SDM COLLEGE OF ENGINEERING AND TECHNOLOGY,
DHAVALAGIRI, DHARWAD - 580002

BREAST CANCER PREDICTION SYSTEM : BIGDATA APPROACH

BY:

NAME	**USN**
NIKITA JOSHI	2SD16IS034
POOJA GUNAGI	2SD16IS035
ANURAG DEVAGIRI	2SD16IS009

Requirements Specification Document

1 Abstract

Mammograms are used for early detection of breast cancer. The breast cancer diagnostic system extracts features from these mammograms and classifies them as malignant or benign. These systems are very helpful to doctors in detecting and diagnosing the disease faster than any other traditional methods.
In this thesis an attempt has been made to classify the extracted features from mammograms as benign or malignant by using Naive Bayes, K-NN, Multilayer Perceptron, Radial Basis Function Networks, Support Vector Machine approaches. Performance variation of the approaches by varying various parameters is studied. Finally the results are compared to find the best performing approaches.

2 Introduction

2.1 Purpose

Cancer detection has always been a major issue for the pathologists and medical practitioners for diagnosis and treatment planning.
The manual identification of cancer from microscopic biopsy images is subjective in nature and may vary from expert to expert depending on their expertise and other factors which include lack of specific and accurate quantitative measures to classify the biopsy images as normal or cancerous one. The objective is to develop a software that can guide and assist pathologists to detect the type .This software takes text inputs and records from patients and predicts the occurrence of Breast Cancer.

2.2 Scope

The research is being carried out with the following objectives:
(i) To study various big data and machine learning approaches for breast cancer diagnosis through their implementation.
(ii) To make a comparative study of the approaches.
For the purpose of research, I have considered only the classification task involved in such systems and used the existing feature space. The extraction of features from the mammographic images is not considered. The machine learning approaches that were considered here could be used for any other classification problem. I have focused mainly on breast cancer diagnosis, a medical domain problem.

2 SYSTEM CONFIGURATION

Hardware Requirements:

- 8GB RAM
- Core i3 processor
- 10GB free memory space
- Hard Disk: 1 TB

Software Requirements:

- 64bit OS
- Python 3 and above
- Operating System: Windows 10

2.5 Developer's Responsibilities

The developer is responsible for
(a) developing the system
(b) installing the software on the client's hardware
(c) conducting any user training that might be needed for using the system
(d) maintaining the system for a period of one year after installation.

3 General Description

3.1 Product Functions Overview

Detection of Cancer is made simple and time effective by the implementation of an interface that prompts the user to enter the values present in the report. On the basis of readings present in the report, values are being entered by the person prediction of the presence of cancerous tumors is done by the system. Since the KNN algorithm when applied to the cancer dataset yielded higher accuracy, This particular algorithm runs in the background to perform the prediction process.

3.2 User Characteristics

The main users of this system can either be hospital management who can maintain this software to perform the prediction process instead of sending the reports of patients to far off places to detect the occurrence of cancer. It can also be used by the individuals and enter the obtained readings from their reports and medicate accordingly.

3.3 General Constraints

The system must contain a minimum of 8GB RAM to accommodate Anaconda Navigator software to run either Spyder or Jupiter Notebook to run several different algorithms of prediction and finally select the one that provides higher accuracy.

4 Specific Requirements

4.1 Inputs and Outputs

Input : Breast Cancer Dataset(Wisconsin) from UCI Repository
Output : Prediction of Breast Cancer tumor cells and occurrence of Breast Cancer depending upon the values entered by the user.

4.2 Functional Requirements

Jupyter Notebook is a platform for accepting(reading) the cancer dataset as input performing several different algorithms and comparing their efficiencies(accuracy) to determine the appropriate algorithm with higher accuracy rate. The output is expected to have a proper and efficient algorithm working in the background and that compares with the entered values by the user.

4.3 External Interface Requirements

External interface requirements require the user to enter the values as per the readings mentioned in their reports.

4.4 System Constraints

Hardware Constraints

The system must contain a minimum of 8GB RAM to accommodate Anaconda Navigator software to run either Spyder or Jupiter Notebook to run several different algorithms of prediction and finally select the one that provides higher accuracy.

2.4 References

[1] Yoichi Murakami, Kenji Mizuguchi: Applying the Nave Bayes classifier with kernel den sity estimation to the prediction of protein-protein interaction sites. Bioinformatics 26(15): 1841-1848 (2010).

[2] George H. John and Pat Langley. Estimating continuous distributions in Bayesian classi fiers. In P. Besnard and S. Hanks, editors, Eleventh Annual Conference on Uncertainty in Artificial Intelligence, pages 338–345, San Francisco, 1995. Morgan Kaufmann Publishers.

[3] Wilbert Sibanda and Philip Pretorius. Article: Novel Application of Multi-Layer Percep trons (MLP) Neural Networks to Model HIV in South Africa using Seroprevalence Data from Antenatal Clinics. International Journal of Computer Applications 35(5):26-31, De cember 2011. Published by Foundation of Computer Science, New York, USA.

[4] Dustin Boswell: Introduction to support vector machines. http://www.work.caltech.edu/boswell/IntroToSVM.pdf.

[5] Cristianini N. and Shawe-Taylor J. 2000. An Introduction to Support Vector Machines. Cambridge University Press, Cambridge, UK.

[6] J. Park, I.W. Sandberg :Approximation and radial basis function networks .Neural Comput, 5 (1993), pp. 305-316.

[7] Domingos,P.A few useful things to know about machine learning. Commun. ACM.55 (10):78-87 (2012).

[8] G.D. Magoulas, A. Prentza, Machine learning in medical applications, in: G. Paliouras, V. Karkaletsis, C.D Spyrpoulos (Eds.), Machine Learning and its Applications, Lecture Notes in Computer Science, Springer-Verlag, Berlin, 2001, pp. 300-307.