Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

Implementing Machine Learning Techniques on WDBC Datasets with Limited and Complete Features: A Comparative Analysis

Yogesh Kumar[1], Himani Choudhary[2]

[1]Assistant Professor, Computer Science & Engineering, School of Computer Science & Engineering, Dev Bhoomi Uttarakhand University, Chakrata Road, Manduwala, Naugaon, Uttarakhand 248007

[2]Assistant Professor, Management, School of Manamgnet & Commerce, Dev Bhoomi Uttarakhand University, Chakrata Road, Manduwala, Naugaon, Uttarakhand 248007

[1]socse.yogesh@dbuu.ac.in, [2]somc.himani@dbuu.ac.in

Article Info
Page Number: 2692-2700
Publication Issue:
Vol. 71 No. 4 (2022)

Article History
Article Received: 25 March 2022
Revised: 30 April 2022
Accepted: 15 June 2022
Publication: 19 August 2022

Abstract

Cancer is one of the most lethal illnesses. Although there are various types of cancer, breast cancer is the most common, particularly among women worldwide. It has been shown that detecting cancer at an early stage increases the odds of survival. This is related to the fact that treatment starts early. As a result, this region requires specific care. Machine learning technologies are gaining traction in the medical profession. Many hardware and software businesses have lately used machine learning methods to get high-quality solutions. In this paper, ML methods were used for the WDBC dataset. A comparison study is performed, demonstrating the differences in results achieved after applying the identical algorithms to WDBC datasets with limited and comprehensive characteristics. Comparison parameters like accuracy, f1- score, and recall are used to demonstrate the performance of the models.

Keywords: Machine Learning, accuracy, f1-score, and recall are all keywords.

1. INTRODUCTION

Machine learning has grown in popularity during the last decade (ML). The cheap cost of computing power and memory has spurred interest. Machine learning is employed in a variety of sectors, including construction, computer vision, healthcare, natural language processing, and education. The purpose of this study is to identify breast cancer. A suite of preoperative testing, including mammography, biopsies, and ultrasound, has assisted in the identification of breast cancer. In any event, the aforementioned diagnostic process has limitations of its own. Patients with breast cancer who are detected early have a lower risk of dying. As a consequence, breakthroughs in currently available approaches are important for breast cancer early detection.

Breast cancer is caused by a malignant cyst that occurs when the development of a cell becomes uncontrollable. Malignant cells spread throughout the body, producing cancer at different stages. Cancers develop when malignant cells and tissues circulate throughout the body. The following are the several forms of breast cancer:

- DCIS (Ductal Carcinoma in Situ): Breast cancer develops when abnormal cells grow outside of the breast.

• Invasive Ductal Carcinoma (IDC): This cancer develops when abnormal breast cells spread across the breast tissues. It is more prevalent in males.

• Mixed tumors breast cancer (MTBC) is an abbreviation for invasive mammary breast cancer. This cancer is caused by irregular duct and lobular cells.

• Lobular Breast Cancer (LBC): Lobular breast cancer is a kind of breast cancer that develops inside the lobule. As a consequence, other forms of invasive cancers are more prone to develop.

• Mucinous Breast Cancer (MBC): Mucinous Breast Cancer (MBC) is caused by invasive ductal cells and is an abbreviation for Colloid Breast Cancer (CBC). It happens when abnormal tissues spread across the duct.

This report outlines a method for identifying breast cancer. The study relies on the Wisconsin Diagnostic Breast Cancer Dataset (WDBC). The paper mentions machine learning methods such as Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Random Forest (RF), and Logistic Regression (LR). These methods may help in the reliable early identification of breast cancer. This also provides as a basis for conducting a comparison of various methodologies. Finally, this study is critical in deciding which machine learning approach to apply when building a unified intelligent model. There are five parts to this research report. Section II is a review of the literature that summarizes the work done so far. The recommended approach was presented in Section III. Section IV uses graphs and tables to illustrate the comparative study of ML approaches utilizing whole and chosen attributes of a dataset. The conclusion reached as a consequence of the suggested framework is stated in Section V.

2. REVIEW OF LITERATURE

Several comparable efforts on breast cancer detection have previously been done in the present area by other researchers. Researchers used machine learning approaches such as SVM, random forest, KNN, decision tree, CNN, and logistic regression on various datasets, with accuracy, recall, AU-ROC, f1-score, sensitivity, specificity, and precision as comparative criteria.

K. Shailaja et al [1] examined the decision tree, SVM, naive Bayes, and K-nearest neighbor performance. The WISCONSIN dataset was utilized in the study. With 96.40 percent accuracy, the data show that SVM was the most suited.

C. Ming et al [2] compare current BCRAT and BOADICEA with machine learning models such as random forest, logistic regression, k-nearest neighbor, and MCMC GLMM utilizing accuracy and AU-ROC as comparative criteria. The accuracy of a machine learning model may be increased by 30 to 35 percent over previous models.

H. Dhahri et al. [3] employ multiple ML approaches to predict and diagnose breast cancer using the WISCONSIN dataset and analysis criteria such as sensitivity, accuracy, specificity, precision, and ROC curves. Using a mix of feature extraction, preprocessing, and classifier methods, genetic programming may determine the best model.

M.D Ganggayah et al [4] investigate the decision tree and random forest on the UMMC (University Malaya Medical Centre) dataset in Kuala Lumpur, Malaysia, using accuracy as the decision parameter. As a consequence, the random forest has an 82.7 percent accuracy.

Md. M. Islam et al [5] used the UCI dataset to compare ML models such as K-NN, ANN, SVM, RF, and LR, using accuracy, precision, and f1-score as decision parameters. The results of the experiments indicated that ANN was capable of achieving maximum accuracy of 98.57 percent, precision of 97.82 percent, and f1-score of 0.9890. C. Gupta et colleagues [6] use well-known machine learning classification methods to diagnose and characterize breast cancer illness on the Wisconsin breast cancer dataset, utilizing accuracy and time as comparison measures. According to the findings, an extreme learning machine has a 99 percent accuracy rate.

S. Bhise et al [7] investigate the performance of machine learning algorithms CNN, SVM, KNN, logistic regression, naive Bayes, and random forest on the BreaKHis 400X dataset, using accuracy and precision as decision parameters. According to the research, CNN outperforms other machine learning algorithms in terms of accuracy and precision.

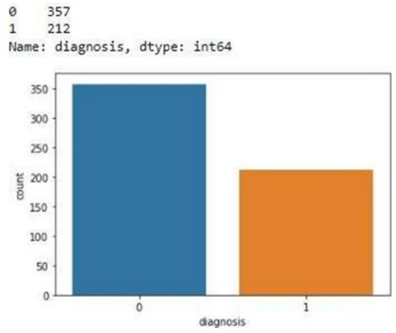
To reduce the total number of dimensions, H. Masood [8] used preprocessing, feature selection, and extraction. The author employs a variety of ML models, including ANN, ELM, SVM, KNN, MLP, NB, and CART. According to the findings, SVM is the best machine learning model.

3. METHODOLOGY PROPOSED

The experimental investigation in this research report concentrated mostly on categorization strategies. The work is divided into two parts: a) In phase one, the learning models instruct the machines on the provided dataset. b) The second step includes testing. Python is a programming language used to develop and test machine learning algorithms.

3.1 Data Collection

Data elicitation is the process of gathering information. The experimental effort made use of the Wisconsin Diagnostic Breast Cancer (WDBC) from www.kaggle.com. The collection includes 569 entries, 357 of which are benign (non-cancerous) and 212 of which are cancerous (malignant). The same is shown in Fig 3.1, where '0' denotes a non-cancerous tumor and '1' denotes a malignant tumor.



```
0    357
1    212
Name: diagnosis, dtype: int64
```

Fig 3.1: Representation of cancerous and non-cancerous cells

3.2 Data Preparation and Selection

Data preparation is conducted on any given dataset to enhance its quality by eliminating extraneous data. This technique is divided into three steps: data cleaning, data transformation, and data reduction. This paper's dataset has 32 properties such as radius, texture, area, perimeter, smoothness, compactness, concavity, symmetry, and many more. All of the dataset's properties are shown in Figure 3.2.

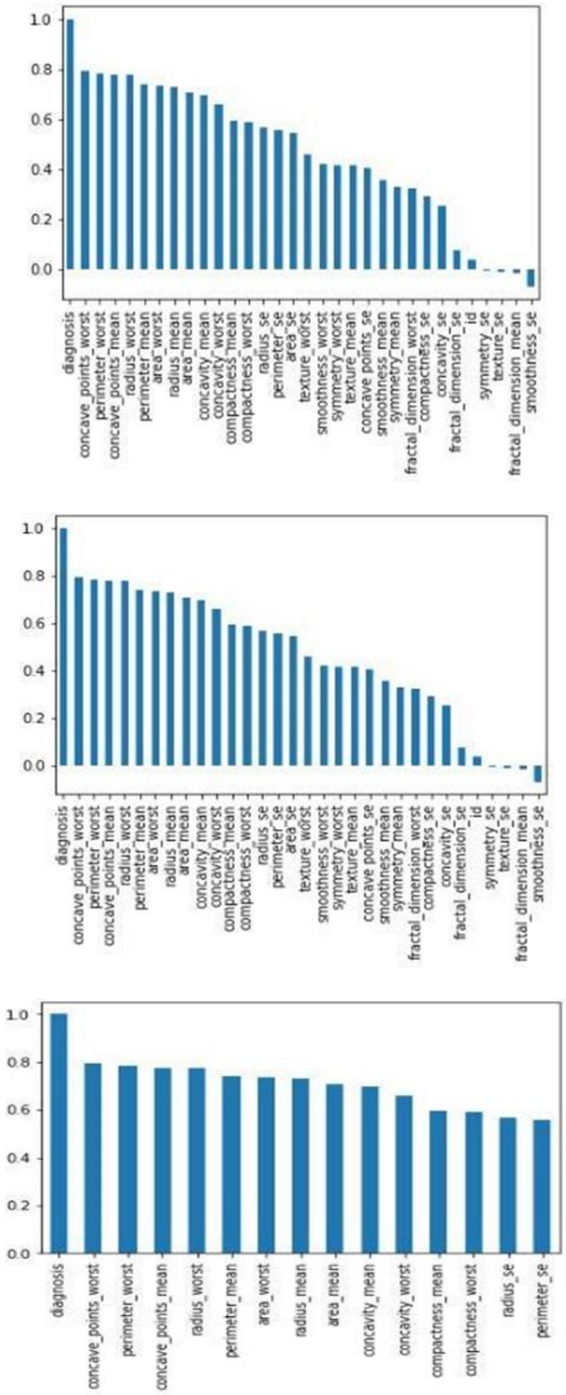


Fig 3.2.1: Dataset represents all attributes

It is crucial to determine if the data is balanced. The dataset is not equally balanced, as seen in Fig. 3.1. The amount of benign cells is about equivalent to that of malignant cells. The next stage includes a heat map to show the relationship between all features.

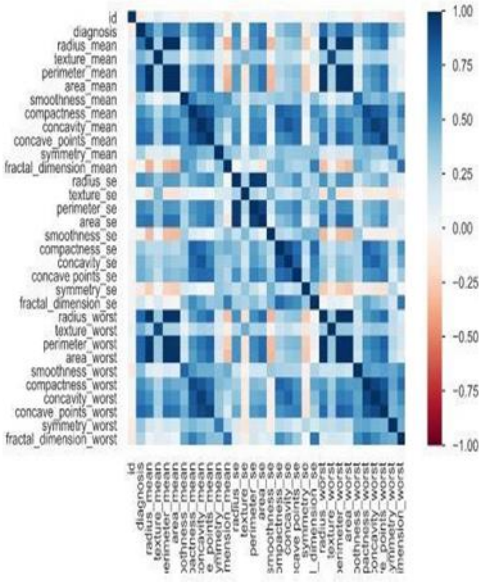


Fig 3.2.2: Independent variables represented by a heat map

Feature selection procedures, on the other hand, are a voluntary process that assists in dimension reduction by shifting features from a larger dimension to a smaller dimension area. The chosen attributes were retrieved when the data selection phase was finished. Fig. 3.3 illustrates the graph with chosen characteristics after some features have been removed.

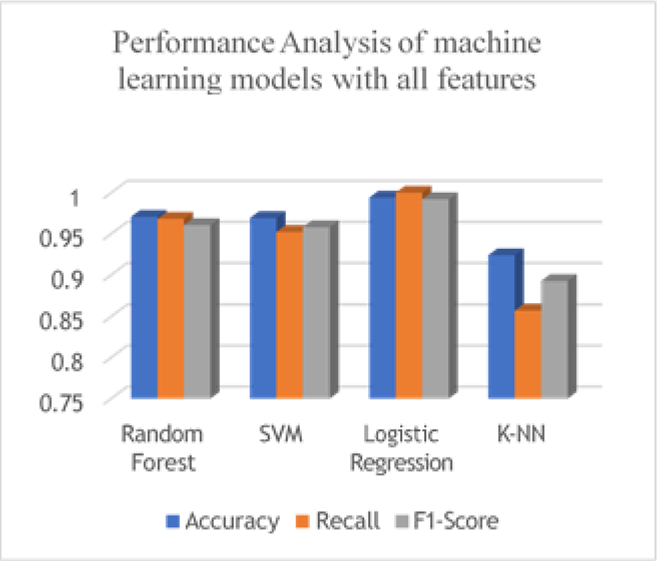


Fig 3.2.3: Dataset represents selected attributes

3.3 Implementing machine learning models

Machine learning models are applied to the given dataset at this step. Machine learning models will be trained and developed utilizing processed data in the present phase. Random forest, support vector machine, logistic regression, and k-nearest neighbor classification models were also investigated. The proposed model is put to the test in order to do a comparative analysis of performance.

4. COMPARATIVE ANALYSIS of ML TECHNIQUES on WDBC DATASET

Machine learning algorithms are utilized to diagnose breast cancer at an early stage using the proposed model. The four machine learning methods Random Forest, Logistic Regression, SVM, and K-NN were used for the breast WDBC dataset in the suggested technique. The accuracy, recall, and f1-score were used to assess the performance. Table 4.1(a) shows the same.

Table 4.1(a): Performance analysis of machine learning models have selected features

Algorithm used	Accuracy	Recall	F1-Score
Random Forest	0.9415	0.9206	0.9206
SVM	0.959	0.9491	0.9411
Logistic Regression	0.9298	0.8888	0.9032
K-NN	0.9239	0.8571	0.8925

The outcomes of different machine learning techniques are viewed and analyzed among each other. With the help of table 4.1(a), the comparison analysis has been drawn in fig 4.1(a) which can help for better understanding.

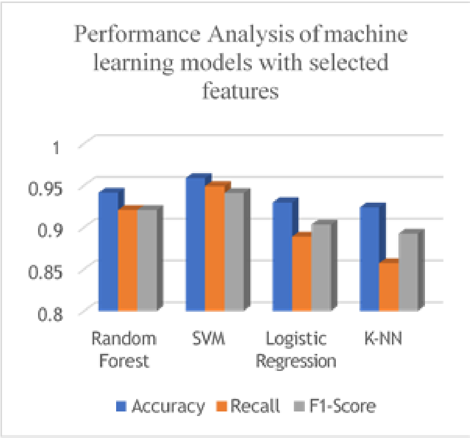


Fig 4.1(a): Performance analysis of machine learning models with selected features

According to the analysis report, it has been observed that SVM gives the best performance with selected features in terms of accuracy (95.90%), recall (94.91%), and f1-score (94.11%).

Table 4.1(b): Performance analysis of machine learning models with all available features

Algorithm Used	Accuracy	Recall	F1-Score
Random Forest	0.9707	0.9682	0.9606
SVM	0.9692	0.9523	0.9580
Logistic Regression	0.9941	1.0	0.9921
K-NN	0.9239	0.8571	0.8925

Machine learning algorithms' output is watched and assessed using all accessible characteristics. The comparative analysis in fig 4.1(b) has been created with the assistance of table 4.1(b), which may aid in comprehension. According to the graph in Fig. 4.1(b), Logistic Regression provides the greatest performance with all characteristics in terms of accuracy (99.41 percent), recall (100 percent), and f1-score (99.21 percent).

RESULTS AND CONCLUSIONS

Breast cancer is one of the most frequent and lethal illnesses that affect women. Machine learning approaches aid in illness diagnosis at an early stage. On the WISCONSIN data set, machine learning methods were used. There were two types of data sets used: one with chosen features and one with all features. Selected features are columns that have been chosen from a dataset using feature extraction or selection procedure. On the two sets of data, Random Forest, SVM, Logistic Regression, and KNN were used. The following characteristics were used to evaluate their performance: accuracy, recall, and F1 score. When all characteristics in the dataset are used, the overall efficiency of the ML algorithm rises. Furthermore, in the case of a restricted feature dataset, the SVM method outperformed the other three techniques. Logistic regression, on the other hand, performed best when all characteristics were included in the dataset. Logistic Regression has grown significantly in terms of accuracy, f1-score, and recall with all characteristics dataset. In none of the two scenarios does the performance of K-NN differ. As a result, K-NN is not reliant on data. When all features' data sets are utilized, Random Forest's performance improves. As a result, it is possible to deduce that the results are data-dependent. Furthermore, although it is true that an algorithm performs better with larger datasets, this is not required. A similar study may be done on datasets from several additional illnesses in the future as part of the scope.

REFERENCES

[1] K. Shailaja et al, “Machine learning in healthcare: A review”, ICECA 2018, IEEE Xplore ISBN:978-1-5386- 0965-1.

[2] Chang Ming et al, “Machine learning techniques for personalized breast cancer risk prediction: comparison with the BCRAT and BOADICEA models”, (2019) 21:75.

[3] Habib Dhahri et al, “Automated breast cancer diagnosis based on machine learning algorithms”, Hindawi Journal of Healthcare Engineering, Vol 2019, Article Id 4253641.

[4] Mogana Darshini Ganggayah et al, “Predicting factors for survival of breast cancer patients using machine learning techniques”, BMC Medical Informatics and Decision Making, (2019) 19:48.

[5] Md. Milon Islam et al, “Breast cancer prediction: A comparative study using machine learning techniques”, SN Computer Science, (2020) 1:290.

[6] Chhaya Gupta, Nasib Singh Gill, “Machine learning techniques and extreme learning machine for early breast cancer detection”, International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN:2278-3075, Vol. 9, Issue 04, February 2020.

[7] Sweta Bhise et al, “Breast cancer detection using machine learning techniques”, International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 10 Issue 07, July 2021.

[8] Hiba Masood, “Breast cancer detection using machine learning algorithm”, International Research Journal of Engineering and Technology (IRJET), e-ISSN: 2395- 0056, p-ISSN: 2395-0072, Vol. 8, Issue: 02, February 2021.

[9] David A. Omondiagbe et al, “Machine learning classification techniques for breast cancer diagnosis”, IOP Conf. Series: Materals Science and Engineering 495 (2019) 012033.

[10] Riyadh M. A1-Tam and Sachin M. Narangale, “Breast cancer detection and diagnosis using machine learning: A survey”, Journal of Scientific Research of The Banaras Hindu University, Vol. 65, Issue 5, 2021.

[11] Mitanshi Rastogi and Neha Goel , “A review of machine learning algorithms and its applications”, Vivekananda Journal of Research - Jan-June 2022, Vol. 12 Issue 1, ISSN 2319-8702(Print) and ISSN 2456-7574 (Online).

[12] Mitanshi Rastogi et al, “Role of machine learning in the healthcare sector”, International Conference on Computational and Intelligent Data Science (ICCIDS- 2022) paper to be published in Elsevier SSRN.

[13] Annina Simon, Mahima Singh Deo, S. Venkatesan, D.R. Ramesh Babu, “An overview of machine learning and its applications”, International Journal of Electrical Sciences & Engineering (IJESE), Vol 1, Issue 1; 2015 pp 22-24.

[14] Ayon Dey, “Machine learning algorithms: A review”, International journal of computer science and information technologies, Vol 7 (3), 2016,1174-1179.

[15] Ozer Celik, Serthan Salih Altunaydin, “A research on machine learning methods and its applications”, Journal of Educational Technology & Online learning, Vol 1, Issue 3, 2018.

[16] Iqbal H. Sarker, “Machine Learning: Algorithms, Real-world applications and Research directions”, SN Computer Science- a springer nature, (2021)2:160.