Breast Cancer Diagnosis & Analysis Using Machine Learning – Decision Tree & Random Forest

Abstract

Breast cancer continues to be a major global health concern, highlighting the importance of early detection by a precise diagnosis. In this paper, we examine machine learning methods for breast cancer diagnosis. Using a dataset from the UCI Machine Learning Repository, which includes 569 patient records with 31 attributes describing features of cell nuclei taken from images of breast masses, we research how well decision tree and random forest models succeed in identifying breast cancer as benign or malignant. Preprocessing techniques, feature selection strategies, model training, along with evaluation are all part of our research. The decision tree has been surpassed by the random forest, which both models have high accuracy rates according to the results. Understanding the basic factors affecting breast cancer diagnosis can be improved by the insights obtained by feature significance analysis. Future research directions and limitations are also discussed, highlighting the potential impact of advanced diagnostic tools in breast cancer care.

Introduction

Around the world, breast cancer is a serious threat to women's health, making detection and treatment methods important for in a clinical perspective. Because it allows for early treatments and targeted techniques, early diagnosis is key to improving the results for patients. By programmed analysis of medical data, machine learning techniques present alternatives for improving the diagnosis of breast cancer. In this work, we use an extensive dataset from the UCI Machine Learning Repository to explore the use of machine learning models for breast cancer diagnosis.

Problem Statement

Accurate diagnosis and immediate care are important for improving the future outlook of breast cancer patients. On the other hand, patient records and data from medical imaging evaluation by computer is costly and highly susceptible to human mistake. Early detection and action could be improved by digitizing the diagnostic process and utilizing machine learning algorithms. The purpose of this study is to assess how well machine learning models perform to recognize cases of breast cancer as benign or malignant using features of the cell nuclei that were taken from images of breast masses.

Dataset Description

The dataset used in this work includes 569 patient records, each of which is described by 31 attributes which focus on elements of the cell nuclei that were taken from photographs of breast masses. Numerical metrics including radius, texture, area, perimeter, smoothness, compactness, concavity, symmetry, and fractal dimension are examples of these attributes or main attributes that have been used in this research. An important resource for creating and evaluating machine learning models for breast cancer diagnosis is the dataset. The dataset, guaranteeing its quality and relevance to the research's goals.

Literature Review

The diagnosis of breast cancer has been investigated with a selection of methodologies in the previous years, including machine learning and traditional statistical techniques. Research has examined the application of combination approaches, decision trees, support vector machines, and neural networks in determining the type of instances of breast cancer according to histological findings. While some studies indicate impressive accuracy rates, others have brought close attention to problems with feature selection, model clarity, and generalization to a range of patient populations. In order to find useful research on breast cancer diagnosis using machine learning techniques, we searched credible internet papers like Google Scholar in addition to peer-reviewed publications. Six papers in all, that includes a range of approaches and actual findings, were selected for review. The research conducted offered helpful explanations of innovative techniques and challenges in breast cancer diagnosis, informing our approach in this study.

Methodology

In this work, we used a dataset. It included the medical records of 569 individuals, each of which had 31 attributes that focused on parts of the cell nuclei that we identified from images of breast masses. Our inquiry on the use of machine learning techniques for breast cancer diagnosis was based on this dataset (it is important to know different dataset or additional information on that dataset will change the result). Preprocessing steps were performed out to ensure the consistency and quality of the data. This involved handling missing values, transforming categorical variables into numerical representation(mainly 0 and 1), and applying Min-Max scaling to features. Additionally, the preparation step was made easier because no missing values were found in the dataset. Finding useful features for model training was made easier by feature selection. Variance filtering and univariate feature selection were the two methods used. Features with little variation were reduced by variance filtering, producing a more refined set of 26 features. Univariate feature selection, utilizing a chi-squared test, identified 10 features with the strongest individual relationship with the target variable.

We selected random forests and decision trees as our two machine learning methods for the classification challenge. Decision trees got selected because of the speed of interpretation and simplicity, while random forests provided greater accuracy and strength by utilizing ensemble learning. To make training and evaluating the model easier, the dataset was divided into training (80%) and testing (20%) sets. On the testing set, models for decision trees and random forests were evaluated using metrics like F1-score, accuracy, precision, and recall. The models were trained on the training set. To make sure the results were accurate, cross-validation techniques were used. A feature importance analysis was performed to learn more about the major variables affecting the diagnosis of breast cancer. Significant findings into the fundamental causes of the categorization task were provided by this investigation, which helped with the interpretation of model predictions.

Results

Our analysis's findings show that decision tree and random forest models perform brilliantly when it comes to identifying cases of breast cancer. The decision tree model has a 95% overall accuracy rate and great recall and precision rates for both benign and malignant patients.(fig 2) Performing better than the decision tree, the random forest model achieves an overall accuracy of 96% and shows robustness when managing complex relationships between attributes and the target variable(fig3). In addition, the significance of features analysis helped understand model predictions by offering useful data about the key variables influencing breast cancer detection.

The sensitive nature of the data and its implications for human life are the primary reasons for using two feature selection techniques: variance filtering and univariate feature selection. Diagnosing breast cancer is an important medical procedure that has a major impact on treatment choices and results for patients. Because of this, it is important to make sure that the features chosen for model training are understandable, reliable, and helpful. Our objective to put in several feature selection strategies was to improve the reliability and consistency of our results. By selecting variables with little variation, variance filtering makes it a point that only important qualities were kept for further analysis. On the other hand, univariate feature selection gave us the opportunity to identify features that had the strongest individual correlation with the target variable, which added to our understanding into the discriminatory power of each attribute. (fig 1)

Using an organized method to feature selection was important because of the seriousness of the task at hand and the potential consequences of mistake. We wanted to maximize the prediction performance of our models while minimizing the possibility of using irrelevant characteristics, so we used a number of strategies. This methodology is in line with machine learning best practices and shows our dedication to strong and open research methods. Our main goal is to create trustworthy and accurate models for diagnosing breast cancer, which will allow doctors to make better decisions while improving patient outcomes.

Discussion

The results of this research explain how machine learning models can be used in order to analyze medical imaging data to improve the diagnosis of breast cancer. Considering the excellent accuracy rates shown by decision tree and random forest models, issues with understanding models and generalization to a variety of populations of patients still exist. Following studies efforts include optimizing model performance, exploring additional feature engineering methodologies, and verifying models on external datasets to determine how well they work in actual healthcare settings. It is important to acknowledge the research's boundaries, which include the use of a single dataset and possible errors in the model's predictions. In order to transform research findings into useful applications and include them into clinical practice, collaborations with medical professionals and experts in the field will be needed.

Conclusion

Using computerized analysis of cell nuclei features taken from breast mass pictures, this study concludes by showing the potential of machine learning techniques in improving breast cancer diagnosis. We used random forest and decision tree models to accurately categorize cases of breast cancer as benign or malignant, using data from the UCI Machine Learning Repository. The results show the importance that early detection and quick response are to bettering patient outcomes, and they also show how machine learning can help create individualized treatment methods. The models' performance is promising, but it's important to recognize their drawbacks and difficulties, such as their interpretability and ability to be applied to a variety of patient populations. To apply these discoveries to clinical practice, partnerships with healthcare providers and additional research are essential. By adding innovative diagnostic instruments into standard medical procedures. We can significantly improve patient outcomes and fight breast cancer by further developing and validating machine learning models and applying them to actual clinical situations.

Future Work

There are numerous possibilities for future study that have a chance to improve the effectiveness and relevance of machine learning methods in the diagnosis of breast cancer. At first, the goal should be to improve model performance through the analysis of advanced algorithms and feature engineering methodologies. To ensure robustness across various demographics and increase model generalization, it can be useful to include extra datasets from a variety of patient populations. Furthermore, the validation of models on outside data sets and their integration into clinical practice require collaborations with medical practitioners and field experts. Developing explainable AI methodologies and investigating transparent machine learning models can improve model transparency and promote confidence among

healthcare professionals. In addition, diverse data sources including genetic data and medical imaging can be integrated with ensemble learning techniques to offer comprehensive insights into breast cancer detection. In more general terms, the goal of upcoming research projects should be to close the gap between clinical practice and machine learning, which will eventually improve patient care and advance the battle against breast cancer.

References

Bennett, Kristin P., and O. L. Mangasarian. "Robust linear programming discrimination of two linearly inseparable sets." *Optimization Methods and Software*.

"It discusses methods for discriminating between two linearly inseparable sets, providing foundational understanding of separation techniques used in machine learning."

Guyon, Isabelle, et al. "Gene selection for cancer classification using support vector machines." *Machine Learning*.

"To highlight the importance of feature selection in cancer classification."

Wolberg, William H., et al. "Machine learning techniques to diagnose breast cancer from fine-needle aspirates." *Cancer Letters*.

"Discusses the use of decision trees and neural networks for diagnosing breast cancer/"

Bennett, Kristin P., and O. L. Mangasarian. "Neural network training via linear programming." *Advances in Neural Information Processing Systems*.

"it discusses neural network training methods via linear programming, contributing to the understanding of training machine learning models for classification tasks."

Street, W. Nick, et al. "Nuclear feature extraction for breast tumor diagnosis." *Proceedings of IS&T/SPIE's Symposium on Electronic Imaging: Science and Technology*, International Society for Optics and Photonics.

"The feature extraction for breast tumor diagnosis, informing our feature selection and extraction methods."

Delen, Dursun, et al. "Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks." *Accident Analysis & Prevention*.

"Even this totally different then our research but it discusses the application of artificial neural networks for predictive modeling."

Fig:1

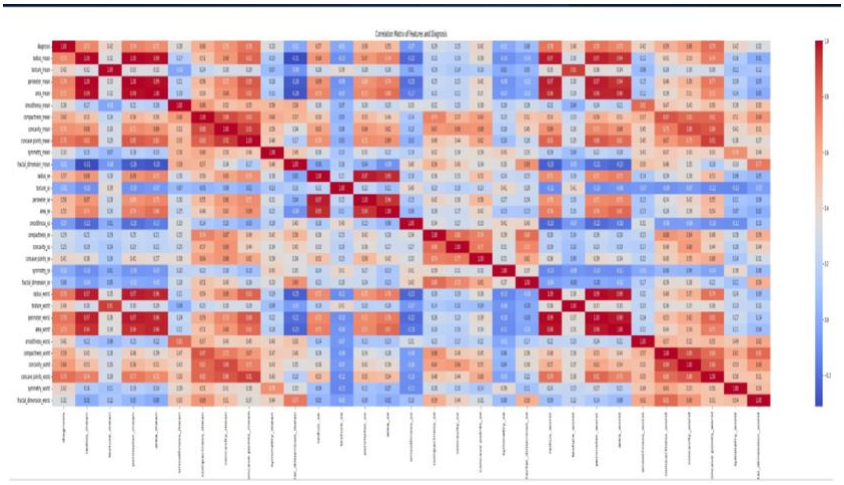

Fig:2

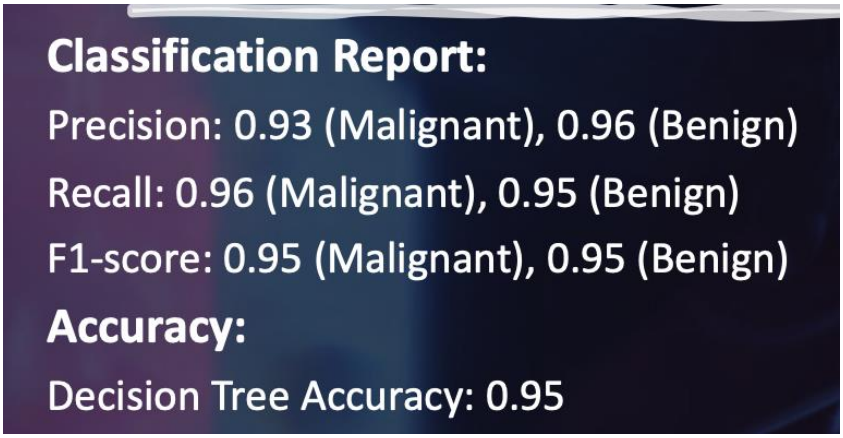

Fig:3

Classification Report:

Precision: 0.96 (Malignant), 0.98 (Benign)

Recall: 0.99 (Malignant), 0.93 (Benign)

F1-score: 0.98 (Malignant), 0.95 (Benign)

Code for the working can be found in GitHub: https://github.com/Huzifa09/Breast-Cancer