Karnataka ReddyJana Sangha(R)
VEMANA INSTITUTE OF TECHNOLOGY
Approved by AICTE-New Delhi, Affiliated to VTU-Belagavi, Recognized by Govt. of Karnataka
#1,Mahayogi Vemana Road, 3rd Block, Koramangala, Bengaluru - 34.
www.vemanait.edu.in
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Academic Year: 2021-2022

PROJECT WORK

SYNOPSIS

On

BREAST CANCER DETECTION USING MACHINE LEARNING

Bachelor of Engineering
in
Computer Science and Engineering
By

***Batch No.:* PJ21CS02**

Adithya Sunder	**:**	**1VI18CS002**
Nithin Kumar B	**:**	**1VI18CS069**
Santhosh B Rao	**:**	**1VI18CS094**
Shashank P	**:**	**1VI18CS099**

Under the supervision of

Mrs. Mary Vidya John
Assistant Professor

Approved By

Guide Sign. with date **Project Coordinator Sign. with date**

TABLE OF CONTENTS

ACKNOWLWDGEMENT

We would like to thank the Head of Department, our Project Coordinators and Project Guides for giving us this opportunity to showcase our ideas and bring them into the real world. The support from them has been immense. They have been open to questions and cleared any misconceptions or doubts we had about the subject. They have opened avenues that we hadn't considered in the past and this has made our base stronger. We aim to develop a solution for the problem at hand and make sure that no stone goes unturned.

ABSTRACT

There have been several empirical studies addressing breast cancer using machine learning and soft computing techniques. Many claim that their algorithms are faster, easier, or more accurate than others are. This study is based on genetic programming and machine learning algorithms that aim to construct a system to accurately differentiate between benign and malignant breast tumors. The aim of this study was to optimize the learning algorithm. In this context, we applied the genetic programming technique to select the best features and perfect parameter values of the machine learning classifiers. The performance of the proposed method was based on sensitivity, specificity, precision, accuracy, and the roc curves. The present study proves that genetic programming can automatically find the best model by combining feature preprocessing methods and classifier algorithms.

INTRODUCTION

Breast cancer is a prevalent cause of death, and it is the only type of cancer that is widespread among women worldwide. Many imaging techniques have been developed for early detection and treatment of breast cancer and to reduce the number of deaths , and many aided breast cancer diagnosis methods have been used to increase the diagnostic accuracy .

In the last few decades, several data mining and machine learning techniques have been developed for breast cancer detection and classification , which can be divided into three main stages: preprocessing, feature extraction, and classification. To facilitate interpretation and analysis, the preprocessing of mammography films helps improve the visibility of peripheral areas and intensity distribution, and several methods have been reported to assist in this process.

Feature extraction is an important step in breast cancer detection because it helps discriminate between benign and malignant tumors. After extraction, image properties such as smoothness, coarseness, depth, and regularity are extracted by segmentation .

Various transform-based texture analysis techniques are applied to convert the image into a new form using the spatial frequency properties of the pixel intensity variations. The common techniques are wavelet transform, fast Fourier transform (FFT) , Gabor transforms , and singular value decomposition (SVD) . To reduce the dimensionality of the feature representation, principal component analysis (PCA) can be applied. Many works have attempted to automate diagnosis of breast cancer based on machine learning algorithms. For example, Malek et al. proposed a method using the wavelet for features extraction and fuzzy logic for classification. Sun et al. studied the problem by comparing features selection methods, whereas Zheng et al. combined K-means algorithm and a support vector machine (SVM) for breast cancer diagnosis. Several works based on clustering and classification have been conducted. Another approach, introduced by Aličković and Subasi applied a genetic algorithm for feature extraction and rotation forest as a classifier.

METHODOLOGY

2.1. The Proposed Method

- **Sourcing and cleaning of datasets:** Storing the medical history of patients and using it as a training set to predict diseases.
- **Developing of neural networks:** Choosing of the specific neural network type based on complexity of datasets.
- **Choosing of algorithms**: Choose specific algorithm and test selected algorithms. After that we repeat the process with our dataset and compare with standard datasets.
- **Training of model against labeled datasets :** Training the neural network based on labeled data
- **Verification of model accuracy and efficiency:** Using a random test case to verify accuracy and efficiency of the model.

2.1.1. Stage 1: Preprocessing

Storing the medical history of patients and using it as a training set to predict diseases. There are two types of data. One is standardized data. Standardized data will give ideal outputs. The other type is data that we possess. This data gives variable results. We need to test our model against both of these to find the accuracy.

2.3.2. Stage 2: Features Selection

Usually, feature selection is applied as a preprocessing step before the actual learning. However, no algorithm can make good predictions without informative and discriminative features; therefore we need to keep the most significant features and reduce the size of the dataset as much as possible or compute for the dataset.

2.3.3. Stage 3: Machine Learning Algorithm

Usually, ensemble machine learning algorithms allow better predictive performance compared with a single model. This can be considered machine learning competition, where the winning solution was used as a model for breast cancer diagnosis.

The following heterogeneous ensembles machine learning algorithms can be used to classify the given data set: support vector machine (SVM), K-nearest neighbor (KNN) , decision tree (DT) , gradient boosting classifier (GB), random forest (RF), logistic regression (LR) , AdaBoost classifier (AB) , Gaussian Naive Bayes (GNB), and linear discriminant analysis (LDA) .

COMPARATIVE ANALYSIS

Reference	Algorithm/ Technique	Platform	Performance Metrics	Advantage	Drawback
[1]	C4.5 decision tree	Web	86.7%	Vital Status Recode (VSR) and Cause of Death (COD) is considered	Analysis does not include records with missing data
[2]	Decision Tree	Web	95.99%	Low computation time	More superior prediction can be achieved
[3]	EM Algorithm	Web	80-90%	High potential to be best predicted system	Can't be implemented in wide scale yet
[4]	Random Forest	Web	85%	False positive cases are reduced	Chances of activation function is higher due to many options

SYSTEM SPECIFICATIONS

Minimum system requirements need to be met in order to showcase this project. Machine learning projects are demanding and hence need high specifications. A processor that is i5/i7 or equivalent,4gb capable Graphics Processing Unit and RAM- 8/16 GB is required.

POSSIBLE OUTCOMES

Breast cancer rates in India have ever been increasing. There is very little knowledge about it among the rural population in particular and urban population seem to be ignorant of the idea. There need to be more awareness drives that help in destigmatizing breast cancer among society. We aim to develop a model that predicts breast cancer with good accuracy and make use of it in the real world to help millions of patients in India.

REFERENCES

- Asymmetry Analysis in Breast Cancer Detection Using Thermal Infrared Images
- Review of Electromagnetic Techniques for Breast Cancer Detection
- Comparison of data mining classification algorithms for breast cancer prediction
- Predicting Breast Cancer Survivability Using Data Mining Techniques

BIBLIOGRAPHY

- Kaggle – Online community of data scientists and machine learning practitioners
- Neural Designer- software tool for machine learning based on neural networks
- Google- Search engine
- Teachable Machine- web-based tool that makes creating machine learning models fast, easy, and accessible to everyone