Southeast University

Department of Computer Science and Engineering

Research Proposal on:

Breast cancer classification using Machine Learning Algorithm

Supervised By:

Md. Mijanur Rahman

Assistant Professor, Department of CSE

Southeast University

Submitted By:

Mashruful Hasan

ID: 20200000000119

Batch: 54

Md Shahadat Hossain

ID: 2020000000117

Batch: 54

Ornab Biswass

ID: 2020000000095

Batch: 54

Abstract

Machine learning algorithms have been widely used in the medical field, particularly in the prediction of breast cancer. This study focuses on three clinical endpoints: the prediction of cancer susceptibility, the prediction of cancer recurrence, and the prediction of cancer survivability. Various machine learning models, such as artificial

neural networks and K-nearest neighbor algorithms, have shown promising results in accurately predicting breast cancer outcomes. Additionally, the combination of mathematical models and machine-learning algorithms has been found to improve the precision of tumor size forecasts. The early detection of breast cancer using these algorithms has the potential to significantly reduce mortality rates and improve the accuracy of cancer prediction. Furthermore, the use of machine learning algorithms in predicting breast cancer recurrence has shown promising results, particularly with the artificial neural network. However, further research is needed to understand the complex relationships between attribute values and classes within the neural network.

Introduction

Breast cancer is the second most common cancer after lung cancer and one of the main causes of death worldwide. Breast cancer is a major health concern, and early detection plays a crucial role in saving lives. Machine learning algorithms have emerged as powerful tools in predicting and diagnosing breast cancer. With the ability to capture complex data relationships, machine learning offers improved accuracy in predicting cancer recurrence, survival rates, and malignant degree of breast lesions.

Machine learning techniques, such as artificial neural networks and deep learning, have shown promising results in tumor risk assessment, lesion detection, prognosis prediction, and treatment response. The use of these algorithms for early detection is considered vital in the medical field, as they can effectively analyze complex datasets and detect critical features in breast cancer patterns. Using machine learning for breast cancer prediction allows for real-time reactions and has been widely utilized in the field.

In this document, we will explore the role of machine learning algorithms in breast cancer prediction and how they have been utilized to develop accurate prediction models for early detection.

Problem Statement:

Breast cancer is a widespread health concern, and early detection is critical for effective treatment. While various prediction models exist, developing and enhancing a reliable breast cancer prediction model using machine learning techniques is necessary. The goal is to create an accurate and interpretable model that outperforms existing studies, providing valuable insights for medical professionals in the diagnosis of breast cancer.

Objectives

1. Develop a machine learning model for predicting breast cancer susceptibility, recurrence, and survivability, surpassing the accuracy of existing studies.
2. Enhance the interpretability of the artificial neural network and K-nearest neighbor algorithms to understand the complex relationships between attribute values and classes.
3. Implement mathematical models and machine learning algorithms to more accurately predict the size of breast cancer tumors.
4. Explore the potential of deep learning in effectively solving medical problems related to breast cancer, such as tumor risk assessment, lesion detection, prognosis prediction, and treatment response.
5. Investigate the use of machine learning in early biochemical recurrence prediction and its impact on improving the accuracy of breast cancer prognosis.
6. Develop a real-time breast cancer prediction model that allows for prompt reactions and up-to-date analysis of complex datasets.
7. Contribute valuable insights to medical professionals by creating an accurate and interpretable model for the diagnosis and early detection of breast cancer.

Primary Literature Review

The primary literature review focused on the utilization of machine learning algorithms in breast cancer prediction and their performance compared to conventional regression models. The authors reviewed various machine learning, deep learning, and data mining algorithms specifically related to breast cancer prediction. They found that the most frequently used machine learning algorithm is the artificial neural network, which has shown promise in predicting the recurrence of breast They found that machine learning techniques, such as artificial neural networks and support vector machines, have shown promising results in improving the accuracy of breast cancer prediction.

Algorithms used for breast cancer prediction

In addition to the artificial neural network and support vector machine, several other machine learning algorithms have been utilized for breast cancer prediction. These include Decision Tree, Random Forest, K-Nearest Neighbors, Naive Bayes, and Logistic Regression. Each of these algorithms has shown potential in improving the accuracy of breast cancer prediction models.

Decision Tree

A Decision Tree is a machine learning algorithm that aids in decision-making by recursively partitioning a dataset into subsets based on the most influential features. It mimics a tree-like structure where each internal node represents a decision based on a specific feature, each branch denotes the outcome of that decision, and each leaf node signifies the final prediction or classification. Decision Trees are widely used for classification and regression tasks and offer interpretability, allowing researchers to analyze and comprehend the decision-making process. They excel in capturing complex relationships within data and are valuable tools for research in various fields, including medicine, where they can be applied to tasks such as breast cancer classification by considering multiple clinical and genomic features. The interpretability of Decision Trees makes them instrumental in extracting actionable insights and contributing to the understanding of complex systems.

Random Forest Classifier

The Random Forest Classifier is a powerful machine learning ensemble method designed to enhance the accuracy and robustness of predictive models. It operates by constructing a multitude of decision trees during the training phase and outputs the mode of the classes (classification) or the average prediction (regression) of the individual trees. The "random" aspect comes from two key sources of variability: random sampling of data points with replacement (bootstrapping) and random feature selection for each split in the trees. This randomness mitigates overfitting, as individual trees may excel in different aspects of the data. By aggregating their predictions, the Random Forest Classifier provides a more stable and accurate overall result, making it particularly effective for complex tasks such as breast cancer classification. Its ability to handle diverse datasets, prevent overfitting, and offer insights into feature importance makes the Random Forest a valuable tool for research, contributing to the advancement of classification methodologies, especially in the context of medical data analysis.

K-Nearest Neighbors

The K-Nearest Neighbors (KNN) algorithm is a versatile and intuitive machine learning approach used for classification and regression tasks. At its core, KNN makes predictions by identifying the majority class (for classification) or averaging values (for regression) among the K nearest data points to a given query point in a feature space. The term "nearest" is defined by a chosen distance metric, commonly Euclidean distance. KNN relies on the assumption that similar instances in the feature space exhibit similar outcomes. The simplicity and flexibility of KNN make it applicable in various research domains, including breast cancer classification. Researchers can leverage KNN to classify patients based on their similarity in clinical and genomic features. However, the algorithm's effectiveness depends on choosing an appropriate value for K and handling the impact of varying feature scales. In research contexts, KNN serves as a valuable tool for exploring patterns in data and contributing insights into the relationships between variables, fostering a deeper understanding of the underlying mechanisms in complex datasets.

Naive Bayes Classifier

The Naive Bayes Classifier is a probabilistic machine learning algorithm that leverages Bayes' theorem to make predictions based on the assumption of feature independence. Despite its simplistic nature, the algorithm has proven effective in various research applications, including text classification and medical diagnosis, such as breast cancer classification. Naive Bayes calculates the probability of a particular outcome given the presence of certain features, assuming that these features are conditionally independent, which simplifies the computations. In the context of breast cancer research, Naive Bayes can be employed to predict the likelihood of a patient belonging to a specific cancer subtype based on relevant clinical and genomic features. Its efficiency in handling high-dimensional data and fast training times make Naive Bayes particularly suitable for large datasets. Researchers benefit from its ease of implementation and interpretation, and the algorithm contributes to uncovering associations between different features in the dataset, fostering insights into the complex relationships within medical data.

Logistic Regression

Logistic Regression has been employed to model the probability of a binary outcome, making it suitable for predicting the likelihood of breast cancer occurrence and recurrence.

By exploring the performance of these diverse machine learning algorithms in breast cancer prediction, this study aims to identify the most effective and interpretable model for early detection and accurate prognosis of breast cancer.

Support Vector Machine (SVM)

The Support Vector Machine (SVM) is a robust and versatile machine learning algorithm widely utilized in research for classification and regression tasks. SVM operates by finding an optimal hyperplane that maximally separates data points belonging to different classes in a high-dimensional feature space. Its unique strength lies in its ability to handle complex relationships within data, making it particularly valuable in research contexts such as breast cancer classification. SVM seeks to establish a decision boundary that not only accurately classifies data points but also generalizes well to unseen instances. The algorithm's versatility extends to various kernel functions, allowing researchers to tailor the model to capture intricate patterns in diverse datasets, including those featuring clinical and genomic data relevant to breast cancer research. SVM's effectiveness in high-dimensional spaces, robustness against overfitting, and ability to handle non-linear relationships make it a key tool for extracting meaningful insights and contributing to advancements in the understanding of complex systems in research endeavors.

Dataset Evaluation and Dataset Selection

Initially, we will collect patient datasets from hospitals and

the internet. The dataset will contain the Breast Cancer Test. We have

planned to collect around 500 patient data.

After collecting the dataset, we will evaluate its quality and ensure that it includes relevant features and sufficient data points to train the machine learning models effectively.

Attribute

1. id
2. diagnosis
3. radius_mean
4. texture_mean
5. perimeter_mean
6. area_mean
7. smoothness_mean
8. compactness_mean
9. concavity_mean
10. concave points_mean
11. symmetry_mean
12. fractal_dimension_mean
13. radius_se
14. texture_se

15. perimeter_se
16. area_se
17. smoothness_se
18. compactness_se
19. concavity_se
20. concave points_se
21. symmetry_se
22. fractal_dimension_se
23. radius_worst
24. texture_worst
25. perimeter_worst
26. area_worst
27. smoothness_worst
28. compactness_worst
29. concavity_worst
30. concave points_worst
31. symmetry_worst
32. fractal_dimension_worst

References

1. Random Forest Modelling of High-Dimensional Mixed-Type Data for Breast Cancer Classification. Quist, J.; Taylor, L.; Staaf, J.; Grigoriadis.
2. 2018 IEEE International Conference on Computational Intelligence and Computing Research: 2018 December 13-15 : venue: Thiagarajar College of Engineering, Madurai, Tamilnadu, India.
3. 2020 IEEE EMBS Conference on Biomedical Engineering and Sciences .Comparison on Some Machine Learning Techniques in Breast Cancer Classification. Nurul Amirah Mashudi , Norulhusna Ahmad, Syaidathul Amaleena Rossli, Norliza Mohd Noor .
4. 2021 International Conference on Artificial Intelligence (ICAI) Islamabad, Pakistan, April 05-07, 2021 Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms.
5. Cancer Burden Rise to 18.1 Million New Cases and 9.6 Million Cancer Deaths in 2018, International Agency for Research on Cancer, World Health Organization, Geneva, Switzerland, 2018, p. 3.
6. Improved Machine Learning-Based Predictive Models for Breast Cancer Diagnosis Abdur Rasool 1,2,, Chayut Bunterngchit 1,3 , Luo Tiejian 1,*, Md. Ruhul Islam 4, Qiang Qu 2 and Qingshan Jiang
7. An optimized K-Nearest Neighbor based breast cancer detection Tsehay Admassu Assegie College of Engineering and Technology, Department of Computing Technology, Aksum University, Aksum, Ethiopia 2020.
8. 2020 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES) Comparison on Some Machine Learning Techniques in Breast Cancer Classification
9. Detection of Breast Cancer using Machine Learning Algorithms –A Survey Harinishree M. S., Aditya C. R.*, Sachin D. N. Dept. of Computer Science and Engineering Vidyavardhaka College of Engineering Mysuru, India 2021.