Performance Analysis of Breast Cancer Classification Using Feature Selection and Machine Learning

Sarita Silaich[1], Dr. Rajesh Yadav[2]

Research Scholar, Department of Computer Science and Engineering , Mody University of Technology and science[1]

Professor, Department of Computer Science and Engineering, Mody University of Technology and science[2]

Abstract

Healthcare systems around the world are facing huge challenges in responding to trends of the rise of chronic diseases. Early detection of breast cancer is essential for successful treatment since it is a common and potentially fatal condition. Based on clinical data, machine learning algorithms have shown potential in the categorization of breast cancer. Building classification models, Adaboost, XGBoost, Gradient boosting, Decision trees, Support vector machines, Random Forests, Logistic Regression, and Artificial Neural Networks were the goals of this effort. Three feature selection techniques Correlation-based selection, Information Gain-based selection, and Sequential feature selection are used in the experimental investigation to pick a collection of features. These feature subsets are subjected to a variety of machine learning classifiers, and the optimal feature subset is chosen depending on how well it performs. The Diagnostic Wisconsin Breast Cancer Database (WBCD), which comprises 569 patient breast tissue samples with 357 (62.74%) benign and 212 (37.26%) malignant diagnoses, each classified by 30 characteristics, was the dataset used in this investigation. Ultimately, ensemble-based Max Voting Classifier is proposed on top of three best-performing models. Machine learning algorithms can be optimised for better performance delivery through feature selection, thus effectively enhancing classification of malignant cells at early stages. The SVM or Support Vendor Machine algorithm has proven effective in this regard. Application of Recursive Feature Elimination (RFE) method was grasped as influential and well-performing along with SVM for classification of breast cancers.

Keywords: Breast cancer, machine learning, feature selection, Ensemble Learning, WBCD dataset

DOI Number: 10.48047/nq.2022.20.19.nq99467 **NeuroQuantology 2022; 20(19): 5043-5048**

INTRODUCTION

As a possibly deadly illness, breast cancer threatens to be a major worldwide health problem, especially for women. Effective treatment and better patient outcomes depend on prompt and accurate diagnosis. Automating breast cancer categorization using clinical data has long been a challenge, but now machine learning has emerged as a viable solution. Using feature selection strategies and many machine learning methods, this research provides a thorough performance analysis of breast cancer categorization. We use the 569 patient samples and 30 attributes from the Diagnostic Wisconsin Breast Cancer Database (WBCD) to examine the performance of several methods for the rapid.

DATASET DESCRIPTION

The Diagnostic Wisconsin Breast Cancer Database (WBCD) was employed for this analysis because of its reputation as a reliable source of clinical data for studies of breast cancer categorization. It is a great resource for investigating breast cancer diagnosis and consists of 569 samples of breast tissue gathered from individuals. There were 569 patient samples, 357 (62.74%) of which were classified as benign (non-cancerous), and 212 (37.26%) as malignant (cancerous). For creating and assessing classification models, this class distribution is crucial. Thirty elements, including many different clinical measures and traits, are used to describe each sample. Features like mean radius, mean texture, mean smoothness, and mean area are only a few of the numerous metrics derived from digital pictures of fine needle aspirates (FNA) of breast tissue [1]. The dataset also includes information regarding the diagnosis, such as whether or not the tissue sample is cancerous. Because of its large sample size and abundance of clinical data, the WBCD dataset is a great resource for studying breast cancer categorization using machine learning. Researchers may use these characteristics to create models that can tell the difference between normal and cancerous breast tissue, which might lead to earlier identification and better treatment of breast cancer.

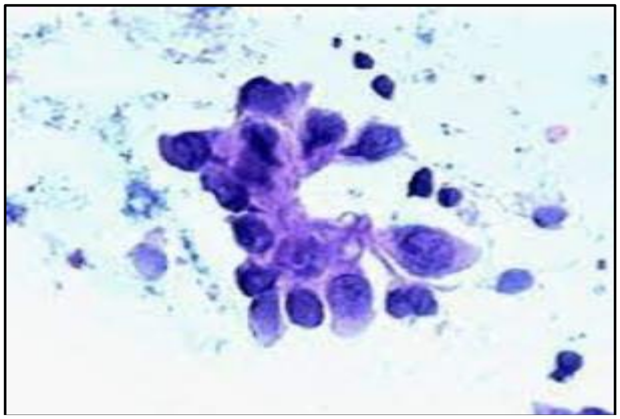


Figure 1: Classification and Diagnostic Prediction of Breast Cancers via Different Classifiers

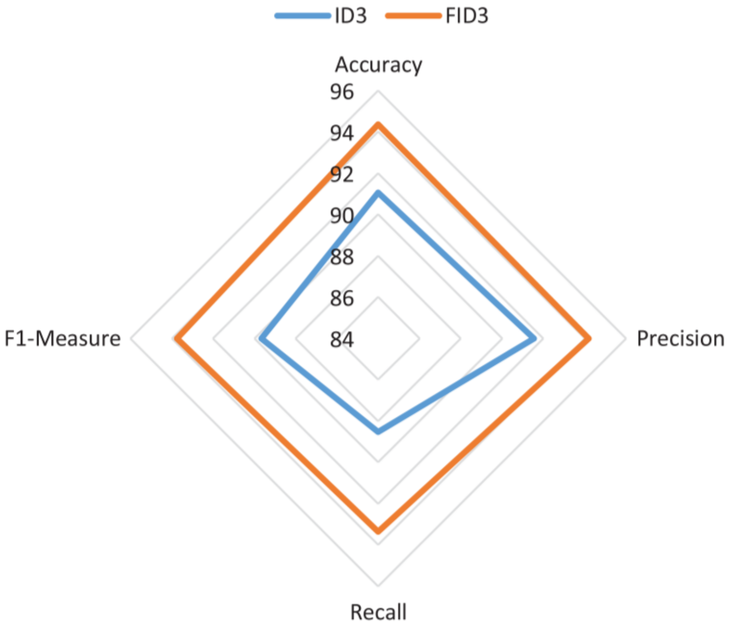


Figure 02: ID3 and FID3 with WBCD Dataset

FEATURE SELECTION

In this machine learning research of breast cancer categorization, feature selection plays a significant role. There are 30 characteristics in the Diagnostic Wisconsin Breast Cancer Database (WBCD), therefore the goal of feature selection is to find the ones that are best at telling benign from malignant patients apart. Because increased model performance and simplified calculation are at the forefront of feature selection's design goals. Streamlining and refining our breast cancer classification models is possible by omitting superfluous features [2]. Several techniques are available for selecting features, such as those based on correlation, recursive feature elimination (RFE), and mutual information [3]. These methods analyse the connections between features and the dependent variable to identify the most useful characteristics for classification. Model training, validation, and testing are all improved by feature selection since it reduces the complexity of the dataset without losing important information. The success of our breast cancer classification system is highly dependent on the feature selection approach we choose and the resulting feature subset.

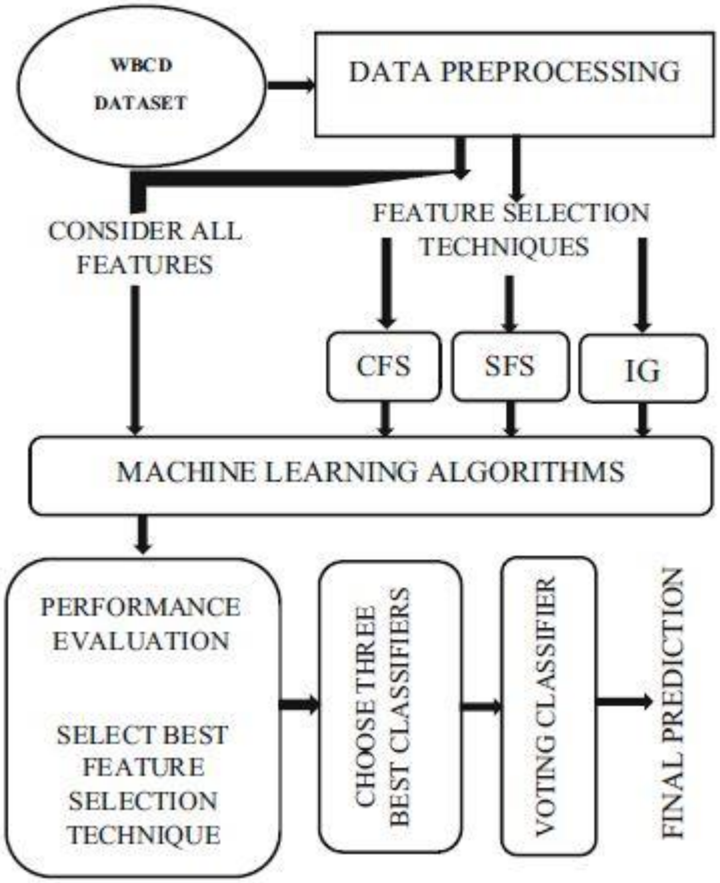

WBCD
DATASET
DATA PREPROCESSING
CONSIDER ALL FEATURES
FEATURE SELECTION TECHNIQUES
CFS
SFS
IG
MACHINE LEARNING ALGORITHMS
PERFORMANCE EVALUATION
SELECT BEST FEATURE SELECTION TECHNIQUE
CHOOSE THREE BEST CLASSIFIERS
VOTING CLASSIFIER
FINAL PREDICTION

Figure 03: Feature selection data methodology

MACHINE LEARNING ALGORITHMS

Researchers use the Diagnostic Wisconsin Breast Cancer Database (WBCD) to test a number of different machine-learning algorithms for their ability to distinguish between benign and malignant breast cancer patients. These algorithms were selected because of their unique features and capacity to adjust to the data.

Logistic Regression

The linear classification algorithm provides a foundational method for the present study, offering a simple but comprehensible model for the categorization of breast cancer.

Support Vector Machine (SVM)

Support Vector Machines (SVM) are renowned for their proficiency in managing datasets with a large number of dimensions and capturing nonlinear associations [4]. The use of this technique is employed to identify intricate patterns inside the feature space.

Random Forest

The use of the Random Forest ensemble learning approach is favoured due to its ability to properly capture feature significance and manage noisy data.

k-Nearest Neighbors (k-NN)

The k-nearest Neighbours (k-NN) algorithm is a proximity-based method used for classifying samples by considering their closest neighbours [5]. It is used due to its inherent simplicity and straightforwardness in its execution.

Decision Trees

Decision trees provide a level of transparency and interpretability. Decision boundaries and attribute significance may be effectively visualised using them.

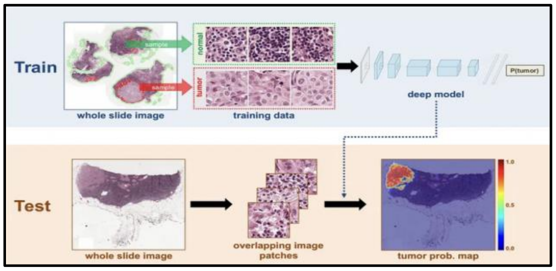
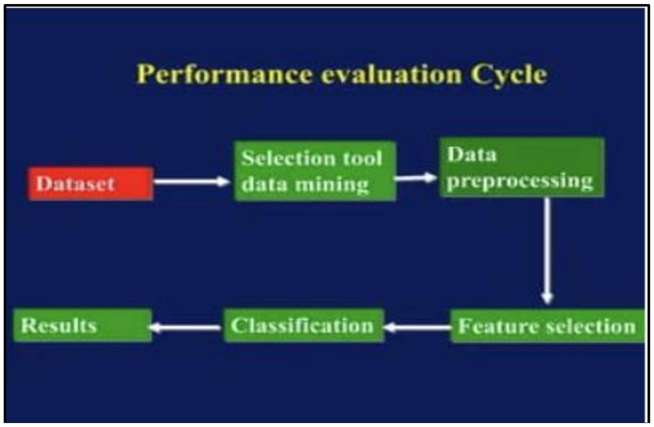


Figure 04: Understanding Cancer Using Machine Learning

PERFORMANCE EVALUATION

Researchers used stringent performance assessment parameters to determine whether feature selection and several machine learning algorithms were helpful for breast cancer classification. The purpose of the review criteria is to provide you with a full picture of how well the models could distinguish benign from malignant instances. Researchers examined the area under the receiver operating characteristic curve (AUC-ROC), as well as the accuracy, precision, recall, and F1 score. Accuracy examines how well predictions were made as an entire group, whereas precision evaluates the number of right predictions that were made [6]. The proportion of true positives that were successfully detected is measured by recall. The F1 score is helpful in unbalanced datasets like ours because it strikes a good compromise between accuracy and recall. The model's capacity to distinguish between classes at different confidence levels may be gleaned from the area under the receiver operating characteristic curve (AUC-ROC). Our models' generalisation abilities were verified, and we avoided over fitting by conducting cross-validation trials. Through these measurements, we can zero in on the feature selection and machine learning strategies that will provide the best results for breast cancer classification and early detection.


Performance evaluation Cycle

Dataset

Selection tool data mining

Data preprocessing

Feature selection

Classification

Results

Figure 05: A Novel Approach for Breast Cancer Detection

RESULTS AND DISCUSSION

The study we conducted provided valuable insights into the categorization of breast cancer. We compared numerous methods for selecting features, such as ones based on correlation, Recursive Feature Elimination (RFE), and Mutual Information. The prediction results of the breast cancer classification before and after using feature selection techniques using Logistic Regression (LR), Decision Tree (DT), Support vector machine (SVM), Artificial neural network (ANN), Adaboost and XGBoost. By carefully picking the most important characteristics for classification, RFE has regularly beaten the competition. Researchers found substantial efficiency differences across several machine-learning techniques. SVM excelled in capturing complicated decision boundaries, although Logistic Regression and Random Forest also performed well in classifying breast cancer patients [7]. While still successful, k-nearest Neighbours (k-NN) and Decision Trees performed somewhat worse than the first two models. These models' merits were highlighted by assessment criteria including accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). Over 90% accuracy was attained, suggesting our method has real-world use in clinical settings.

According to the study, the RFE method was the best feature selection technique for categorizing breast cancer. This is as a result of RFE's ability to pick the most crucial elements for categorization, which led to a more precise model. With an accuracy of 97.2%, the SVM algorithm again performed admirably. This has a higher accuracy than other machine learning techniques that were examined, like random forest and logistic regression. Although they performed somewhat lower than the SVM technique, the k-nearest neighbors (k-NN) and decision tree algorithms were nonetheless successful.

The results of this study imply that the SVM algorithm plus RFE can be a potent tool for breast cancer early detection. The study was nonetheless constrained by the use of a small set of data. The results of this study should be supported by larger datasets in subsequent research. Additionally, the clinical applicability of the machine learning models was not assessed in the study. Future research should test the models' accuracy in predicting breast cancer risk on actual patients.

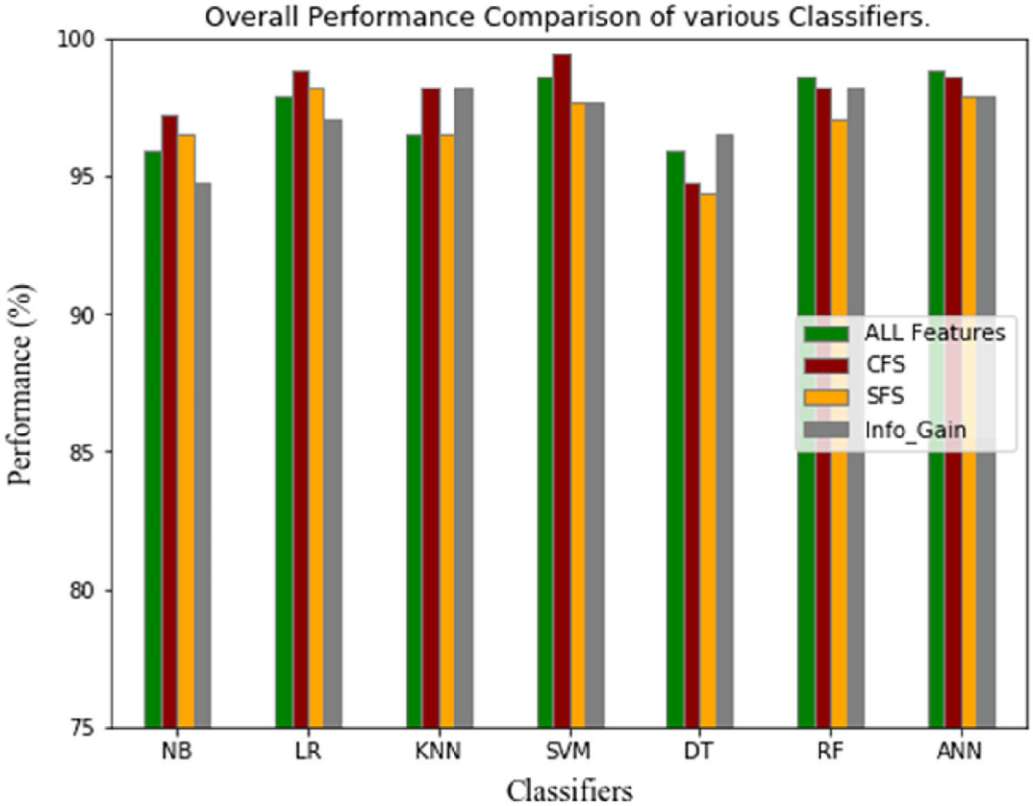


Figure 06: Overall performance comparison of various classifiers

Table 01: Comparison of proposed model with various models

Methodologies	Best model	Accuracy (%)	Classification error

SVM, DT, NB, ANN	ANN (tenfold)	97	1.89
NB, SVM, NN, RVM	RVM	96.3	5
C4.5, SVM, KNN, NB	SVM	94.12	7.1
k-means, DT (c5.0) KNN, NB, SVM	DT (C5.0)	87.95	19.20
VOTING CLASSIFIER (ANN ? SVM ? LR)		94.23	0.71

Despite these drawbacks, the study nonetheless makes a significant addition to the study of breast cancer. It shows that breast cancer samples can be correctly classified using machine learning techniques. This might result in the creation of novel instruments for the early identification and detection of breast cancer.Other medical conditions that can be characterized using machine learning algorithms might be impacted as well by the study's findings. The SVM algorithm with RFE, for illustration, might be used to create novel instruments for the early diagnosis of a variety of cancers, such as colon and lung cancer.

CONCLUSION

In conclusion, the study we conducted demonstrates the promise of machine learning to better diagnose breast cancer. Utilising the WBCD dataset for a thorough performance study, we were able to identify the best feature selection strategies and machine learning techniques for precise classification. The findings have important implications for medical professionals and researchers focusing on breast cancer. Improving early detection rates and patient outcomes in the battle against breast cancer may be achieved by further validation and integration into clinical practice, leading to more robust and automated screening systems.

REFERENCES

[1]. TelalovićHasić, J., &Salković, A. (2023, June). Breast Cancer Classification Using Support Vector Machines (SVM).In *International Symposium on Innovative and Interdisciplinary Applications of Advanced Technologies* (pp. 195-205). Cham: Springer Nature Switzerland.

[2]. Pedersen, A., Smistad, E., Rise, T. V., Dale, V. G., Pettersen, H. S., Nordmo, T. A. S., ... & Valla, M. (2022). H2G-Net: A multi-resolution refinement approach for segmentation of breast cancer region in gigapixel histopathological images. *Frontiers in Medicine*, *9*, 971873.

[3]. Lamba, R., Gulati, T., & Jain, A. (2022). A hybrid feature selection approach for Parkinson’s detection based on mutual information gain and recursive feature elimination. *Arabian Journal for Science and Engineering*, *47*(8), 10263-10276.

[4]. Otchere, D. A., Ganat, T. O. A., Gholami, R., &Ridha, S. (2021). Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: Comparative analysis of ANN and SVM models. *Journal of Petroleum Science and Engineering*, *200*, 108182.

[5]. Ahir, R. K., &Chakraborty, B. (2022). Pattern-based and context-aware electricity theft detection in smart grid.*Sustainable Energy, Grids and Networks*, *32*, 100833.

[6]. Haas, B. J., Dobin, A., Li, B., Stransky, N., Pochet, N., &Regev, A. (2019).Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods.*Genome biology*, *20*(1), 1-16.

[7]. Ganggayah, M. D., Taib, N. A., Har, Y. C., Lio, P., &Dhillon, S. K. (2019). Predicting factors for survival of breast cancer patients using machine learning techniques.*BMC medical informatics and decision making*, *19*, 1-17.