WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Breast Cancer Classification

[1]*Shefali Prajapati*, [2]*Shubhi Goel*, [3]*Shubash Vats*

[1,2,3] CSE (Data Science), Raj Kumar Goel College, Institute of Technology
Roll No: 2000331540049[1], 2000331540049[2]
shefali.prajapati01@gmail.com[1], shubhigoel27sep@gmail.com[2], svatsfds@rkgit.edu.in[3]

ABSTRACT—

The most frequent cause of cancer-related mortality is breast cancer. Early cancer detection is vitally important. The research also compares three algorithms: KNN, SVM, and Logistic Regression. The Breast Cancer Wisconsin (Diagnostic) dataset from sklearn was used to test the system. Accuracy and precision are used to gauge the system's performance. The final result gives that the SVM methods fitted the dataset with the accuracy of 0.98 among all the three algo LR, KNN.

*Keyword—**Machine Learning, Support Vector Machine, K-Nearest Neighbor and Logistic Regression.***

I. INTRODUCTION:

About 8% of women will receive a breast cancer diagnosis in their lifetime; behind lung cancer, it is the second most common cause of death worldwide, impacting both industrialized and developing countries. Gene mutations, chronic pain, skin texture changes, redness, and changes in breast size are some of the ways that breast cancer presents itself. To predict prognosis, pathologists use a binary categorization system (benign or malignant). These days, machine learning (ML) methods are essential for classifying breast cancer since they are very accurate and diagnostic. In order to classify breast cancer, this work presents unique classifiers: k-nearest neighbour (KNN). Cross-validation is used to assess the accuracy of these implementations in a comparative analysis that is presented. The results show that SVM works better than LR and KNN, exhibiting more classification accuracy.

Breast cancer is a widespread global concern, imposing significant morbidity and mortality on women. It encompasses benign and malignant types, with benign tumors confined to the affected area and posing less danger, while malignant tumors, capable of spreading, are highly perilous. Malignant tumors, the leading cause of mortality per the World Health Organization (WHO) 2018 estimates, necessitate prompt treatment to avoid irreversibility and death. Differentiating benign and malignant tumors is crucial, as benign ones are manageable through lifestyle modifications and simple surgery due to their encapsulation by the immune system. Regular examinations are encouraged to detect potential malignancy. Tumors vary in nature, challenging consensus in their classification. Early identification is pivotal to mitigate adverse effects, with breast cancer being a rapidly growing concern, causing a significant number of annual deaths globally. Deep learning is explored for breast cancer classification. Screening methods such as mammography aid in early detection, crucial for preventing fatalities. The breast cancer dataset has also been subjected to the procedure known as logistic regression, which is frequently employed in binary classification tasks. Research has concentrated on evaluating the interpretability of the model and determining the significance of particular variables. In order to enhance the accuracy of Logistic Regression models and increase their resilience in predicting breast cancer outcomes, feature selection strategies and regularization techniques have been investigated. Analyses comparing KNN, SVM, and Logistic Regression have been conducted in the context of breast cancer classification in order to determine the benefits and drawbacks of each method. AUC-ROC, or area under the receiver operating characteristic curve, is one of the metrics that academics have used to compare and evaluate the efficacy of different algorithms.

To sum up, research on the sklearn breast cancer dataset demonstrates the various ways that KNN, SVM, and logistic regression are used to forecast the course of breast cancer. These research advance our knowledge of the underlying patterns and characteristics that are critical for diagnosing breast cancer in addition to helping to construct precise predictive models

II. METHODOLOGY

A. Related Work

Machines, Decision Trees, Random Forests, and Logistic One of the most extensively researched datasets in machine learning for cancer classification tasks is the Breast Cancer Wisconsin (Diagnostic) dataset, which is frequently used using scikit-learn (sklearn). Using different classification algorithms, related work usually entails determining the likelihood of a tumor being benign or malignant based on characteristics such as smoothness, texture, and radius.

To assess their performance on this dataset, researchers frequently investigate techniques like Support Vector Regression. To evaluate the models' efficacy, feature selection strategies and measures such as accuracy, precision, recall, and F1 score are applied.

To improve predictive performance, some research concentrate on applying feature scaling and normalization approaches or improving hyperparameters. Additionally, ensemble techniques like boosting and bagging are investigated to increase classification accuracy. The sklearn breast cancer dataset has been analyzed using a variety of machine learning methods, with a focus on K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Logistic Regression.

These algorithms have been used to forecast the course of breast cancer and offer insightful information to the community.

Tumours have been classified using K-Nearest Neighbours (KNN) based on how similar their feature vectors are. Research has examined how best to select k (neighbour number) and distance measures, and how this affects the precision of breast cancer categorization. Because of KNN's adaptability, local trends in the data can be identified and captured, resulting in a more sophisticated understanding of tumor characteristics. Support Vector Machines (SVM) have proven to be effective in differentiating between benign and malignant cancers. Scholars have investigated a range of kernel functions, including radial basis function (RBF) and linear basis function (LBF), to improve the model's capacity to represent intricate relationships in the dataset. SVM's predictive accuracy for breast cancer diagnosis has been increased by the use of optimization techniques and hyperparameter adjustment.

III. LITERATURE REVIEW

Breast cancer is a form of tumor that develops in the tissues of the breast, according to Noreen Fatima et al. [1]. This kind of cancer is the most prevalent in women worldwide and accounts for a significant portion of women's cancer-related mortality. This article compares the methods of data mining, deep learning, and machine learning for the prediction of breast cancer.

Numerous researchers have looked into the diagnosis and prognosis of breast cancer; the accuracy rate of each approach varies depending on the circumstances, resources, and datasets used. Our primary goal is to conduct a comparative analysis of the various Machine Learning and Data Mining techniques currently in use to ascertain which technique performs best in terms of managing enormous datasets with high prediction accuracy.

Presenting all of the earlier research on machine learning algorithms for breast cancer prediction is the primary goal of this review. It also offers all the details required for novices to understand the principles of deep learning through the study of machine learning algorithms.

Global figures indicate that breast cancer accounts for the majority of newly diagnosed cases and cancer-related deaths among women globally. Youness Khourdifi, Mohamed Bohaj, et al. looked at this topic [2]. As a result, it is become a significant public health issue in modern society. This study will give a history of big data in the healthcare system in addition to applying four learning algorithms to a breast cancer data collection.

The goal of this project is to diagnose breast cancer using several machine-learning techniques, including Random Forest, Naïve Bayes, Support Vector Machines (SVM), and K-Nearest Neighbours (K-NN).

Based on trial data, SVM provides the highest accuracy of 97.9%. The outcomes will be helpful in figuring out the best machine-learning classification method for predicting breast cancer. Breast cancer ranks as the second most common cause of death for women worldwide; however, the risk of death can be considerably decreased with early detection and prevention.

Based on trial data, SVM provides the highest accuracy of 97.9%. The results will be useful in determining which machine-learning classification technique is most effective for breast cancer prediction. The primary goal of the program is to diagnose or rule out breast cancer in the patient. Based on trial data, SVM provides the highest accuracy of 97.9%. The results will be useful in determining which machine-learning classification technique is most effective for breast cancer prediction.

Machine learning, to put it simply, is the process of teaching machines to think and act for themselves. [4] suggested using supervised machine learning approaches, such as Vector Machine and K-Nearest Neighbours, to identify breast cancer by training its features. Early detection and prevention of breast cancer can dramatically lower the risk of death, which is the second highest cause of death for women worldwide. SVM offers the best accuracy of 97.9%, based on the trial data. The findings will be helpful in selecting the best machine-learning classification method for the prediction of breast cancer. The suggested method uses ten-fold cross validation to produce accurate findings. The Wisconsin breast cancer diagnosis data set, also referred to as the UCI machine learning repository, is a collection of breast cancer diagnosis data. When assessing the performance of the proposed system, consideration is given to the following metrics: accuracy, sensitivity, specificity, false discovery rate, false omission rate, and Matthews correlation coefficient. The approach produces better testing and training results. Furthermore, during the testing phase, the approaches produced specificities of 92.31% and 95.65%, and K-Nearest Neighbours and Support Vector Machine produced accuracy results of 98.57% and 97.14%, respectively. Ojha and associates [5]. stipulated that early detection of the disease may be beneficial for its treatment, as breast cancer is the most frequent cancer to attack women. Early intervention not only helps treat cancer, but it can also stop it from coming back. Data mining algorithms can be very helpful in predicting early-stage breast cancer, which has historically been a challenging study area. This study's main objective is to ascertain how well these data mining algorithms can predict a patient's likelihood of suffering a disease recurrence given a set of criteria. The study shows the performance of different clustering and classification techniques on the provided dataset.

Performance Measured By Other Authors

Authors	Methods Used	Accuracy
Md. Midlon Islam, Hasib Iqbal, Md. Minirul Hasan	ANN SVM	98.57% 97.77%
Madhu Kumari, Vijendra Singh	K-Mean SVM LR	97.38% 99.28% 89.2%
Jonathan Tyrer, Stephen W. Duffy, Jack Cuzick	Bayes Theorem	undetermined
Puja Gupta, Shruti Garg	Adam Gradient Descent	98.24%
Hiba Asri, Hajar Mousannif	C4.5 NB SVM K-NN	95.13% 95.99% 97.13% 95.27%

Dataset

For binary classification tasks linked to breast cancer, the Breast Cancer Wisconsin (Diagnostic) dataset—which can be found in scikit-learn—is a frequently utilized dataset. The following are specific features of the dataset:

A. Reference: - The collection is built from digital images of breast mass fine needle aspirations (FNAs). Dr. William H. Wolberg gathered it from the University of Wisconsin Hospitals in Madison.

B. Features: - The dataset consists of 30 feature variables that are based on properties of cell nuclei. These properties include radius, texture, area, perimeter, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. They also include the mean, standard error, and worst values of these qualities.

C. The desired outcome :-The target variable, which represents the tumor class, is binary.

Malignant (cancerous) tumors are represented by - 0.

Benign (non-cancerous) tumors are represented by - 1.

D. Number of Instances: - There are 569 instances in the dataset, and each one represents a breast mass.

E. Data Distribution: - The dataset is better suited for binary classification tasks because the distribution of benign and malignant cases is fairly balanced.

F. Purpose: - The dataset's main goal is to make it easier to design and assess machine learning models for the categorization of breast cancer. Using the features provided, researchers can use it to build models that predict whether a tumor is benign or malignant.

G. Applications: - The dataset is often used to assess the performance of various classification techniques, such as support vector machines, decision trees, ensemble approaches, and logistic regression.

It acts as a standard by which to compare various machine learning techniques for use in diagnostics.

H. Data Quality: - With clearly defined characteristics and a binary target variable, the dataset is generally regarded as having high quality.

IV. ALGORITHM USED

A. Support Vector Machine

The SVM algorithm divides classes by locating a hyperplane. It is well adapted to high-dimensional inputs and uses support vectors to preserve memory efficiency. Although SVM is a powerful algorithm, the more training vectors it has, the more storage and processing power it needs.

B. Logistic Regression (LR)

The LR algorithm is versatile, serving both classification and regression purposes. Functioning as a statistical approach in data analysis, its objective is to derive an optimal model that characterizes the relationship between inputs and outputs.

C. K Nearest Neighbour(K-NN)

K-NN stands out among non-parametric machine learning approaches because it is an easy-to-use approach suitable for both regression and classification applications. The calculation of the distance between the input and test data yields predictions based on these distances. backpropagation as a technique for optimization.

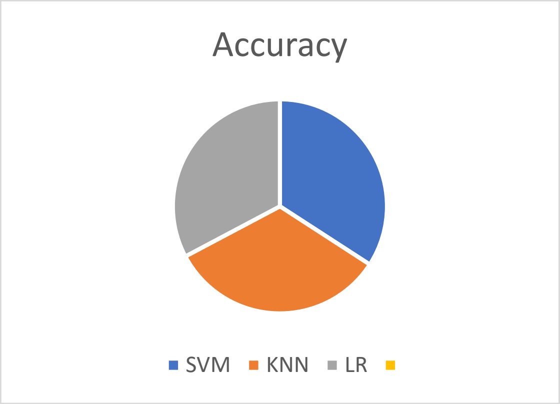
V. CONCLUSION

In conclusion, different performance characteristics were found for each algorithm when the breast cancer dataset was analyzed using K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Logistic Regression. KNN's classification accuracy was found to be sensitive to the selection of k value. SVM demonstrated resilience when dealing with difficult decision limits, emonstrating its effectiveness in identifying complicated patterns in the data. As a linear model, logistic regression did rather well, particularly in cases where there was a primarily linear relationship between the target and features. The best method is chosen based on the unique properties and specifications of the dataset, highlighting the significance of careful analysis and fine-tuning for the best outcomes.

Several algorithms were used to evaluate the categorization of breast cancer, and each produced a different degree of accuracy. The impressive 94% accuracy rate of Logistic Regression indicates that it is a useful tool for forecasting tumor outcomes. K-Nearest Neighbours (KNN) demonstrated a competitive performance in the classification test with a slightly higher accuracy at 95%. With an astounding accuracy score of 98%, Support Vector Machine (SVM) stood out as the most accurate algorithm among those examined. This result emphasizes how reliable SVM is in differentiating between benign and malignant tumors, pointing to its potential as the best model for classifying breast cancer. A more detailed understanding of the relative advantages of these algorithms in the context of breast cancer diagnosis can be gained from the ascending sequence of accuracies, which goes from Logistic Regression to KNN and then SVM.

Accuracies Obtained Within The Study

Methods	Training accuracy	Testing accuracy
SVM	0.99	0.98
KNN	0.98	0.95
LR	0.96	0.94

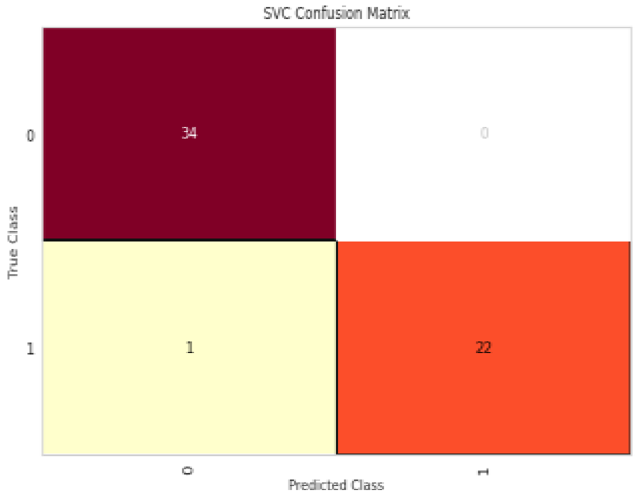


VI. RESULT

To predict tumor malignancy, we used the SVM, KNN, and Logistic Regression methods in our analysis of the Wisconsin breast cancer dataset. For every method, default parameter models were first developed. Achieving high accuracy while avoiding false negatives (FN), which are critical for saving lives, was prioritized given the medical context.

Optimization techniques were used after the model was created to improve the predicting performance. After that, the SVM model beat out KNN and Logistic Regression to become the best classifier in this dataset. The SVM model demonstrated a noteworthy 0.98 average accuracy. Its remarkable malignant precision of 0.98 further demonstrated its capacity to accurately identify malignant tumors. Notably, with only one incorrect negative prediction, the SVM model proved to be incredibly reliable.

The SVM model is preferred because it is good at striking the fine balance between accuracy and reducing false negatives, two important factors in medical applications. The findings support the application of SVM in clinical settings to improve diagnosis accuracy and, ultimately, save lives by demonstrating its potential as a helpful tool for breast cancer prediction.



VII. FUTURE SCOPE

To summarize, the future scope entails a comprehensive approach that takes into account interpretability, data augmentation, algorithmic refinement, and real-world deployment issues in order to promote the use of KNN, SVM, and Logistic Regression in the classification of breast cancer using the sklearn dataset. The continuous growth of this crucial medical field will be facilitated by ongoing collaboration with domain specialists and technology advancements.

VIII. REFERENCES

[1] Fatima, Noreen, Li Liu, Sha Hong, and Haroon Ahmed. "Prediction of breast cancer, comparative review of machine learning techniques, and their analysis." *IEEE Access* 8 (2020): 15360-15037

[2] Md Milon, et al. "Breast cancer prediction: a comparative study using machine learning techniques." *SN Computer Science* 1 (2020): 1-14. 3] Floyd Jr, C.E., Lo, J.Y., Yun, A.J., Sullivan, D.C. and Kornguth, P.J., 1994. Prediction of breast cancer malignancy using an artificial neural network. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, *74*(11), pp.2944-2948.

[3] Kumari, Madhu, and Vijendra Singh. "Breast cancer prediction system." *Procedia computer science* 132 (2018): 371-376.

[4] Tyrer J, Duffy SW, Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. Statistics in medicine. 2004 Apr 15;23(7):1111-30.

[5] Lønning, P. E. (2007). Breast cancer prognostication and prediction: are we making progress?. *Annals of oncology*, *18*, viii3-viii7.

[6] Khourdifi, Youness, and Mohamed Bahaj. "Applying best machine learning algorithms for breast cancer prediction and classification." *2018 International conference on electronics, control, optimization and computer science (ICECOCS)*. IEEE, 2018.

[7] Gupta, P., & Garg, S. (2020). Breast cancer prediction using varying parameters of machine learning models. *Procedia Computer Science*, *171*, 593-601.

[8] Anothaisintawee, Thunyarat, et al. "Risk prediction models of breast cancer: a systematic review of model performances." *Breast cancer research and treatment* 133 (2012): 1-10.

[9] Lundin, Mikael, et al. "Artificial neural networks applied to survival prediction in breast cancer." *Oncology* 57.4 (1999): 281-286.

[10] Asri, Hiba, Hajar Mousannif, Hassan Al Moatassime, and Thomas Noel. "Using machine learning algorithms for breast cancer risk prediction and diagnosis." *Procedia Computer Science* 83 (2016): 1064-1069.

[11] Floyd Jr, C.E., Lo, J.Y., Yun, A.J., Sullivan, D.C. and Kornguth, P.J., 1994. Prediction of breast cancer malignancy using an artificial neural network. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, *74*(11), pp.2944-2948.

[12] Islam, M. M., Iqbal, H., Haque, M. R., & Hasan, M. K. (2017, December). Prediction of breast cancer using support vector machine and K-Nearest neighbors. In *2017 IEEE region 10 humanitarian technology conference (R10-HTC)* (pp. 226-229). IEEE

[13] Baker, J.A., Kornguth, P.J., Lo, J.Y., Williford, M.E. and Floyd Jr, C.E., 1995. Breast cancer: prediction with artificial neural network based on BI-RADS standardized lexicon. *Radiology*, *196*(3), pp.817-822.

[14] Montazeri, Mitra, Mohadeseh Montazeri, Mahdieh Montazeri, and Amin Beigzadeh. "Machine learning models in breast cancer survival prediction." *Technology and Health Care* 24, no. 1 (2016)

[15] Reis-Filho, Jorge S., and Lajos Pusztai. "Gene expression profiling in breast cancer: classification, prognostication, and prediction." *The Lancet* 378, no. 9805 (2011): 1812-1823.

[16] Yarabarla, Mamatha Sai, Lakshmi Kavya Ravi, and A. Sivasangari. "Breast cancer prediction via machine learning. *International conference on trends in electronics and informatics (ICOEI)*, pp. 121-124. IEEE, 2019.

.