Machine Learning Classifiers for Breast Cancer Diagnosis

Dik Sharma
PG Scholar, Department of MCA
Dayananda Sagar College of Engineering
Bengaluru, India
diksharma040@gmail.com

***Abstract*— The most frequent type of cancer in women is breast cancer. Effective therapy depends on early diagnosis and discovery. It has been demonstrated that ML classifiers are useful for detecting breast cancer. In this study, we examine the Wisconsin Diagnosis Breast Cancer dataset performance of three ML classifiers: logistic regression, k-nearest neighbors (KNN), and support vector machines (SVM). The SVM model performs better than the other two models, according to the experimental findings.**

Keywords— K-nearest Neighbor; Logistic Regression; Machine Learning; Support Vector Machine

I. Introduction

The most frequent type of cancer found in women is breast cancer. In 2020, there were an estimated of around 2.2 million new cases of breast cancer and 685,000 deaths from the disease. Effective therapy depends on early diagnosis and discovery.

It has been demonstrated that machine learning (ML) can be helpful in diagnosing breast cancer. ML classifiers can be trained on a dataset of historical data to learn the patterns that distinguish between benign and malignant tumors. Once trained, ML classifiers can be used to precisely identify new tumors. For example, using the results of the biopsy to determine whether the patient needs surgery or not. In cases where surgery is conducted to remove malignant cells, there are instances where it is later revealed that the cells are benign, meaning they are non-cancerous. Patients end up having needless, uncomfortable, and expensive procedures as a result. For healthcare-related datasets like pictures, x-rays, and blood samples, machine learning algorithms have various advantages. Different approaches are better suited for small or large datasets, and some techniques may run into problems with data noise.

In this paper, we compare the performance of three ML classifiers, logistic regression, k-nearest neighbors (KNN), and support vector machines (SVM), on the Wilson Diagnosis Breast Cancer dataset.

II. Dataset

A publicly accessible dataset with information from 569 breast cancer patients is used; it is called the Wisconsin Diagnosis Breast Cancer dataset. Tumor size, tumor grade, and lymph node status are among the characteristics. The classification of breast cancer—benign or malignant—is the target variable. The dataset was cleaned before being randomly split into two sections using the holdout method. 80% of the dataset needed for training with 455 observations is in the first part.The second part contains 20% of the dataset to present testing with 114 observations.

In order to perform feature selection, 12 distinct variables have been discovered to be suitable for the categorization following comprehensive research.

	mean texture	mean smoothness	mean compactness	mean symmetry	texture error	smoothness error	compactness error	symmetry error	worst texture	worst smoothness	worst compactness	worst symmetry	target
0	10.38	0.11840	0.27760	0.2419	0.9053	0.006399	0.04904	0.03003	17.33	0.1622	0.6656	0.4601	0.0
1	17.77	0.08474	0.07864	0.1812	0.7339	0.005225	0.01308	0.01389	23.41	0.1238	0.1866	0.2750	0.0
2	21.25	0.10960	0.15990	0.2069	0.7869	0.006150	0.04006	0.02250	25.53	0.1444	0.4245	0.3613	0.0
3	20.38	0.14250	0.28390	0.2597	1.1560	0.009110	0.07458	0.05963	26.50	0.2098	0.8663	0.6638	0.0
4	14.34	0.10030	0.13280	0.1809	0.7813	0.011490	0.02461	0.01756	16.67	0.1374	0.2050	0.2364	0.0

Fig 1: Dataset (sample)

III. Objectives of the Proposed Method

The technique of identifying breast cancer has been enhanced using machine learning. The main goal is to get an accurate diagnosis.

- To get the correct diagnosis by removing human error.
- To reduce time and human resources.
- To encourage the use of machine learning technologies in healthcare.

IV. Review of Literature

A literature review showed that there have been several studies on the survival prediction problem using statistical approaches and artificial neural networks. However, there aren't many studies that use data mining techniques like decision trees to diagnose [10,11]. Delen et al. used artificial neural networks, decision trees and logistic regression to develop prediction

models for breast cancer survival by analyzing a large dataset, the SEER cancer incidence database [11]. Lundin et al. used ANN and logistic regression models to predict 5, 10, and 15 - year breast cancer survival. They studied 951 breast cancer patients and used tumour size, axillary nodal status, histological type, mitotic count, nuclear pleomorphism, tubule formation, tumour necrosis, and age as input variables [12]. Pendharker patterns in breast cancer. In this study, researchers demonstrated how data mining could be a useful tool for spotting trends in breast cancer cases that could be used for diagnosis, prognosis, and therapy purposes. [9]. Ahmad LG et al. used DT, ANN and SVM to predict recurrence in patients who were followed -up for two years and used advanced data mining techniques to discover hidden patterns and relationships[8]. These studies are some examples of researches that apply data mining to medical fields for prediction of diseases.

V. METHODOLOGY

A. Logistic Regression:

Logistic Regression is a statistical technique used to make predictions about a binary outcome, such as a yes or no decision, by analysing previous observations from a dataset. A logistic regression model analyses the relationship between one or more existing independent factors to predict a dependent data variable [3].

Logistic regression predicts the probability of the default class and transforms the probability into a binary value (0 or 1) for classification using the" sigmoid" function as:

$$f(x) = \frac{1}{1 + e^{-x}}$$

B. K-Nearest Neighbor (KNN):

KNN is a non-parametric method used for classification. It is also among the most well-known classification algorithms. The basic idea is that predetermined features form a space in which known data are ordered. K-nearest neighbor assigns a case to the class that is most common among its k nearest neighbors.[5] The distance between the case and its neighbor is measured by using distance functions like Euclidean:

$$D_{Euclidean} = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$$

C. Support Vector Machine (SVM):

A versatile model used for regression and classification problems is the Support Vector Machine (SVM). It is suitable for a variety of practical applications since it can handle both linear and non-linear issues [7]. The basic principle of SVM is to create a line or hyperplane that effectively separates the data into distinct classes. By utilizing the kernel trick, the algorithm applies data transformations that enable the identification of an optimal boundary between different output possibilities. This allows SVM to effectively address complex classification challenges.

$$F(x, xj) = sum(x.xj)$$

Here, x, xj represents the data that needs to be classified.

VI. FLOW DIAGRAM FOR THE PROPOSED APPROACH

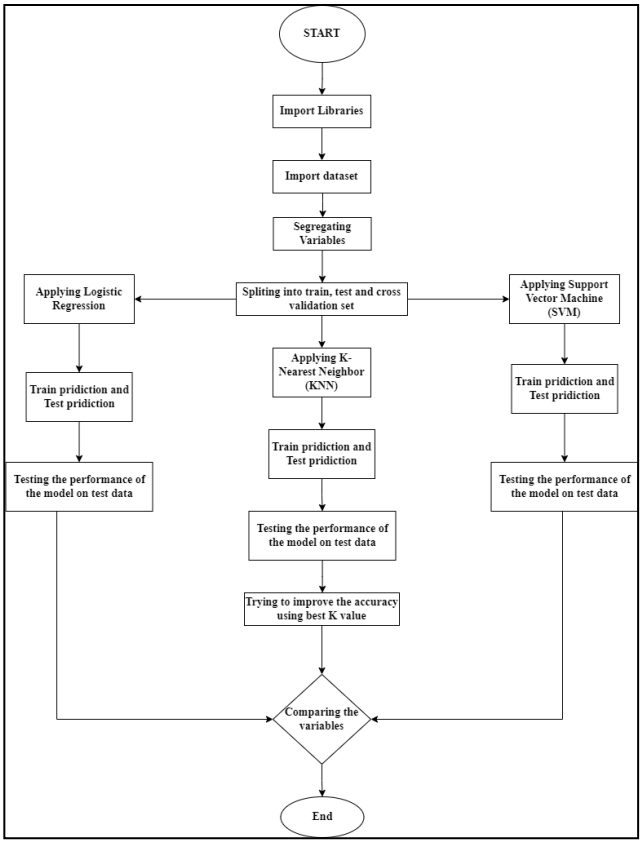

Fig.2: Flow Diagram

VII. Results

A. Logistic Regression:

Table 1: Logistic Regression Results.

Logistic Regression	Precision	Recall	F1-score	**Support**
0.0	0.80	0.81	0.80	48
1.0	0.86	0.85	0.85	66
Accuracy			0.83	114
Macro avg	0.83	0.83	0.83	114
Weighted avg	0.83	0.83	0.83	114

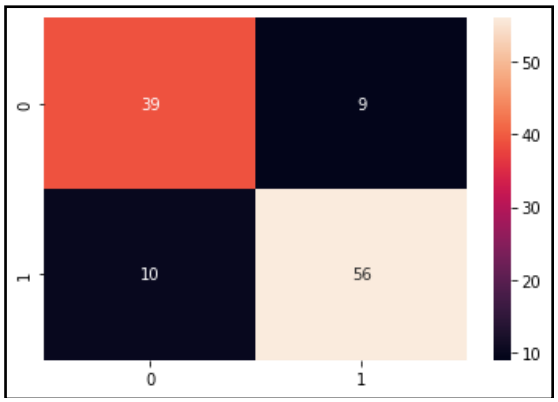

Fig. 1: Confusion Matrix for Logistic Regression

A. K-Nearest Neighbor (KNN):

Table 2: K-Nearest Neighbor Results.

KNN	Precision	Recall	F1-score	Support
0.0	0.64	0.79	0.71	48
1.0	0.82	0.68	0.74	66
Accuracy			0.73	114
Macro avg	0.73	0.74	0.73	114
Weighted avg	0.74	0.73	0.73	114

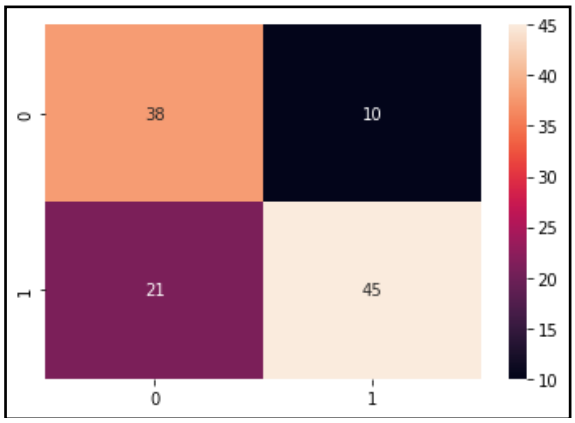

Fig. 2: Confusion Matrix for K-Nearest Neighbor

B. Support Vector Machine:

Table 3: Support Vector Machine Results.

Support Vector Machine	Precision	Recall	F1-score	Support
0.0	0.82	0.83	0.82	48
1.0	0.88	0.88	0.87	66
Accuracy			0.85	114
Macro avg	0.85	0.85	0.85	114
Weighted avg	0.85	0.85	0.85	114

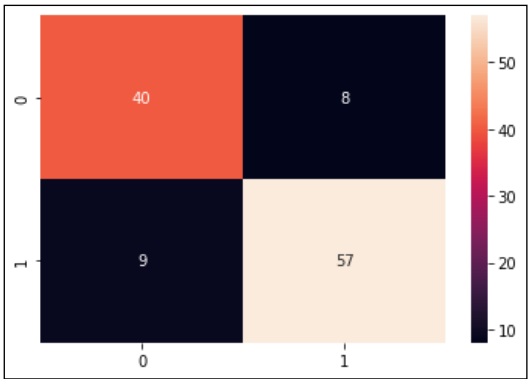

Fig. 3: Confusion Matrix for Support Vector Machine

To compare the models, the data from the Wisconsin Breast Cancer dataset were analyzed. Table 1, Table 2 and Table 3 shows the summaries of accuracy, precision and F1-score for different classification techniques.

VIII. CONCLUSION

In this paper, we have compared the performance of three ML classifiers, logistic regression, KNN, and SVM, on the Wilson Diagnosis Breast Cancer dataset. We have shown that two of the three classifiers are able to achieve high accuracy in classifying breast cancer cases. SVM performs better than the other classifier, though, with an accuracy of 85%.

FUTURE WORK

In future work, we plan to explore other ML classifiers and hyperparameters for breast cancer diagnosis. We also plan to evaluate the performance of the ML classifiers on other breast cancer datasets.

Acknowledgment

We would like to thank the Wisconsin Diagnosis Breast Cancer dataset for providing the data for this study.

REFERENCES

[1] Mangasarian, O. L., & Street, W. N. (1995). Breast cancer diagnosis and prognosis via linear programming. Operations Research, 43(4), 570-577. doi:10.1287/opre.43.4.570.

[2] Wolberg, W. H., & Mangasarian, O. L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. Proceedings of the National Academy of Sciences, 87(23), 9193-9196. doi:10.1073/pnas.87.23.9193

[3] N. Chakrabarty, S. Chowdhury, and S. Rana, "A Statistical Approach to Graduate Admissions," Chance Prediction, pp. 145-154, (2020).

[4] L. Lei, "Research on Logistic Regression Algorithm of Breast Cancer Diagnose Data by Machine Learning," International Conference on Robots & Intelligent Systems (ICRIS), pp. 157-160, July 2018. DOI: 10.1109/ICRIS.2018.00049.

[5] N. Gupta, A. Sawhney, and D. Roth, "Will I Get in? Modelling the Graduate Admission Process for American Universities," IEEE Int.

[6] Hsu, C. W., & Lin, C. J. (2002). A comparison of methods for multiclass support vector machines. IEEE transactions on neural networks, 13(2), 415-425.

[7] Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine learning,46(1-3),389-422.

[8] A. LG, E. AT,Porrebrahumi A., Ebrahimi M., & Razavi AR (2013).Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence,J Heal Med Informatics, 04 (2013).pp. 1-4.

[9] Pendharkar PC, Rodger JA, Yaverbaum GJ, Herman N, Benner M (1999).Association, statistical, mathematical and neural approaches for mining breast cancer patterns. Expert Systems with Applications 17: 223-232.

[10] Zhou ZH, Jiang Y (2003) Medical diagnosis with C4.5 Rule preceded by artiicial neural network ensemble. IEEE Trans Inf Technol Biomed 7: 37-42.

[11] Delen D, Walker G, Kadam A (2005) Predicting breast cancer survivability: a comparison of three data mining methods. Artiicial Intelligence in Medicine 34: 113-127.

[12] Lundin M, Lundin J, Burke HB, Toikkanen S, Pylkkanen L, et al. (1999) Artiicial neural networks applied to survival prediction in breast cancer. Oncology 57: 281-286.