Volume 3, Issue 1, 2024
PP. 498-505

International Journal of Futuristic Innovation in Arts, Humanities and Management (IJFIAHM)

BREAST CANCER PREDICTION AND PROGNOSIS USING MACHINE LEARNING

Dooman Maitry, Nikita Pandey, Tushar Sahu

U. G. Scholar, Department of Information Technology,

J. S. Government Engineering College, Jagdalpur, Bastar (C. G.), India

doomanmaitry287@gmail.com

U. G. Scholar, Department of Information Technology,

J. S. Government Engineering College, Jagdalpur, Bastar (C. G.), India

nikkipandey06460@gmail.com

U. G. Scholar, Department of Information Technology,

J. S. Government Engineering College, Jagdalpur, Bastar (C. G.), India

tusharsahu0212@gmail.com

Abstract: Nowadays, Breast cancer is the most commonly diagnosed life-threatening cancer in women and the leading root of cancer death among women. In the last two years, research linked to breast cancer has conducted remarkable progression in our understanding of the disease, resulting in more efficient and less toxic treatments. Increased public awareness and improved screening have led to earlier diagnosis at stages amenable to complete surgical resection and curative therapies. Consequently, survival rates for breast cancer have improved significantly, particularly in younger women. This article addresses the types, causes, clinical symptoms, and various approaches both non-drug (such as surgery and radiation) and drug treatment (including chemotherapy, gene therapy, etc.) of breast cancer.

Keywords: Breast Cancer, Tumor, Chemotherapy, Gene Therapy

I. INTRODUCTION

According to the World Health Organization, the top two causes of cancer death in 2023 were breast cancer and lung cancer [1]. Approximately, 27 percent of all cancer deaths are attributed to breast cancer [2]. Many cancers (such as breast, lung, mouth, brain, blood, and bladder), and early-stage diseases cause no important indication but treatment is approved because of a prediction that a danger would progress and scare a patient's quality of life or survival [3]. The production of oxygen radicals a role in the growth of cancers and iron chelators (removal of iron in the body) have defeated the cell growth through the body tissue cell, which is dangerous to breast cancer patients.

[4] There are different types of breast cancer; (1) Ductal Carcinoma in Situ (DCIS), (2) Invasive Ductal Carcinoma (IDC), (3) Mixed Tumours Breast Cancer (MTBC), (4) Lobular Breast Cancer (LBC), (5) Mucinous Breast Cancer (MBC) and (6) Inflammatory Breast Cancer (IBC) [18].

A. Breast Cancer Diagnosis and Research:

Breast cancer is a condition when the cells in a woman's body grow out of control and migrate to other areas of her body [5]. Since there are billions of cells in the human body, cancer can begin in any place of the body. Genetics has a role in some malignancies. Stated differently, the individual's parents' DNA may provide a genetic predisposition to cancer, hence increasing the parent's risk of contracting the illness. Additionally, unprofitable treatment planning is seen in multiplex and other disorders.

1. Computation Analysis for Breast Cancer Research:

Pathologists are essential to the medical business due to the rise in various forms of breast cancer and other illnesses, and doctors rely on them to make accurate and well-organized diagnosis. However, even with experience, the histological analysis is laborious and prone to human error when performed by a pathologist [6]. Over the next few decades, to apply fundamental research on breast cancer to computational and mathematical methods.

2. Role of ML Techniques in Cancer Predication and Prognosis:

A worldwide search of ML techniques in breast cancer reactivity, outbreak, and survival deviation was conducted. According to their survey based on ML use of cancer prediction, we have seen a large rise in documents released in the last decade [7]. ML algorithms suitable to the prediction result of breast cancer patients, which is mainly used; first as SVM classifiers and KNN classifiers.

B. Machine Learning Methodology:

Machine learning (ML) is a subsection of artificial intelligence (AI) that prioritizes developing approaches that learn—or upgrade the performance—found in the data they consume. And, it gives the computer that makes it more related to humans that is capability to study/learn. [8] There are two main kinds of Machine Learning methods (i) supervised learning and (ii) unsupervised learning. In supervised learning, a labeled arrange of data is used to consider or chart/plan the input data to the required output., In the unsupervised learning methods, no labeled examples are if and only if there is no concept of the output in the course of the learning process [9]. These obtain attributes of classifiers to apply ANN, but it may be any Machine Learning algorithm such as Decision Tree, Random Forest, SVM, etc. Ultimately, the outcomes are forwarded to the doctor to operate a second point of view.

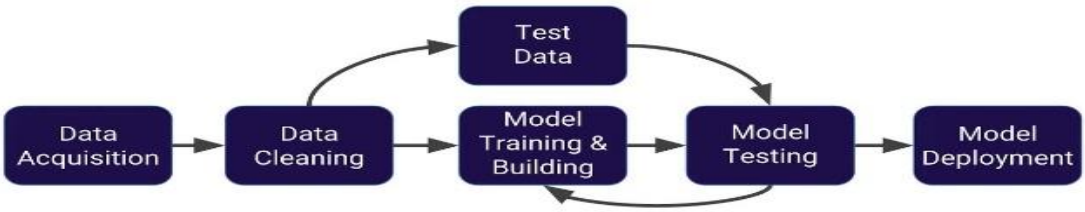

FIG I. BLOCK DIAGRAM OF MODEL BUILDING

1. Data Acquisition: In machine learning, data acquisition refers to the process of collecting raw data from various sources such as databases, sensors, files, or APIs. This step is crucial as the quality and quantity of data significantly impact the machine learning model's performance. The acquired data serves as the foundation for subsequent steps in the machine-learning pipeline.

2. Data Cleaning: Data cleaning, also known as data preprocessing, is a fundamental step in machine learning where the acquired raw data is processed to remove inconsistencies, errors, and outliers. This process involves tasks such as handling missing values, standardizing data formats, normalizing features, and removing noise. The goal of data cleaning is to ensure that the data is accurate, complete, and suitable for training machine learning models.

3. Test Data*:* Test data, in the context of machine learning, refers to a subset of the dataset that is reserved for evaluating the performance of the trained model. It is separate from the training data and assesses how well the model generalizes to unseen data. Test data helps estimate the model's performance metrics, such as accuracy, precision, recall, and F1-score, providing insights into the model's effectiveness.

4. Model Training and Building**:** Model training and building involve the process of developing a predictive or descriptive model using machine learning algorithms. This process starts with selecting an appropriate algorithm based on the problem domain and the characteristics of the data. The selected algorithm is then trained on the training data, where it learns patterns and relationships to make predictions or classifications. Model training also involves techniques such as feature engineering, hyper-parameter tuning, and cross-validation to optimize the model's performance.

5. Model Testing*:* Model testing is the phase where the trained model is evaluated using the test data to assess its performance and generalization capabilities. The model makes predictions or classifications on the test data during testing, and the outcomes are compared with the actual known labels. Performance metrics such as accuracy, precision, and recall are computed to evaluate the model's effectiveness and identify any potential issues such as over-fitting or under-fitting.

6. Model Deployment*:* Model deployment is the final stage of the machine learning process where the trained model is integrated into a production environment to make predictions or classifications on new, real-world data. This involves creating an application or service that exposes the model's functionality via an API or other interface, allowing other systems or users to interact with the model in real-time. The model deployment also includes scalability, reliability, security, and monitoring to ensure that the deployed model performs effectively in production.

C. Machine Learning Algorithms for Breast Cancer Prediction:

General Algorithm using Machine Learning to predict/detect the Breast Cancer are as follows:

Artificial Neural Network is commonly inspired by human neural structure [10]. The neural network is made up of an input layer, an output layer, and a hidden layer. These methods are used to remove the pattern that is too complicated [11]. This Algorithm is based on a data mining process, image recognition, speech recognition, and natural language processing.

Decision Tree is based on the classification and regression model of the supervised learning algorithm [12]. The dataset is split up into a minimum number of subgroups. These minimum subgroups of data can be detected with the top level of accuracy. The decision tree method is used to solve decision-related problems [13].

Support Vector Machine (SVM) is based on a supervised learning algorithm that is used for both classification and regression problems [14]. It contains a lot of conceptual and mathematical functions to work out the regression problem [15]. It gives the highest perfection speed while making predictions of huge datasets. It is a tough machine learning algorithm that is built from 2 dimensions and 3 Dimensions [16], [17].

Random Forest algorithm [18] is also based on supervised learning [19] that is used to work out classification as well as regression issues. It is a structure block of machine learning that is used for the prediction of the latest data on the root of the last dataset [16].

II. EXPERIMENT

In sequence to compare the conduct of SVM, Logistic Regression, and Random Forest, we conducted an experiment that focused on assessing both the effectiveness and the efficiency of the algorithms [20].

1. Experiment Environment: All experiments on the classifiers described in this paper were conducted using libraries from the Python machine-learning environment. Python contains a collection of machine-learning algorithms for data preprocessing, classification, regression, clustering, and association rules. Machine learning techniques implemented in Python are applied to various real-world problems. The program offers a well-defined framework for experimenters and developers to build and evaluate their models.

2. Breast cancer dataset: The Wisconsin Breast Cancer (original) datasets20 from the UCI Machine Learning Repository are used in this study [25]. Breast-cancer-Wisconsin has 569 instances (Benign: 357 Malignant: 212), 2 classes (37.25% malignant and 62.74% benign), and 30 integer-valued attributes.

III. EXPERIMENT RESULT

In this section, the results of the data analysis are reported [18]. To apply our classifiers and evaluate them, we split the original set into a training sample to train the model, and a test set to evaluate it. After applying the pre-processing and preparation methods, we try to analyze the data visually and figure out the distribution of values in terms of effectiveness and efficiency.

1. Precision is the ratio of correctly predicted positive observations to the total predicted positives. It measures the accuracy of the positive predictions made by the classifier.

2. Recall: Recall, also known as sensitivity or true positive rate, is the ratio of correctly predicted positive observations to all observations in actual class. It measures the ability of the classifier to find all the positive samples.

3. F1-score: The F1-score is the harmonic mean of precision and recall. F1-score reaches its best value at 1 and worst at 0.

4. Accuracy: Accuracy is the ratio of correctly predicted observations to the total observations. It measures the overall correctness of the classifier across all classes.

5. Mean Absolute Error (MAE): MAE measures the average absolute difference between the predicted values and the actual values. It gives an idea of the magnitude of the errors without considering their direction.

6. Root Mean Square Error (RMSE): RMSE is the square root of the average of the squared differences between the predicted values and the actual values. It penalizes large errors more heavily than MAE due to the squaring operation.

TABLE I
Comparison of accuracy measures for Logistic Regression, SVM, and Random Forest

Model	Category	Precision	Recall	F1-Score
Logistic Regression	Benign	0.97	0.97	0.97
	Malignant	0.95	0.95	0.95
SVM	Benign	0.99	0.97	0.98
	Malignant	0.95	0.98	0.97
Random Forest	Benign	0.96	0.96	0.96
	Malignant	0.93	0.93	0.93

In this section, we evaluate the precision and accuracy of the following algorithms. The results are shown in Table 2.

TABLE II
Analysis of Following Algorithm

Algorithm	Accuracy	Precision
Logistic Regression	96.49%	95.34%
SVM	97.36%	95.45%
Random Forest	94.73%	93.02%

In this section, we evaluate the Error Report of the following algorithm. The error reports are shown in Table 3.

TABLE III
Error Report of the following Algorithm

Evaluation Criteria	Logistic Regression	SVM	Random Forest
Mean Absolute Error	0.035	0.026	0.052
Root Mean Square Error	0.187	0.162	0.229

CONCLUSION

In our breast cancer prediction and diagnosis research utilizing the Wisconsin Breast Cancer dataset, we evaluated three machine learning models—Logistic Regression, SVM, and Random Forest. Among these, SVM emerged as the most effective, achieving the highest accuracy of 97.36% and precision of 95.45%, with minimal errors reflected by MAE of 0.026 and RMSE of 0.162. This indicates SVM's robustness in distinguishing between benign and malignant cases and its potential as a reliable tool for early diagnosis. The clinical implications are substantial, with the SVM model showing promise as an assisting tool for clinicians in diagnosing breast cancer, thus emphasizing the importance of validation and integration into clinical practice to improve patient outcomes.

REFERENCES

[1] Ferlay J, Ervik M, Lam F, Colombet M, Mery L, Piñeros M, et al. Global Cancer Observatory: Cancer Today. Lyon: International Agency for Research on Cancer; 2020.

[2] F.J. Shaikh, D.S. Rao, Materials Today: Proceedings "Prediction of Cancer Disease using Machine Learning Approach" 2022.

[3] Andrew J. Vickers Ph.D. : CA Cancer J Clin, "Prediction models in cancer care"; 2011.

[4] NOREEN FATIMA, LI LIU, SHA HONG, AND HAROON AHMED (Student Member, IEEE): "Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis"; 2020.

[5] National Cancer Institute: "Metastatic Cancer: When Cancer Spreads"; 2020.

[6] Yari, Yasin, Hien Nguyen and Thuy V. Nguyen. "Accuracy improvement in binary and multi-class classification of breast histopathology images" 2020 IEEE Eighth International Conference on Communications and Electronics (ICCE), IEEE 2021.

[7] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.

[8] Joseph A. Cruz, David S. Wishart: "Applications of Machine Learning in Cancer Prediction and Prognosis"; 2006.

[9] Arun Solanki, Fadi Al-Turjman, Meenu Gupta, Rachna Jain: "Cancer Prediction for Industrial IoT 4.0 - Google Books"; 2021.

[10] Y. Uzun and G. Tezel, ''Rule learning with machine learning algorithms and artificial neural networks,'' J. Seljuk Univ. Natural Appl. Sci., vol. 1, no. 2, pp. 1–11, 2012.

[11] P. Singhal and S. Pareek, ''Artificial neural network for prediction of breast cancer,'' in Proc. 2nd Int. Conf. I-SMAC (IoT Social, Mobile, Anal. Cloud)(I-SMAC), 2018, pp. 464–468.

[12] H. Sharma and S. Kumar, ''A survey on decision tree algorithms of classification in data mining,'' Int. J. Sci. Res., vol. 5, no. 4, pp. 2094–2097, 2016.

[13] T. Evgeniou and M. Pontil, ''Support vector machines: Theory and applications,'' in Advanced Course on Artificial Intelligence. Berlin, Germany: Springer, 2005, pp. 249–257.

[14] Milecia McGregor: "SVM Machine Learning Tutorial – What is the Support Vector Machine Algorithm, Explained with Code Examples";2020

[15] H. Tran, ''A survey of machine learning and data mining techniques used in multimedia system,'' Dept. Comput. Sci., Univ. Texas Dallas Richardson, Richardson, TX, USA, Tech. Rep., Sep. 2019.

[16] Y. Yang, J. Li, and Y. Yang, ''The research of fast SVM classifier method,'' in Proc. 12th Int. Comput. Conf. Wavelet Act. Media Technol. Inf. Process. (ICCWAMTIP), Dec. 2015, pp. 121–124.

[17] T. O. Ayodele, ''Types of machine learning algorithms,'' New Adv. Mach. Learn., vol. 3, pp. 19–48, Feb. 2010.

[18] Y. Li and H. Wu, ‘‘A clustering method based on K-means algorithm,’’ Phys. Procedia, vol. 25, pp. 1104–1109, Jan. 2012.

[19] Hiba Asri, Hajar Mousannif, Hassan Al Moatassime, Thomas Noel: “Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis”; 2016

[20] K. P. Bennett and O. L. Mangasarian: "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets", Optimization Methods and Software 1, 1992, 23-34]