BREAST CANCER DETECTION : A COMPARATIVE STUDY OF MACHINE LEARNING MODELS

V. Rama, Abhishek Adik Nirmal, Chaitanya Sanjiv Wakodkar, Sanyam Raj
Department of Electrical and Electronics Engineering, National Institute of Technology, Warangal, 506004
Emails: aa21eeb0b02@student.nitw.ac.in, wc21eeb0b66@student.nitw.ac.in, sr21eeb0b55@student.nitw.ac.in

ABSTRACT –

Breast cancer remains a significant global health concern, and early detection plays a crucial role in improving patient outcomes. In this study, we aim to investigate the performance of various machine learning models for breast cancer prediction. Specifically, we employ logistic regression, decision trees, random forest classifier, and support vector classifier(SVC) algorithms to analyze a dataset acquired from Kaggle's Breast Cancer Wisconsin (Diagnostic) DataSet.

The main objective of this research is to identify the algorithm that demonstrates superior performance in classifying breast cancer as benign or malignant. We evaluate the models based on their accuracy, precision, recall, and F1-score metrics, employing cross-validation techniques to ensure reliable results. Additionally, we explore feature selection methods to identify the most relevant features contributing to accurate predictions.

Our experimental results indicate that the machine learning models exhibit varying levels of performance in breast cancer prediction. Among the algorithms investigated, we found that the random forest classifier achieved the highest accuracy of 95.10% and F1-score of 0.92, indicating its potential as a reliable predictor for breast cancer. Besides, other models, such as logistic regression and SVC, also demonstrated promising results.

In conclusion, this research highlights the effectiveness of machine learning models in breast cancer detection. The findings suggest that the random forest classifier could be a robust predictive model for identifying breast cancer. These results contribute to the ongoing efforts in developing accurate and efficient tools for early breast cancer diagnosis, ultimately aiding in better patient management and improved survival rates.

Keywords : Breast Cancer, Machine Learning, Classifiers, Benign, Malignant

1. INTRODUCTION

Breast cancer is the most prevalent disease among women, according to the Centers for Disease Control and Prevention (CDC). The large variations in breast cancer survival rates are caused by numerous factors. Additionally, cancer candevelop in your breast's fat tissue or fibrousconnective tissue. In addition to frequently invading healthy breast tissue, unchecked cancer cells can also go to the lymph nodesunder the arms.

According to medical professionals, breast cancer was caused by breast cells that grew abnormally and then spread to the lymph nodes or other parts of the body. In order to prevent the effects of the following phase, it is crucial to identify and stop the proliferation of these undesirable cells as soon as feasible. The first thing a doctor does after diagnosing a tumor is to determine if it is benign or malignant because the two tumors have different treatment and preventative approaches. While malignant cells can travel to other parts of the body, benign cells are not carcinogenic and cannot do so.

The absence of prognosis models makes it challenging for doctors to develop a therapeutic strategy that might increase patient's survival time. Therefore, it takes time to design the method that produces the least amount of error in order to increase accuracy.

There was a need for a computerized diagnostic system that employed machine learning technique because the existing tests to diagnose breast cancer, such as mammography, ultrasound, and biopsy, were time-consuming. Algorithms used in this methodology help classify tumors, identify cells more precisely, and do so in a quicker manner.

2. LITERATURE SURVEY

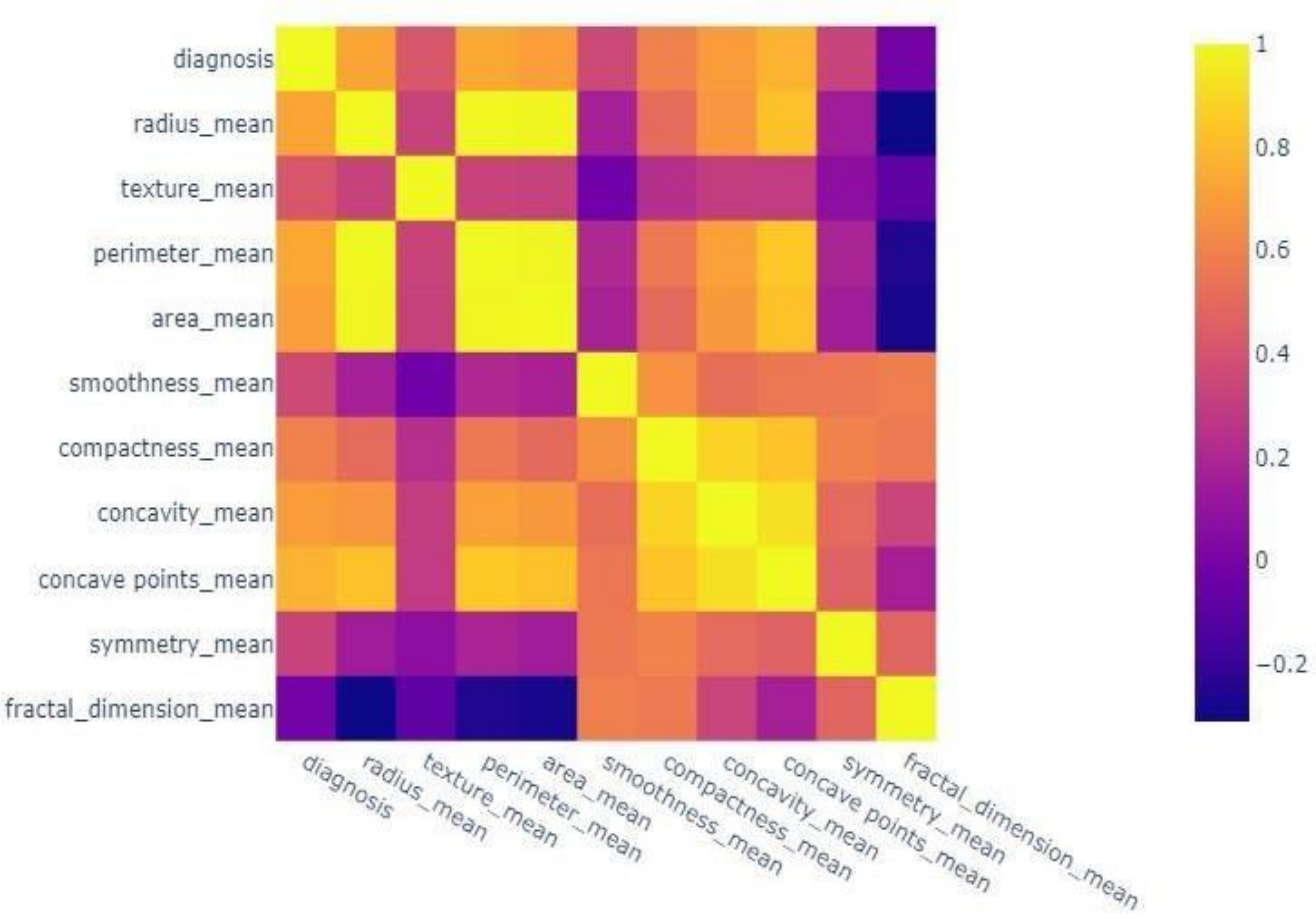
Vazirani [1] recommends two NN models, BPNN (Back Propagation Neural Network) and RBFN (Radial Basis Function). The expansion is completed with the help of a probabilistic total guideline. The assessed neural system currently provided a precision of 95% over preparing information and 95.22% over testing information, which was provisionally determined to be superior than solid neural systems.

M. Karabatak et al. [2] proposed an automated breast cancer detection technique based on AR (Association Rules) and NN (Neural Network). In this case, AR is employed to lower the measurement of intelligent classification. AR reduces the length of the input feature space from nine to four. During the testing phase, a 3-fold cross validation technique is used on the WBC database to examine the predicted system's accuracy, which is 95%. This researcher discovered that AR can be used to reduce the length of features and that the suggested AR+NN model may be utilized to create quick automatic diagnosis systems for more illnesses. In therapeutic fields where information and examination driven research are inextricably linked, fresh and distinct research bearings were considered to further accelerate the facility and natural.

In order to speed up the diagnosis process and improve the accuracy of classifying breast masses as benign or malignant, Afzan Adam [3] et al. combined a genetic algorithm with a back propagation neural network to create a computerized breast cancer diagnosis. The dataset underwent two distinct cleaning methods. While set B was trained using the standard statistical cleaning approach to identify any noisy or missing values, set A simply deleted records with missing values. Set A provided an accuracy of 100%, while Set B provided an accuracy of 83.36%. As a result, the author came to the conclusion that medical data are best retained in their original form since they have a higher accuracy rate than data that has been altered.

3. DATASET

Our dataset (Breast Cancer Wisconsin (Diagnostic) Data Set) was obtained from Kaggle. The dataset was created by photographing a needle-tip-wide breast mass with biopsy by Dr. William Wolberg,



a Wisconsin Hospital staff, then digitizing pictures by William Nick Street, a University of Wisconsin Computer Sciences Department researcher. Our dataset has 569 samples in total. Our samples are distinguished by 32 characteristics, the first of which is the sample's ID, the second is its class, and the remaining 30 are features that include diverse information about the cells.

Our samples can be classified as malignant (M) or benign (B). These are medical words for tumor cells we discussed before. The properties have no missing values. 357 of our samples are benign, with the remaining 212 being malignant .

4. METHODOLOGY

4.1 Classifiers:

Logistic Regression :

Logistic regression is a statistical model used for classification tasks. It predicts the probability of an outcome belonging to a certain class based on input variables. It uses a sigmoid function to map inputs to probabilities between 0 and 1. The model estimates coefficients using maximum likelihood estimation and uses a decision boundary to classify instances. It can handle binary or multi - class classification problems. Logistic regression is simple, interpretable and efficient but assumes a linear relationship between variables and the outcome.

Decision Tree Classifier :

A machine learning system called a Decision Tree Classifier creates a tree-like model to make predictions. A series of if-else conditions are created once the data is divided based on the input features. For classification or regression, each leaf node stands in for a continuous value or a class label. Based on the feature values, the algorithm predicts the result by following the path from the root node to a leaf node. Decision trees can overfit and require approaches like pruning or ensemble methods to improve generalization, but they are interpretable and can handle different feature types.

Random Forest Classifier:

Random Forest Classifier is a decision tree-based ensemble learning technique. It chooses subsets of data and features at random to train each tree individually. Each tree makes a forecast during prediction, and the ultimate conclusion is chosen by majority vote. Random Forests can handle a wide range of data sources, capture complicated patterns, and are resistant to outliers. They give measurements of feature significance. They are less interpretable and computationally costly than individual trees, yet they are popular for classification problems because to their accuracy and resilience.

Support Vector Classifier (SVC):

A machine learning system called a support vector classifier(SVC) chooses the most effective border to divide several classes of data points. It utilizes support vectors to determine the decision border and maximizes the margin between classes. By employing kernel functions, it can deal with both linear and nonlinear data. The system successfully handles large datasets and produces reliable classification outcomes. It concentrates on the data points that are the most instructive by determining the crucial support vectors. Support vector classifiers are strengthened by this method and are able to handle a variety of classification problems in a high-dimensional feature space.

4.2 Preprocessing:

The dataset used in this study consists of 569 samples, which have been divided into two categories: benign and malignant. The benign category comprises 357 samples, while the malignant category comprises 212 samples. Each sample in the dataset is described by 30 features, including radius mean, texture mean, perimeter mean, area mean, and others. These features provide information about various aspects of the samples, which will be used for further analysis and modelling.

4.3 Feature Extraction:

Six dependent and independent features namely - radius mean, perimeter mean, area mean, symmetry mean, compactness mean, and concave points mean are utilized to make predictions from the aforementioned 30 features.

4.4 Train-Test Split:

To evaluate the performance of the developed models accurately, the dataset should be divided into training and testing subsets. The training set will be used to train the models, while the testing set will be used to assess their generalization ability. The split ratio is 70-30.

4.5 Model training :

The 5 models are trained using the training set, and they are taught to recognize the patterns and traits that are present in the training set and signal the existence of breast cancer.

4.6 Hyper-tuning:

It involves systematically searching through different combinations of hyper-parameters to find the best configuration that maximizes the model's performance on a validation set. This process helps to enhance the model's accuracy, generalization, and overall effectiveness in making predictions.

4.7 Evaluation:

It involves measuring how well the model generalizes to new, unseen data and how accurately it makes predictions. Model evaluation typically includes metrics such as accuracy, precision, recall, F1 score, and area under the curve (AUC). By evaluating the model, we can determine its strengths, weaknesses, and limitations, allowing us to make informed decisions about its deployment and potential improvements.

5. Results and analysis:

Classification report of logistic regression:

	Precision	Recall	F1-Score	Support
B	0.90	0.96	0.93	115
M	0.92	0.84	0.88	73

Classification report of Random Forest :

	Precision	Recall	F1-Score	Support
B	0.92	0.96	0.94	115
M	0.93	0.88	0.90	73

Classification report of Decision Tree:

	Precision	Recall	F1-Score	Support
B	0.90	0.96	0.93	115
M	0.92	0.84	0.88	73

Classification report of SVC:

	Precision	Recall	F1-Score	Support
B	0.90	0.97	0.93	115
M	0.94	0.84	0.88	73

The accuracy and score given to each machine learning algorithm are displayed in the table below.

Model	**Score**	**Accuracy**
Logistic Regression	0.91	93.01%
Random Forest	0.99	95.10%
Decision Tree	1.00	91.61%
SVC	0.91	94.41%

6. Conclusion:

In order to distinguish between benign and malignant cells for breast cancer detection, this paper investigated the efficacy of various machine learning models in analyzing a particular dataset. The results showed that Random Forest had the highest accuracy of the models tested, followed by Support Vector Machine (SVC) and Logistic Regression. This indicates that given its higher performance, Random Forest might be thought of as a promising option for tasks involving the identification of breast cancer. The outcomes also suggest that machine learning models can analyze the supplied dataset successfully, highlighting their potential to help medical practitioners make an accurate diagnosis of breast cancer. These results add to the growing body of knowledge on machine learning's use in healthcare, notably in the area of breast cancer diagnosis.

7. Limitations and Future Scope:

Limitations:

I. **Dataset Bias:** The accuracy and representativeness of the dataset play a critical role in how well machine learning model's function. The generalizability of the results may be impacted by potential biases in the dataset, such as uneven class distribution or a lack of diversity, which should be addressed.

II. **Performance Metrics:** Accuracy is a frequently used performance indicator, although it might not be enough to fully assess the models' performance. Precision, recall, and F1-score are three additional metrics that can be used to evaluate the models' capacity to distinguish between benign and malignant cells more comprehensively.

III. **Limited Model Comparison:** Although the study only compares the Random Forest, SVC, Logistic Regression, and Decision Tree classifier models, there are a number of additional machine learning algorithms that could be included in the comparison to provide a more thorough assessment of the models' efficacy for detecting breast cancer.

Future Scope:

I. **Ensemble Models**: Investigate the potential benefits of ensemble models that combine the strengths of multiple algorithms, such as Random Forest and SVC, to further improve the accuracy and robustness of breast cancer detection.

II. **Transfer Learning**: Investigate the use of transfer learning techniques to leverage pre-trained models from other medical domains, such as lung cancer or skin cancer detection, and fine-tune them for breast cancer detection. This could potentially overcome limitations related to limited dataset size.

III. **Real-Time Application**: Explore the feasibility of implementing the developed models in real-time clinical settings, such as integrating them into medical imaging systems or mobile applications, enabling early and efficient detection of breast cancer.

IV. **Multimodal Data Integration**: Investigate the fusion of different types of data, such as genetic data, histopathology images, and clinical notes, to develop a multimodal approach for breast cancer detection. This could provide a more comprehensive analysis and potentially enhance the accuracy of the models.

References :

[1]. 8. H. R. Mhaske and D. A. Phalke, "Melanoma skin cancer detection and classification based on supervised and unsupervised learning," 2013 International conference on Circuits, Controls and Communications (CCUBE), Bangalore, 2013, P. 1-5.

[2] . Y. Tsehay et al., "Biopsy-guided learning with deep convolutional neural networks for Prostate Cancer detection on multiparametric MRI," 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017),
Melbourne, VIC, 2017, P. 642-645

[3] . Afzan Adam1 Khairuddin Omar2 "Computerized Breast Cancer Diagnosis with Genetic Algorithm and Neural Network"
fitt.mmu.edu.my/caiic/papers/afzaniCAIET.pdf

https://www.kaggle.com/datasets/uciml/breast-cancer- wisconsin-data