifier. The dataset is split into sections: testing and

Table 2 Statistical analysis using Mann–Whitney U Test

Feature	Asymptotic significance
Perimeter worst	<0.001
Texture worst	<0.001
Radius worst	<0.001
Fractal dimension se	<0.001
Symmetry se	**0.028**
Concave points se	<0.001
Concavity se	<0.001
compactness se	<0.001
Smoothness se	**0.214**
Area se	<0.001
Perimeter se	<0.001
Texture se	**0.644**
Radius se	<0.001
Fractal dimension mean	**0.537**
Symmetry mean	<0.001
Concave points mean	<0.001
Concavity mean	<0.001
Compactness mean	<0.001
Smoothness mean	<0.001
Area mean	<0.001
Perimeter mean	<0.001
Texture mean	<0.001
Radius mean	<0.001
Area worst	<0.001
Smoothness worst	<0.001
Compactness worst	<0.001
Concavity worst	<0.001
Concave points worst	<0.001
Symmetry worst	<0.001
Fractal dimension worst	<0.001

training, under K-fold cross-validation protocol. Value of $k = 10$ is taken to compute the performance of the system. Table 3 shows evaluation measures used to evaluate the classifiers' sensitivity, specificity, and classification accuracy. True positive (tp)represents the quantity of patients who have been correctly classified, while the number of patients who have been correctly classified as negative class is represented by true negative (tn). False positive (fp)represents the number of incorrectly predicted patients, whereas false negative (fn)indicates the number of incorrectly predicted patients.

Table 3 Performance measure

Performance measures	Definition
Classification accuracy	$\frac{tp + tn}{tp + fn + tn + fp} \times 100$
Sensitivity	$\frac{tp}{tp + fn} \times 100$
Specificity	$\frac{tn}{tn + fp} \times 100$

Table 4 shows the experimental result in terms of accuracy, specificity, and sensitivity obtained by applying ML algorithms with feature selection using Mann–Whitney U Test. And, Table 5 shows the result obtained without feature selection. As shown in Tables 4 and 5, performance measure under ten-fold cross-validation is evaluated. It is observed that employing Random Forest as a classifier increases accuracy on selecting features. Accuracy of 99.5%, sensitivity of 98.8%, and specificity of 94.5% are obtained.

A comparative analysis of the proposed technique with prior relevant work on the WDBC dataset is shown in Table 6. Accuracy rate of 97.2% for the proposed method was obtained. Asri et al. [25] used C4.5, SVM, NB, and KNN for classification. Saravana Kumar et al. [26] proposed multi-layer perceptron based on deep learning. Performance comparison of proposed work with aforementioned related work is mentioned in Table 6.

Table 4 Performance metric obtained with feature selection

Classification technique	Accuracy (%)	Sensitivity (%)	Specificity (%)
Logistic regression	96.5	97.7	94.4
Decision free	94.4	97.6	89.4
Random forest	97.2	98.8	94.5

Table 5 Performance metric obtained without feature selection

Classification technique	Accuracy (%)	Sensitivity (%)	Specificity (%)
Logistic regression	95.1	96.6	92.5
Decision tree	93.7	97.6	87.9
Random forest	96.5	97.7	94.4

Table 6 Comparison with related work

Dataset	Method	Accuracy (%)
Wisconsin dataset [25]	C4.5	95.13
	SVM	97.13
	NB	95.99
	KNN	95.27
Wisconsin dataset [26]	MLP	97
Wisconsin dataset	**Proposed method**	**97.2**

4 Conclusion and Future Scope

This study offered a thorough methodology for ultrasound-based breast cancer diagnosis. The study's primary contributions are as follows: Firstly, the WBCD dataset were taken and some unwanted columns were removed for better result. Secondly, to pick relevant features, an effective statistical approach Mann–Whitney U Test was used. Thirdly, features are trained using different machine learning classifiers to differentiate class labels. For future scope of this work, we plan to use a substantial dataset to test our proposed study and also to use data augmentation for increasing data size and optimization techniques for feature selection. In conclusion, the potential for the proposed technique to classify breast tumors is apparent, though better optimization techniques and big datasets are still needed.

References

1. James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning, vol 112. Springer, New York, p 18
2. Huang MW, Chen CW, Lin WC, Ke SW, Tsai CF (2017) SVM and SVM ensembles in breast cancer prediction. PLoS ONE 12(1):e0161501
3. Moon WK, Huang CS, Shen WC, Takada E, Chang RF, Joe J, Nakajima M, Kobayashi M (2009) Analysis of elastographic and B-mode features at sonoelastography for breast tumor classification. Ultrasound Med Biol 35(11):1794–1802
4. Moon WK, Huang YS, Lee YW, Chang SC, Lo CM, Yang MC, Bae MS, Lee SH, Chang JM, Huang CS, Lin YT (2017) Computer-aided tumor diagnosis using shear wave breast elastography. Ultrasonics 78:125–133
5. Okagbue HI, Adamu PI, Oguntunde PE, Obasi EC, Odetunmibi OA (2021) Machine learning prediction of breast cancer survival using age, sex, length of stay, mode of diagnosis and location of cancer. Health Technol 1–7
6. Amrane M, Oukid S, Gagaoua I, Ensari T (2018) Breast cancer classification using machine learning. In: 2018 electric electronics, computer science, biomedical engineering's meeting (EBBT), IEEE, pp 1–4
7. Mušić L, Gabeljić N (2019) Predicting the severity of a mammographic tumor using an artificial neural network. In: International conference on medical and biological engineering. Springer, Cham, pp 775–778

8. Sengar PP, Gaikwad MJ, Nagdive AS (2020) Comparative study of machine learning algorithms for breast cancer prediction. In: 2020 Third international conference on smart systems and inventive technology (ICSSIT). IEEE, pp 796–801
9. Vaka AR, Soni B, Reddy S (2020) Breast cancer detection by leveraging machine learning. ICT Express 6(4):320–324
10. Alarabeyyat A, Alhanahnah M (2016) Breast cancer detection using k-nearest neighbor machine learning algorithm. In: 2016 9th international conference on developments in esystems engineering (DeSE). IEEE, pp 35–39
11. Bazazeh D, Shubair R (2016) Comparative study of machine learning algorithms for breast cancer detection and diagnosis. In: 2016 5th international conference on electronic devices, systems and applications (ICEDSA). IEEE, pp 1–4
12. Sadhukhan S, Upadhyay N, Chakraborty P (2020) Breast cancer diagnosis using image processing and machine learning. In: Emerging technology in modelling and graphics. Springer, Singapore, pp 113–127
13. Adel M, Kotb A, Farag O, Darweesh MS, Mostafa H (2019) Breast cancer diagnosis using image processing and machine learning for elastography images. In: 2019 8th international conference on modern circuits and systems technologies (MOCAST). IEEE, pp 1–4
14. Kaklamanis MM, Filippakis ME (2019) A comparative survey of machine learning classification algorithms for breast cancer detection. In: Proceedings of the 23rd panhellenic conference on informatics, pp. 97–103
15. Pérez NP, López MAG, Silva A, Ramos I (2015) Improving the Mann-Whitney statistical test for feature selection: an approach in breast cancer diagnosis on mammography. Artif Intell Med 63(1):19–31
16. MacFarland TW, Yates JM (2016) Mann–whitney u test. In Introduction to nonparametric statistics for the biological sciences using R. Springer, Cham, pp 103–132
17. Bhavsar H, Ganatra A (2012) A comparative study of training algorithms for supervised machine learning. Int J Soft Comput Eng (IJSCE) 2(4):2231–2307
18. Lu HJ, Zou N, Jacobs R, Afflerbach B, Lu XG, Morgan D (2019) Error assessment and optimal cross-validation approaches in machine learning applied to impurity diffusion. Comput Mater Sci 169:109075
19. Fatima M, Pasha M (2017) Survey of machine learning algorithms for disease diagnostic. J Intell Learn Syst Appl 9(01):1
20. Salama GI, Abdelhalim M, Zeid MAE (2012) Breast cancer diagnosis on three different datasets using multi-classifiers. Breast Cancer (WDBC) 32(569):2
21. Kirk R (2007) Statistics: an introduction. Nelson Education
22. Singh BK (2019) Determining relevant biomarkers for prediction of breast cancer using anthropometric and clinical features: a comparative investigation in machine learning paradigm. Biocybernetics Biomed. Eng. 39(2):393–409
23. De Mántaras RL (1991) A distance-based attribute selection measure for decision tree induction. Mach Learn 6:81–92
24. Breiman L (2001) Random forests. Mach Learn J Paper 45:5–32
25. Asri H, Mousannif H, Al H, Noel T (2016) Using machine learning algorithms for breast cancer risk prediction and diagnosis. Proc Comput Sci 83:1064–1069
26. NM SK, Tamilselvi S, Hariprasath K, Kaviyavarshini N, Kavinya A (2022) An efficient multilayer perceptron neural network-based breast cancer prediction. In: Principles and methods of explainable artificial intelligence in healthcare. IGI Global, pp 211–231