# **Breast Cancer Detection Using Machine Learning**

A Project report submitted in partial fulfillment of the requirements for the award of the degree of

***BACHELOR’S OF TECHNOLOGY***

#### in

***Computer Science and Engineering***

**By – DONGURGIKAR SUNDER RAJ**

**MIS – 112015042**

**Semester: IV**



**Computer Science and Engineering**

**Indian Institute of Information and Technology Pune**

**Near Bopdev Ghat, Kondhwa Annexe, Yewalewadi, Pune, Maharashtra 411048**

**April 2022**

**B****ONAFIDE CERTIFICATE**

This is to certify that the project report entitled **“Breast Cancer Detection Using Machine Learning”** issubmitted by **D SUNDER RAJ** bearing the **MIS No - 112015042**, in completion of his project work under the guidance of **Dr. Chandrakant Narayan Guled** is accepted for the project report submission in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering in the Department of Computer Science and Engineering , Indian Institute of Information Technology, Pune, during the academic year 2021-22.

**Dr. Chandrakant Narayan Guled**

**Head of CSE Department**

Project viva-voce held on - 29/04/2022

**Internal Examiner External Examiner**

**ACKNOWLEDGEMENT**

This project would not have been possible without the help and cooperation of many. I would like to thank the people who helped me directly and indirectly in the completion of this project work.

First and foremost, I would like to express my gratitude to our beloved director,

**Dr. Anupam Shukla**, for providing his kind support in various aspects.

I would like to express my gratitude to my project guide **Dr. Chandrakant Narayan Guled**, Head of Department CSE, excellent guidance, encouragement, inspiration, constant and timely support throughout this project.

I would also like to thank all the faculty members in the Dept. of CSE and my classmates for their steadfast and strong support and engagement with this project.

Index

1. **Abstract**
2. **Introduction**
3. **Objective**
4. **Review of Literature/Related Work**
5. **Methodology**
6. **DATA FLOW DIAGRAM**
7. **Conclusion and Future Work**
8. **Reference**

**Abstract**

Breast Cancer 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 issue in today's society. The early diagnosis of BC can improve the prognosis and chance of survival significantly, as it can promote timely clinical treatments to patients. This project aims to detect the type of Breast Cancer (Malignant or Benign classes) using K-Nearest Neighbors (K-NN) a Machine Learning Algorithm by taking cell parameters as input. The quality of the results depends largely on the distance and

value of parameter “k” which represents the number of nearest neighbors. This project also

aims to achieve maximum accuracy in detection of Breast Cancer using BC data sets. This study is conducted on Wisconsin breast cancer dataset (WBCD) obtained by the university of Wisconsin Hospital from UCI repository

**Introduction**

Breast cancer (BC) is the most common cancer in women, affecting about 10% of all women at some stages of their life. Over the past few decades, ML techniques have been widely used in intelligent healthcare systems, especially for breast cancer (BC) detection and diagnosis. The breast is made up of different tissue, ranging from very fatty tissue to very dense tissue. Within this tissue is a network of lobes. Each lobe is made up of tiny, tube-like structures called lobules that contain milk glands. Tiny ducts connect the glands, lobules, and lobes, carrying milk from the lobes to the nipple. The nipple is located in the middle of the areola, which is the darker area that surrounds the nipple. Blood and lymph vessels also run throughout the breast. Blood nourishes the cells. The lymph system drains bodily waste products. The lymph vessels connect to lymph nodes, the tiny, bean-shaped organs that help fight infection. It has been identified that one of the leading causes of death in developing countries as breast cancer. Earlier detection of cancer can reduce the death rate and reduce the treatment phase. As breast cancer is a medical scenario it requires a medical diagnosis to detect it. To make the process must simpler computer aided tools are adopted. The objective of this work is to classify the given data set into different types of breast cancer i.e. Benign or Malignant.

A mass of abnormal tissue is known as tumor. Breast cancer tumors are classified into two types,1. Benign, those that are non-cancerous, and Malignant, those that are cancerous.

**Benign Tumors**

: Generally, these tumors are not aggressive toward surrounding tissue, they may continue to grow occasionally.

**Malignant Tumors**

: Malignant tumors are cancerous and aggressive because they invade and damage surrounding tissue. `K-Nearest Neighbors (K-NN) is one of the most prominent classification algorithms because it is simple, effective and more accurate than many other classification algorithms. This algorithm does not require any assumption for detection

**Objective**

Even though modernization in medical science is increased in terms of technology, there is lot to achieve. Similarly Breast Cancer is one of the most discussed and researched disease in medical science. Detection and diagnosis of BC is very though due to reason that early detection of BC is difficult because symptoms do not show effects in early stages. Every year around 42,000 women die due to Breast Cancer, mostly mature women are victims of it.

The main objective behind this project is to help women by helping them diagnose Breast Cancer (BC) using their cell reports and detect about the type of breast cancer, if it non-dangerous and can be treated or dangerous so as to help them know what to measures they can take at an early stage. The project helps women to identify the type of Breast Cancer, if it is malignant or benign. By adding data from their cell report, the machine learning algorithm can help todiagnose it. The K-NN algorithm uses cell data to detect the nearest cells which are cancerous and detects the type, so they can take preventive measures early stages

To help women diagnose cancer type we are making a application which

will help women’s to identify the cancer type using machine

 learning techniques so as they can get to check in early stages of disease just by putting the credentials asked on application with the help of their reports and identify type, which can decrease the death rate due to breast cancer.

**Review of Literature/Related Work**

In Breast cancer detection field there are many studies with many concepts and methods were used to be a useful methods. Many researchers present many methods and algorithms to detect the breast cancer disease; here we discussed some of them.

**A. Image Processing:**

Digital mammographic images are effortlessly available on internet which can be downloaded from the respective web address. Digital database for screening mammography(DDSM) is one of the databases available from joint efforts of Massachusetts General Hospital, Sandia National Laboratories and the University of South Florida Computer Science and Engineering Department gives approximately 2,500 case studies. The Mammographic Image Analysis Society (MIAS) Mini-Mammographic Database is another source of digital mammographic images accessible easily. The mammographic image analysis society is an organization of UK research group has created a database of digital mammogram.

**1. Mammography:**

Mammography is the most common method of breast imaging. It uses low-dose amplitude-X-rays to inspect the human breast. Cancerous masses and calcium deposits look brighter on the mammogram.

This method is good for identifying Ductal Carcinoma In Situ(DCIS) and calcifications. Currently, mammography is the gold standard method to identify early stage breast cancer before the lesions develop clinically palpable. Mammography has assisted to decrease the mortality rate by 25%-30% in screened women when compared with a control group after 5 to 7 years.

**2. MRI:**

MRI uses the hydrogen nucleus (single proton) for imaging determinations because this nucleus is abundant in water and fat. The magnetic property of the hydrogen nucleus is used to yield detailed images from any part of the body. The patient who is inspected using MRI is placed in a magnetic field and a radio frequency wave is applied to generate high contrastimages of the breast. In dynamic contrast enhanced-MRI (DCE-MRI), a contrast agent is inserted before the images are captured. This technique has been found to be more complex than mammography

**B. Artificial Neural Network (ANN):**

Artificial neural networks (ANN) or connectionist systems are computing systems vaguely inspired by the biological neural networks that constitute animal brains. The neural network itself isn't an algorithm, but rather a framework for many different machine learning algorithms to work together and process complex data inputs. Such systems "learn" to perform tasks by considering examples, generally without being programmed with any task-specificrules. For example, in image recognition, they might learn to identify images that contain cats by analyzing example images that have been manually labled as "cat" or "no cat" and using the results to identify cats in other images. The use of the ANN proved to give better diagnostic performance than the radiologists

when the network output was compared to the radiologists’ categorical

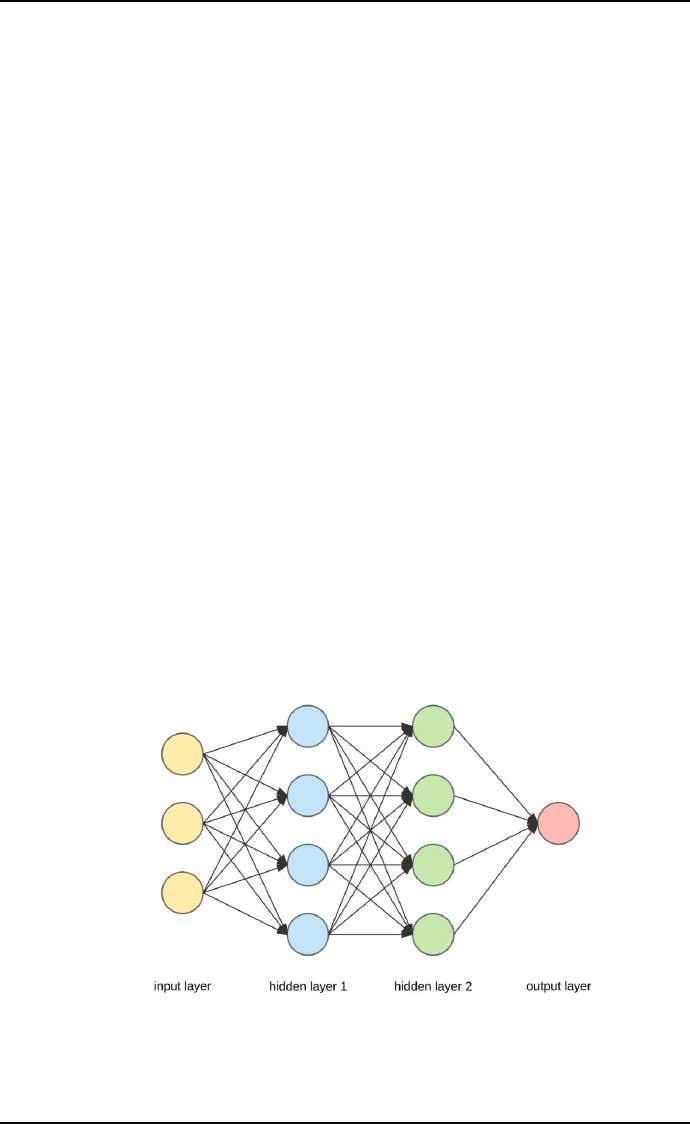
assessment. Both utilized ANN to predict malignancy using different mammographic elements as inputs. The accuracies were significant and improved by 3

 –5% compared with conventional experts’

 judgment. Thus, more than twenty years ago, ANN has been proved excellent in BC diagnosis and prognosis. Although ANN has shown a good predictor of results in pattern classification problems, the ANN is not easily explained as ANN has been considered as a

series of “black box”. By having a better understanding of ANN, a

three-phase algorithm has been proposed to unveil the ANN workings by building a weight-decay BP network, deleting insignificant connections and extracting rules by recursively discretizing the activation values of the hidden unit. The rules from this pruned network keep the accuracy as high as the rules from the standard ANN through a series of tests. After a year, based on the previous method, a modified pruned network was presented with fewer connections between each neuron, and higher accuracy .Fig: 2.1 : Artificial Neural Networks



Classification is a kind of complex optimization problem. Many ML techniques have been applied by researchers in solving this classification problem. In the following sections, a comprehensive explanation of different classification methods applied to BC will be given. We focus on k-nearest neighbor (k-NNs) techniques as they are the main methods used in BC diagnosis and prognosis. Scientists strive to find the best algorithm to achieve the most accurate classification result, however, data of variable quality will also influence the classification result. Further, the rarity of data will influence the number of algorithm applications as well. Overall, most ML techniques are first tested in open source databases. Over time, a bench mark dataset has arisen in the literature: Wisconsin breast cancer diagnosis (WBCD). There are also many other BC benchmark data sets, for instance Wisconsin Prognostic Breast Cancer Chemotherapy (WPBCC), Wisconsin Diagnostic Breast Cancer (WDBC) and so on. ML techniques that have been used on the WBCD database in BC diagnosis and prognosis show different levels of accuracy that ranged between 94.36% and 99.90%. Similarly, there are results with differently modified algorithms relating to BC databases. This review attempts to provide readers with the essential elements of BC diagnosis and prognosis using ML techniqueson WBCD. By using ML techniques to analyse the WBCD database, BC can be diagnosed accurately based on 9 attributes

**Clump thickness**

: Benign cells tend to be grouped in monolayers, while cancerous cells are often grouped in multilayers.

**Uniformity of cell size/shape**

: Cancer cells tend to vary in size and shape. That is why these parameters are valuable in determining whether the cells are cancerous or not.

**Marginal adhesion**

: Normal cells tend to stick together. Cancer cells tends to loos this ability. So loss of adhesion is a sign of malignancy.

**Single epithelial cell size**

: Is related to the uniformity mentioned above. Epithelial cells that are significantly enlarged may be a malignant cell.

**Bare nuclei**

: This is a term used for nuclei that is not surrounded by cytoplasm (the rest of the cell). Those are typically seen in benign tumours.

**Bland Chromatin**

: Describes a uniform "texture" of the nucleus seen in benign cells. In cancer cells the chromatin tend to be more coarse.

**Normal nucleoli**

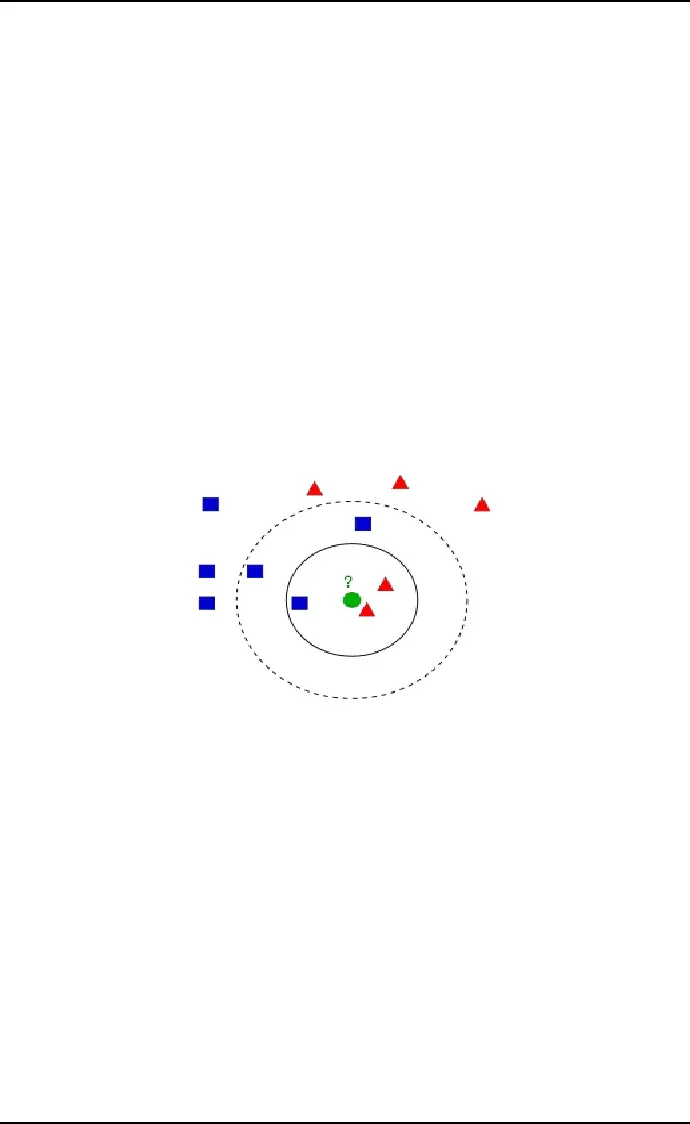
: Nucleoli are small structures seen in the nucleus. In normal cells the nucleolus is usually very small if visible at all. In cancer cells the nucleoli become more prominent, and sometimes there are more of them.

**K-Nearest Neighbor:**

The k-nearest neighbors algorithm is one of the most used algorithms in machine learning. Itis a learning method bases on instances that does not required a learning phase. The training sample, associated with a distance function and the choice function of the class based on the classes of nearest neighbors is the model developed. Before classifying a new element, we must compare it to other elements using a similarity measure. Its k-nearest neighbors are then considered, the class that appears most among the neighbors is assigned to the element to be classified. The neighbors are weighted by the distance that separate it to the new elements to classify. k-NN is one of the most central ML techniques in classification. k-NN is a non- parametric lazy learning algorithm used for classification, which classifies the objects using

their “k” nearest neighbours. k

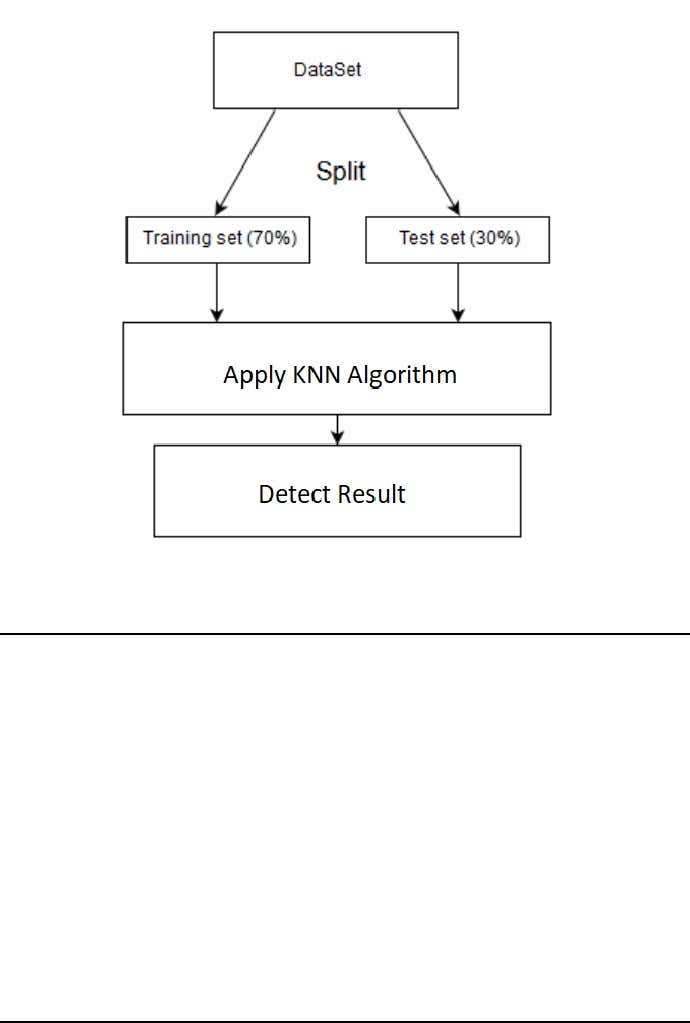
-NN only considers the neighbours around the object, not the underlying data distribution. Additionally, there is no training phase with the training data. In image, an example of k-NN structure is presented for determining BC diagnosis and prognosis if k = 3, the test sample (circle) is assigned to malignant BC (square) because there are 2 squares and only 1 triangle inside the inner circle. In image KNN Malignant and Benign Tumours



If k = 5, the test sample is assigned to benign BC (triangle).k-Nearest neighbor for breast cancer diagnosis. Green circle means the test sample, red triangle means the malignant BC and bluesquare means the benign BC. k-NN related algorithms have a number of applications in BC diagnosis and prognosis. The quality of the classification depends on the selection of k. In2000, the k-NN and fuzzy k-NN algorithms were implemented to classify the WBCD. The different values of k from 1 to 15 were considered, and the best performance was when k equalled 1.

**Proposed Methodology:**

The proposed methodology is shown in Figure. The main idea is to use k-NN algorithm to predict the class labels in the test set. Then for each classifier, the con formal predictionalgorithm is applied to calculate the non-con formality score for each prediction and use it tocalculate the confidence. The con-formal prediction algorithm is fully described.

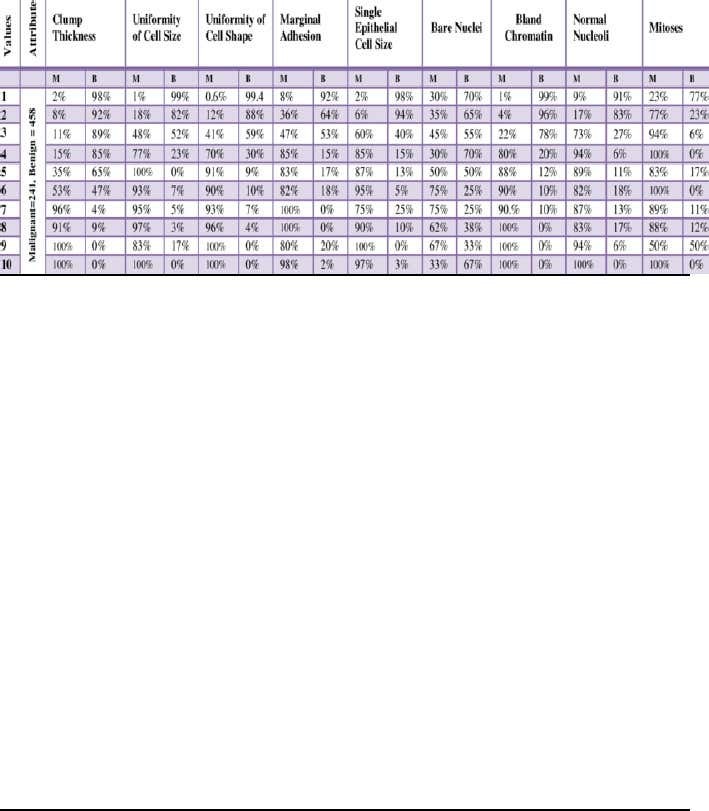


Proposed method flowchart

K-Nearest Neighbors (KNN): which is a lazy classifier that is widely used in datamining applications. In this work, we implement the KNN algorithm using Euclidian distance as a similarity measure with K=7.

**Breast Cancer Dataset**

Studies in this paper are conducted on Wisconsin breast cancer dataset (WBCD) from UCI repository. This dataset has 699 clinical cases, each one labeled as malignant (cancerous) or benign (non-cancerous). The number of malignant and benign cases are 241(24.5%) and 458(65.5%), respectively. This dataset has 16 samples (cases) with some missing values. Removing these samples from dataset decreases the sample size to 683. Every sample has 11features (Table 1). The first feature is sample id, and the last one is a class label that keeps two values: 2 for benign and 4 for malignant. Practically is proved that stratified 10-fold cross validation is one of the best methods due to low bias and variance. After dividing the dataset into ten folds, first fold is selected for testing and the combination of the other nine folds for training. The numbers of test and train samples are equal to 69 and 614 in each run. The numbers of positive (malignant) and negative (benign) train-samples are equal to 215 and 399.The standard deviations of positive and negative classes are 8.269 and 3.143, respectively. After that, every sample in the test fold is classified by finding K nearest samples from the training set. Now, the values of accuracy, sensitivity, and specificity are measured for the selected test fold. This process is repeated ten times by selecting each fold exactly once for testing. At this point, we have ten values for accuracy, sensitivity, and specificity. In order to increase the correctness of outcome, these steps are repeated 100 times by considering that the samples are randomly reassigned to the folds again Wisconsin Breast Cancer Dataset Attributes



**User Breast Cancer Data**

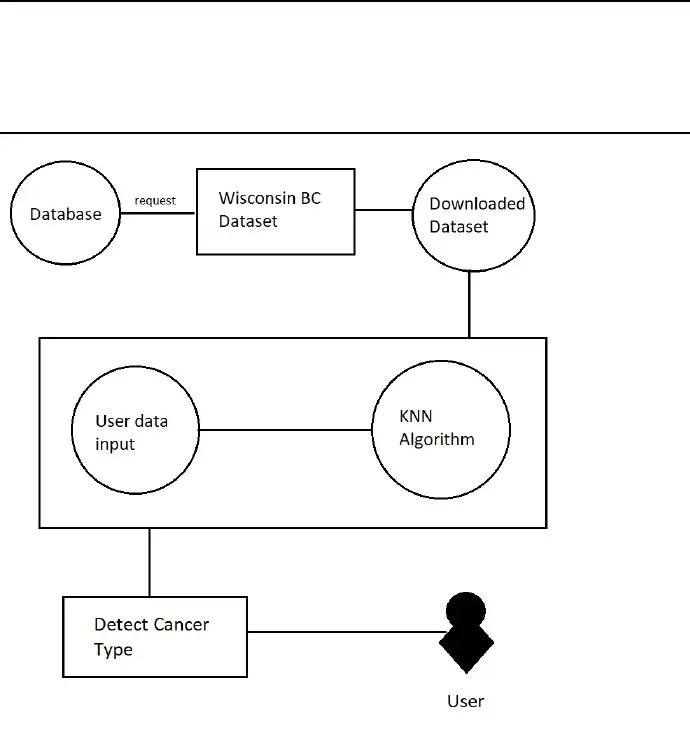
The user will input the parameters of the cell mentioned in the report. These parameters will decide the type of cancer.

**User Data Analysis**

 At first, the values given by the user specificity, the accuracy for different values of K between1 and 614 is reported to show the individual performance of the classifier on positive and negative classes. After that, the maximum values, minimum values, and standard deviations of positive and negative classes are examined to show the stability of classifier over positive and negative classes. Then the system detects the cancer type depending upon the training dataset and user dataset. There are various factors and variable that define cancer cells. The genomic data is collected with biological knowledge and stored in a database which is collective called the dataset. This

module’s purpose is to connect to dataset so that it can be processed to predict cancer. There are 32 variables that contribute to the tumor’s initiation and progression, which are recorded and stored in the dataset; the variables include radius, texture, volume, size, etc. of the cancer cells.

**DATA FLOW DIAGRAM**

****

**Conclusion and Future Work**

We have proposed Breast Cancer Detection System Based on KNN Algorithm which is a machine learning technology. This approach increases the chances of detecting breast cancer in early stages so as women can start treatment to be cured or decrease the chances of death in this cases. This will in turn increase the mortality rate and help women to spread awareness regarding breast cancer diagnosis

**References**

1. The Performance of K-Nearest Neighbors on Malignant and Benign Classes, Arash Roshanpoor, Reza Safdari [2017]

2. Cancer Prediction Using KNN,

Dheeraj.R, Hariprasath.R, Akshay Kannan.V, NishanthKumar.S

 [2018]

3. Breast Cancer Detection Using K-Nearest Neighbor Machine Learning Algorithm, Moh'dRasoul Al-Hadidi, Abdulsalam Alarabeyyat, Mohannad Alhanahnah [2016]

4. Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules Seyyid Ahmed Medjahed, Tamazouzt Ait Saadi, Abdelkader Benyettou

5. H. Zhang, T. Arslan, B. Flynn, A Single Antenna Based microwave System for BreastCancer Detection: Experimental Results, IEEE, 2013.

6. Medjahed SA, Saadi TA, Benyettou A. Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules. Int J Comput Appl. 2013;62

7. Gupta S, Kumar D, Sharma A. Data Mining Classification Techniques Applied For Breast Cancer Diagnosis And Prognosis