Breast Cancer Risk Factor Prediction Using SVM and Extra-Tree-Based Feature Selection

Prashant Rajput,
Department of Computer Science and Engineering
Galgotias University,Greater Noida Uttar Pardesh
203201, INDIA
Email: rajputprashantsyau@gmail.com

ABSTRACT

According to the world health organization (WHO) Breast cancer is the most frequent cancer among women, impacting 2.1million women each year, and also causes the greatest number of can cerrelated deaths among women. In 2018, it is estimated that 627,000 women died from breast cancer – that is approximately 15% of all cancer deaths among women. While breast cancer rates are higher among women in more developed regions, rates are increasing in nearly every region globally. In order to improve breast cancer outcomes and survival, early detection is critical. There are two early detection strategies for breast cancer: early diagnosis and screening. Limited resource settings with weak health systems where the majority of women are diagnosed in late stages should prioritize early diagnosis programs based on awareness of early signs and symptoms and prompt referral to diagnosis and treatment. Early diagnosis strategies focus on providing timely access to cancer treatment by reducing barriers to care and/or improving access to effective diagnosis services. The goal is to increase the proportion of breast cancers identified at an early stage, allowing for more effective treatment to be used and reducing the risks of death from breast cancer. Since early detection of cancer is key to effective treatment of breast cancer we use various machine learning algorithms to predict if a tumor is benign or malignant, based on the features provided by the data.

Keyword:: : *Breast cancer; Breast Cancer diseasedetection; convolutional neural network; segmentation; early blight; late blight; deep learning*

1. INTRODUCTION

Breast cancer (BC) is one of the most common cancers among women worldwide, representing the majority of new cancer cases and cancer-related deaths according to global statistics, making it a significant public health problem in today’s society. The early diagnosis of BC can improve the prognosis and chance of survival significantly, as it can promote timely clinical treatment to patients. Further accurate classification of benign tumors can prevent patients undergoing unnecessary treatments. Thus, the correct diagnosis of BC and classification of patients into malignant or benign groups is the subject of much research. Because of its unique advantages in critical features detection from complex BC datasets, machine learning (ML) is widely recognized as the methodology of choice in BC pattern classification and forecast modeling. Some Risk Factors for Breast Cancer The following are some of the known risk factors for breast cancer. However, most cases ofbreast cancer cannot be linked to a specific cause. Talk to your doctor about your specific risk. Age. Thechance of getting breast cancer increases as women age. Nearly 80 percent of breast cancers are found in women over the age of 50. Personal history of breast cancer. A woman who has had breast cancer in one breast is at an increased risk of developing cancer in her other breast. Family history of breast cancer. A woman has a higher risk of breast cancer if her mother, sister or daughter had breast cancer, especially at a young age (before 40). Having other relatives with breast cancer may also raise the risk. Genetic factors. Women with certain genetic mutations, including changes to the BRCA1 and BRCA2 genes, are at higher risk of developing breast cancer during their lifetime. Other gene changes may raise breast cancer risk as well. Childbearing and menstrual history. The older a woman is when she hasher first child, the greater her risk of breast cancer. Also at higher risk are Role of Machine Learning In Detection of Breast Cancer

A mammogram is an x-ray picture of the breast. It canbe used to check for breast cancer in women who haveno signs or symptoms of the disease. It can also be used if you have a lump or other sign of breast cancer. Screening mammography is the type of mammogram that checks you when you have no symptoms. It can help reduce the number of deaths from breast cancer among women ages 40 to 70. But it can also have drawbacks. Mammograms can sometimes find something that looks abnormal but isn't cancer. Thisleads to further testing and can cause you anxiety.Sometimes mammograms can miss cancer when it is there. It also exposes you to radiation. You should talk to your doctor about the benefits and drawbacks of mammograms. Together, you can decide when to start and how often to have a mammogram. Now while its difficult to figure out for physicians by seeing only images of x-ray that weather the tumor is toxic or nottraining a machine learning model according to the identification of tumour can be of great help.

2. LITERATURE REVIEW

Twenty-four recent research articles have been reviewed to explore the computational methods to predict breast cancer. The summaries of them are presented below. Chaurasia et al. developed prediction models of benign and malignant breast cancer. Wisconsin breast cancer data set was used. The dataset contained 699 instances, two classes (malignant and benign), and nine integervalued clinical attributes such as uniformity of cell size. The researchers removed the 16 instances with missing values from the data set to become the data set of 683instances. The benign were 458 (65.5%) and malignant were 241 (34.5%). The experiment was analyzed by the Waikato Environment for Knowledge Analysis (WEKA). Naive Bayes, RBF Network, and J48 are the three most popular data mining algorithms were used to develop the prediction models. The researchers used 10- fold cross-validation methods to measure the unbiased estimate of the three prediction models for performance comparison purposes. The models' performance evaluation was presented based on the methods‟ effectiveness and accuracy. Experimental results showed that the Naive Bayes had gained the best performance with a classification accuracy of 97.36%; followed by RBF Network with a classification accuracy of 96.77% and the J48 was the third with a classification accuracy of 93.41%. In addition, the researchers conducted sensitivity analysis and specificity analysis of the three algorithms to gain insight into the relative contribution of the independent variables to predict survival. The sensitivity results indicated that theprognosis factor „„„Class""" was by far the most important predictor.

3. PROPOSED METHODOLOGY

Breast Cancer is one of the leading cancer developedin many countries including India. Though the endurance rate is high – with early diagnosis 97% women can survive for more than 5 years. Statistically, the death toll due to this disease has increased drastically in last few decades. The main issue pertaining to its cure is early recognition. Hence, apart from medicinal solutions some Data Science solution needs to be integrated for resolving the death causing issue. This analysis aims to observe which features are most helpful in predicting malignant or benign cancer and to see general trends that may aid us in model selectionand hyper parameter selection. The goal is to classify whether the breast cancer is benign or malignant. To achieve this i have used machine learning classification methods to fit a function that can predict the discreteclass of new input

.

3.1 PROPOSED SOLUTION & RESULTANALYSIS

In this project we will use Data Mining and Machine Learning Algorithms to detect breast cancer, based off of data. Breast Cancer (BC) is a common cancer for women around the world. Early detection of BC can greatly improve prognosis and survival chances by promoting clinical treatment to patients. We will use the UCI Machine Learning Repository for breastcancer dataset.

Url:http://archive.ics.uci.edu/ml/datasets/breast+can cer+ wisconsin+%28diagnostic%29 The dataset usedin this story is publicly available and was created byDr. William H. Wolberg, physician at the UniversityOf Wisconsin Hospital at Madison, Wisconsin, USA.To create the dataset Dr. Wolberg used fluid samples, taken from patients with solid breast masses and aneasy-to-use graphical computer program called Xcyt,which is capable of perform the analysis of cytological features based on a digital scan. The program uses a curve-fitting algorithm, to computeten features from each one of the cells in the sample,than it calculates the mean value, extreme value and standard error of each feature for the image, returning a 30 real-valuated vector Attribute Information: 1. IDnumber 2) Diagnosis (M = malignant, B = benign) 3–
32) Ten real-valued features are computed for each cell nucleus: 2. radius (mean of distances from centerto points on the perimeter) 3. texture (standard deviation of gray-scale values) 4. perimeter 5. area 6.smoothness (local variation in radius lengths) 7. compactness (perimeter² / area — 1.0) 8. concavity (severity of concave portions of the contour) 9. concave points (number of concave portions of the contour) 10. symmetry 11. fractal dimension ("coastline approximation" — 1) The mean, standarderror and "worst" or largest (mean of the three largestvalues) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

4. DATA MINING AND MACHINE LEARNING

In The term "data mining" is a misnomer, because the goal is the extraction of patterns and knowledge from large amounts of data, not the extraction (mining) of data itself. It also is a buzzword and is frequently applied to any form of large-scale data or information processing (collection, extraction, warehousing, analysis, and statistics) as well as any application of computer decision support system, including artificial intelligence (e.g., machine learning) and business intelligence. The book Data mining: Practical machine learning tools and techniques with Java[8] (which covers mostly machine learning material) was originally to be named just Practical machine learning, and the term data mining was only added for marketing reasons. Often the more general terms (large scale) data analysis and analytics – or, when referring to actual methods, artificial intelligence and machine learning – are more appropriate. In this project we use the following machine learning algorithms: Decision treealgorithms : Decision tree algorithms are successful machine learning classification techniques. They are the supervised learning methods which use information gained and pruned to improve results. Moreover, decision tree algorithms are commonly used for classification in many research, for example,in the medicine area and health issues. There are many kinds of decision tree algorithms such as ID3 and C4.5.However, J48 is the most popular decision tree algorithm. J48 is the implementation of an improved version of C4.5 and isan extension of ID3.K-nearest-neighbours (kNN) algorithm: It is a simplesupervised learning algorithm in pattern recognition.It is one of the most popular neighborhood classifiers due to its simplicity and efficiency in the field of machine learning. KNN algorithm stores all cases and classifies new cases based on similarity measures; it searches the pattern space for the ktraining tuples that are closest to the unknown tuples.The performance depends on the optimal number of neighbors (k) chosen, which is different from one data sample to another.

Support Vector Machine (SVM): It is a supervisedlearning method derived from statistical learning theory for the classification of both linear and nonlinear data. SVM classifies data into two classes over a hyperplane at the same time avoiding over- fitting the data by maximizing the margin ofhyperplane separating.
Naïve Bayes (NB) It is a probabilistic classifier: It isone of the most efficient classification algorithms based on applying Bayes' theorem with strong (naïve) independent assumptions. It assumes the value of the feature is independent of the value of any other features

4.1 DATA MINING AND MACHINE LEARNING

The term "data mining" is a misnomer, because the goal is the extraction of patterns and knowledge fromlarge amounts of data, not the extraction (mining) of data itself. It also is a buzzword and is frequently applied to any form of large-scale data or informationprocessing (collection, extraction, warehousing, analysis, and statistics) as well as any application of computer decision support system, including artificial intelligence (e.g., machine learning) and business intelligence. The book Data mining: Practical machine learning tools and techniques with method of Keras library in Python to overcome overfitting and enhance the dataset's diversity. Thecomputational cost was reduced using the smaller pixel values and the same range; for this purpose, itused scale transformation. Therefore, every pixel value was ranged from 0 to 1 using the parameter value (1./255). Images were rotated to a specific angleusing the rotation transformation; therefore, 25◦ wasemployed to rotate the images. Images can shift randomly either towards the right or left by using thewidth shift range transformation; selected a 0.1 valueof the width shift parameter.

Training, Validation and Testing ::
The entire PLD dataset was divided into three parts, training, validation and testing. The training dataset was used to train the PDDCNN model, while we utilised the validation and test dataset to evaluate the proposed model's performance. Therefore, we split the training, validation and testing datasets with 80%,10% and 10%, respectively. For the PLD dataset,3257, 403 and 403 images for training, validation andtesting were used, respectively. Different data augmentation techniques performed on the training set, i.e., rescaling, rotation, width shift, height shift, shear range, zoom range, horizontal flip, brightness and channel shift with the fill mode nearest to increasethe diversity and enhance the dataset. It would overcome the overfitting problem, thus ensuring the generalisation of the model

Objective of Research:;
The main objectives of this project is to provide solution for farmers those who grows potato to save from the economical loses & women can detect thesedisease early and apply appropriate treatment so it can save a lot of waste from the potato disease like early blight & late blight to detect the disease accurately identify what kind of disease in the potato.

Conclusion

In this project in python, we learned to build a breast cancer tumour predictor on the wisconsin dataset and created graphs and results for the same. It has been observed that a good dataset provides better accuracy. Selection of appropriate algorithms with good home dataset will lead to the development of predictionsystems. These systems can assist in proper treatmentmethods for a patient diagnosed with breast cancer. There are many treatments for a patient based on breast cancer stage; data mining and machine learning can be a very good help in deciding the line of treatment to be followed by extracting knowledgefrom such suitable databases

Reference

1. DeSantis C, Siegel R, Bandi P, Jemal A. Breast cancer statistics, 2011. CA Cancer J Clin. learning methods.” 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT) (2018): 1-3.2011;61(6):409-418. doi:10.3322/caac.20134.

2. Y. Lu, J.-Y. Li, Y.-T. Su, and A.-A. Liu, ‘‘A review of breast cancer detection in medical images,’’ in Proc. IEEE Vis. Commun. Image Process. (VCIP), Dec. 2018, pp. 1–4.

3. Turgut, Siyabend et al. “Microarray breast cancer data classification using machine Varalatchoumy and M. Ravishankar, "Comparative study of four novel approaches developed for early detection of breast cancer and its stages," 2017 International Conference on Inventive Computing and Informatics (ICICI), Coimbatore, 2017, pp. 411416, doi: 10.1109/ICICI.2017.8365384.

4. M. Ravishankar and M. Varalatchoumy, "Four novel approaches for detection of region of interest in mammograms — A comparative study," 2017 International Conference on Intelligent Sustainable Systems (ICISS), Palladam, 2017, pp. 261-265, doi: 10.1109/ISS1.2017.8389410.

5. Ammu P K and Preeja V. Article: Review on Feature Selection Techniques of DNA Microarray Data. International Journal of Computer Applications 61(12):39-44, January 20.

6. Bing Nan Li, Chee Kong Chui, Stephen Chang, S.H. Ong,Integrating spatial fuzzy clustering with level set methods for automated medical image segmentation, Computers in Biology and Medicine, Volume 41, Issue 1, 2011, Pages 1-10, ISSN 0010-4825.

7. Reddy, V. Anji, and Badal Soni. "Breast Cancer Identification and Diagnosis Techniques." Machine Learning for Intelligent Decision Science. Springer, Singapore, 2020. 49-70.

8. C. Yanyun, Q. Jianlin, G. Xiang, C. Jianping, J. Dan and C. Li, "Advances in Research of Fuzzy C-Means Clustering Algorithm," 2011 International Conference on Network Computing and Information Security, Guilin, 2011, pp. 28-31, doi: 10.1109/NCIS.2011.104.

9. V. Pali, S. Goswami and L. P. Bhaiya, "An Extensive Survey on Feature Extraction Techniques for Facial Image Processing," 2014 International Conference on Computational Intelligence and Communication Networks, 2014, pp. 142- 148,doi:

10. Bing Nan Li, Chee Kong Chui, Stephen Chang, S.H. Ong,Integrating spatial fuzzy clustering with level set methods for automated medical image segmentation, Computers in Biology and Medicine, Volume 41, Issue 1, 2011, Pages 1-10, ISSN 0010-4825.