

**Project Report**  
**on**  
**Breast Cancer Prediction System**

**Submitted By**

**Bharti Ghildiyal-2100290140051**  
**Diksha Gupta- 2100290140060**  
**Anupam Gupta - 2100290140033**  
**Ayush Gupta - 2100290140044**

**Submitted in partial fulfilment of the**  
**Requirements for the Degree of**

**MASTER OF COMPUTER APPLICATIONS**

**Under the Supervision of**

**Dr. Sangeeta Arora**

**Associate Professor**



**Submitted to**

**Department of Computer Applications,**  
**KIET Group of Institutions,**  
**Delhi-NCR, Ghaziabad**  
**Uttar Pradesh – 201206**  
**(February 2023)**

# **CERTIFICATE**

Certified that **Bharti Ghildiyal→2100290140051,Diksha Gupta→2100290140060,Anupam Gupta→2100290140033,Ayush Gupta→2100290140044** have carried out the project work having “**Breast Cancer Prediction**” for Master of Computer Application from Dr. A.P.J. Abdul Kalam Technical University (AKTU) (formerly UPTU), Technical University, Lucknow under my supervision. The project report embodies original work, and studies are carried out by the student himself / herself and the contents of the project report do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

**Date:**

**Bharti Ghildiyal (2100290140051)**

**Diksha Gupta(2100290140060)**

**Anupam Gupta(2100290140033)**

**Ayush Gupta(2100290140044)**

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

**Date:**

**Dr. Sangeeta Arora**

**Associate Professor**

**Department of Computer Applications**

**KIET Group of Institutions, Ghaziabad**

**Signature of Internal Examiner**

**Signature of External Examiner**

**Dr. Arun Tripathi**

**Head, Department of Computer Applications**

**KIET Group of Institutions, Ghaziabad**

# ABSTRACT

Breast cancer affects the majority of women worldwide, and it is the second most common cause of death among women. However, if cancer is detected early and treated properly, it is possible to be cured of the condition. There are around 2000+ new cases of breast cancer in men each year, and about 2,30,000 new cases in women every year. Diagnosis of this disease is crucial so that women can get it treated faster. Early detection of breast cancer can dramatically improve the prognosis and chances of survival by allowing patients to receive timely clinical therapy. Machine Learning (ML) techniques can help in the detection of breast cancer. We can use these techniques to make tools for doctors that can be used as an effective mechanism. Furthermore, precise benign tumour classification can help patients avoid unneeded treatment.

This project study uses K-Nearest Neighbour (KNN) for numerical dataset, whose features are obtained from digitised image of breast mass, as to forecast and analyse cancer database. The dataset will be analysed, evaluated, and model is trained as part of the process. Finally, only numerical test data will be used for prediction.

# ACKNOWLEDGEMENTS

Success in life is never attained single handedly. My deepest gratitude goes to my thesis supervisor, **Dr. Sangeeta Arora** for her guidance, help and encouragement throughout my research work. Their enlightening ideas, comments, and suggestions.

Words are not enough to express my gratitude to **Dr. Arun Kumar Tripathi, Professor and Head, Department of Computer Applications**, for his insightful comments and administrative help at various occasions.

Fortunately, I have many understanding friends, who have helped me a lot on many critical conditions.

Finally, my sincere thanks go to my family members and all those who have directly and indirectly provided me moral support and other kind of help. Without their support, completion of this work would not have been possible in time. They keep my life filled with enjoyment and happiness.

**Bharti Ghildiyal**  
**(2100290140051)**

**Diksha Gupta**  
**(2100290140060)**

**Anupam Gupta**  
**(2100290140033)**

**Ayush Gupta**  
**(2100290140044)**



# **CONTENTS**

<b>CERTIFICATE</b>	<b>I</b>
<b>ABSTRACT</b>	<b>I</b>
<b>ACKNOWLEDGEMENT</b>	<b>II</b>
<b>LIST OF FIGURES</b>	<b>IV</b>
<b>LIST OF TABLES</b>	<b>V</b>
<b>1. INTRODUCTION</b>	
1.1. DOMAIN SPECIFIC	<b>1</b>
1.2. OBJECTIVES OF THE PROPOSED PROJECT WORK	<b>2</b>
1.3. PROBLEM DEFINITION	<b>3</b>
1.4. PROJECT FEATURES	<b>4</b>
<b>2. LITERATURE SURVEY</b>	
2.1. DOMAIN SPECIFIC	<b>5</b>
2.2. EXISTING SYSTEM	<b>8</b>
2.3. LITERATURE REVIEW	<b>9</b>
2.4. PROPOSED SYSTEM	<b>18</b>
<b>3. REQUIREMENT ANALYSIS</b>	
3.1. METHODOLOGY FOLLOWED	<b>25</b>
3.2. FUNCTIONAL AND NON FUNCTIONAL REQUIREMENTS	<b>25</b>
3.3. HARDWARE REQUIREMENTS	<b>27</b>
3.4. SOFTWARE REQUIREMENTS	<b>28</b>
3.5. FEASIBILITY STUDY	<b>29</b>
<b>4. DESIGN</b>	
4.1. DESIGN GOALS	<b>30</b>
4.2. SYSTEM ARCHITECTURE	<b>30</b>
4.3. DATA FLOW DIAGRAM	<b>31</b>
4.4. EXPECTED OUTCOME	<b>32</b>
<b>5. CODING(MODULE WISE)</b>	
5.1. PREDICTION FORM	<b>34</b>
5.2. DEPLOYING BREAST CANCER PREDICTION MODEL USING FLASK APIs	<b>38</b>
5.3. RESULT	<b>40</b>
<b>6. CONCLUSION</b>	<b>42</b>
<b>7. REFERENCES</b>	

## **LIST OF FIGURES**

<b>Fig no.</b>	<b>Fig.Name</b>	<b>Page no.</b>
2.1.	ML Architecture	6
2.2	Places ML is used	7
2.3	Digitized images of FNA, Benign, Malignant	10
2.5	K-NN	15
4.1	System Architecture	26
4.2	Data Flow Diagram	26
4.3	Expected Output	27

## **LIST OF TABLES**

<b>Table no.</b>	<b>Table Name</b>	<b>Page no.</b>
2.1	Description of features used in the dataset	11



# CHAPTER 1

## INTRODUCTION

### 1.1 DOMAIN SPECIFIC

Cancer is a disease that occurs when there are changes or mutations that take place in genes that help in cell growth. These mutations allow the cells to divide and multiply in a very uncontrolled and chaotic manner. These cells keep increasing and start making replicas which end up becoming more and more abnormal. These abnormal cells later on form a tumor. Tumors, unlike other cells, don't die even though the body doesn't need them.

The cancer that develops in the breast cells is called breast cancer. This type of cancer can be seen in the breast ducts or the lobules. Cancer can also occur in the fatty tissue or the fibrous connective tissue within the breast. These cancer cells become uncontrollable and end up invading other healthy breast tissues and can travel to the lymph nodes under the arms.

There are two types of cancers. Malignant and Benign. Malignant cancers are cancerous. These cells keep dividing uncontrollably and start affecting other cells and tissues in the body. They spread to all other parts of the body and it is hard to cure this type of cancer. Chemotherapy, radiation therapy and immunotherapy are types of treatments that can be given for these types of tumors. Benign cancer is non-cancerous. Unlike malignant, this tumor does not spread to other parts of the body and hence is much less risky than malignant. In many cases, such tumors don't really require any treatment.

Breast cancer is most commonly diagnosed in women of ages above 40. But this disease can affect men and women of any age. It can also occur when there's a family history of breast cancer. Breast Cancer has always had a high mortality rate and according to statistics, it alone accounts for about 25% of all new cancer diagnoses and 15% of all cancer deaths among women worldwide. Scientists know about the dangers of it from very early on, and hence there's been a lot of research put into finding the right treatment for it.

Breast cancer detection is done with the help of mammograms, which are basically X-rays of the breasts. It's a tool which can help detect and diagnose breast cancer. But, detection is not easy due to different kinds of uncertainties in using these mammograms. The result of a mammogram are images that can show any calcifications or deposits of calcium in the breasts. These don't always have to be cancerous. These test can also find cysts which are fluid-filled sacs that are very normal during some women's menstrual cycles — and any cancerous or noncancerous lumps.

Mammograms can cost around ₹15,000 - ₹40,000 or more based on the hospital, location, and area of body to be covered. This is very expensive and not many can afford it.

It is always best for an early diagnosis so that the treatment process can also be started earlyon.

## 1.2 OBJECTIVES OF THE PROJECT

Breast cancer is a disease which we hear about a lot nowadays. It is one of the most widespread diseases. There are around 2000+ new cases of breast cancer in men each year, and about 2,30,000 new cases in women every year. Diagnosis of this disease is crucial so that woman can get it treated faster. It is best for a correct and early diagnosis.

The main objective of this project is to help doctors analyze the huge datasets of cancer data and find patterns with the patient's data and that cancer data available. With this analysis we can predict whether the patient might have breast cancer or not.

Machine learning algorithms will help with this analysis of the datasets. These techniques will be used to predict the outcome. The outcome can be either that the cancer is benign or malignant. Benign cancer is the cancer which doesn't spread whereas malignant cancer cells spread across the body making it very dangerous.

This prediction can help doctors prescribe different medical examinations for the patients based on the cancer type. This helps save a lot of time as well as money for the patient.

## 1.3 PROBLEM DEFINITION

Over the years, a continuous evolution related to cancer research has been performed. Scientists used various methods, like early stage screening, so that they could find different types of cancer before it could do any damage. With this research, they were able to develop new strategies to help predict early cancer treatment outcome.

With the arrival of new technology in the medical field, huge amount of data related to cancer has been collected and is available for medical research. But, physicians find the accurate prediction of the cancer outcome as the most interesting yet challenging part.

For this reason, machine learning techniques have become popular among researchers. These tools can help discover and identify patterns and relationships between the cancer data, from huge datasets, while they are able to effectively predict future outcomes of a cancer type. Patients have to spend a lot of money on different tests and treatments to check whether they have breast cancer or not. These tests can take a long time and the results can be delayed. Also, after confirmation that the patient has cancer, more tests need to be done to check whether the cancer is benign or malignant.

In this project, We will be using machine learning technique (KNN) to analyze the data given in the dataset. This analysis will help us predict whether the cancer is benign or malignant.

Benign cancer is the cancer which doesn't spread whereas malignant cancer cells spread across the body making it very dangerous.

Machine learning algorithm will help with this analysis of the datasets. This technique will be used to predict the outcome. The outcome can be either that the cancer is benign or malignant. Benign cancer is the cancer which doesn't spread whereas malignant cancer cells spread across the body making it very dangerous.

This prediction can help doctors prescribe different medical examinations for the patients based on the cancer type. This helps save a lot of time as well as money for the patient.

## 1.4 PROJECT FEATURES

This project scheme was developed to reduce some amount of work for the physicians and other doctors so that they don't have to conduct many tests on the patients. It also helps minimize the amount of time and money spent by the patients undergo these tests. As everything is digitalized and based on data analysis, it takes less amount of time to get results. Based on the results, further action can be taken. It also helps researchers in the medical as well as IT sector to understand how different algorithms can predict different outcomes.

This scheme normally requires a huge amount of data about different patient history and the cancer details. This data has been collected by many doctors for a long period of time and will be used to do the analysis. This reduces the computational time required to gather all the data necessary.

## **CHAPTER 2**

### **LITERATURE SURVEY**

#### **2.1 DOMAIN SPECIFIC**

Literature survey is a very important step in the software development process. Before building any new tool, we need to check the time factor, economy and company strength. When these things are fulfilled, at that point following stages is to figure out which working framework and language can be utilized for building up the device. A lot of help is required for building the tool, internal as well as external. Senior programmers can help and provide this support to the developers from various sources like research papers, books or online websites. Before building the framework the above thought are considered for building up the proposed framework.

#### **Machine Learning**

Machine Learning is a sub category of Artificial Intelligence which allows systems to automatically learn and understand data from experience without the system being programmed to do so. It helps software applications become better at predicting outcomes for various types of problems. The basic idea of ML is to take in input data and use different algorithms to help it predict outcomes and also update these outcomes when new data is available as input.

The procedures used with machine learning are like that of data mining and predictive modeling. Both require scanning through huge amounts of data to search for any type of pattern in the data and then modify the program accordingly.

Machine Learning has been seen many a times by individuals while shopping on the internet. They are then shown ads based on what they were searching earlier on. This happens because many of these websites use machine learning to customize the ads based on user searches and this is done in real time. Machine learning has also been used in other various places like

detecting fraud, filtering of spam, network security threat detection, predictive maintenance and building news feeds.

### Machine Learning methods:

- **Supervised** learning – Here both the input and output is known. The training dataset also contains the answer the algorithm should come up with on its own. So, a labeled dataset of fruit images would tell the model which photos were of apples, bananas and oranges. When a new image is given to the model, it compares it to the training set to predict the correct outcome.
- **Unsupervised learning** – Here input dataset is known but output is not known. A deep learning model is given a dataset without any instructions on what to do with it. The training data contains information without any correct result. The network tries to automatically understand the structure of the model.
- Semi-supervised learning – This type comes somewhere between supervised and unsupervised learning. It contains both labelled and un-labelled data.
- **Reinforcement learning** – In this type, AI agents are trying to find the best way to accomplish a particular goal. It tries to predict the next step which could possibly give the model the best result at the end.

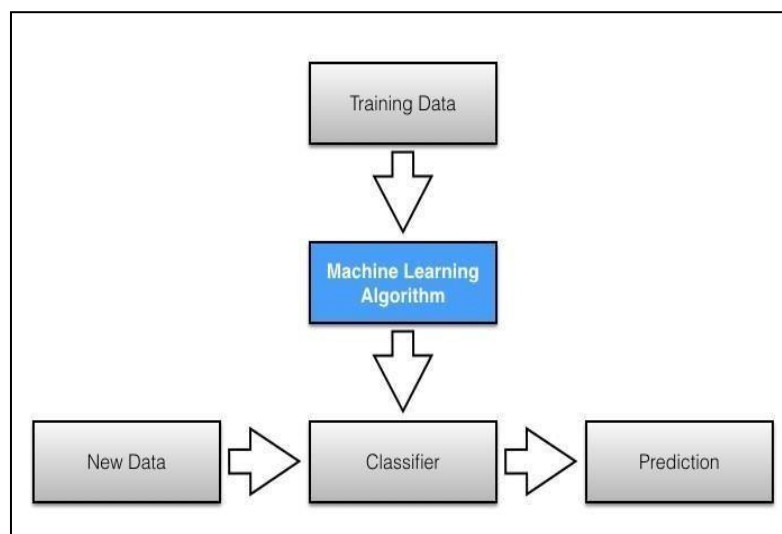


Fig 2.1 ML Architecture

## Advantages of ML:

- It gives fast and real time predictions to problems.
- Efficiently utilizes the resources.
- Helps in automation of different tasks.
- It is used a lot in different sectors of life like business, medicine, sports etc.
- Helps interpret previous behavior of model.

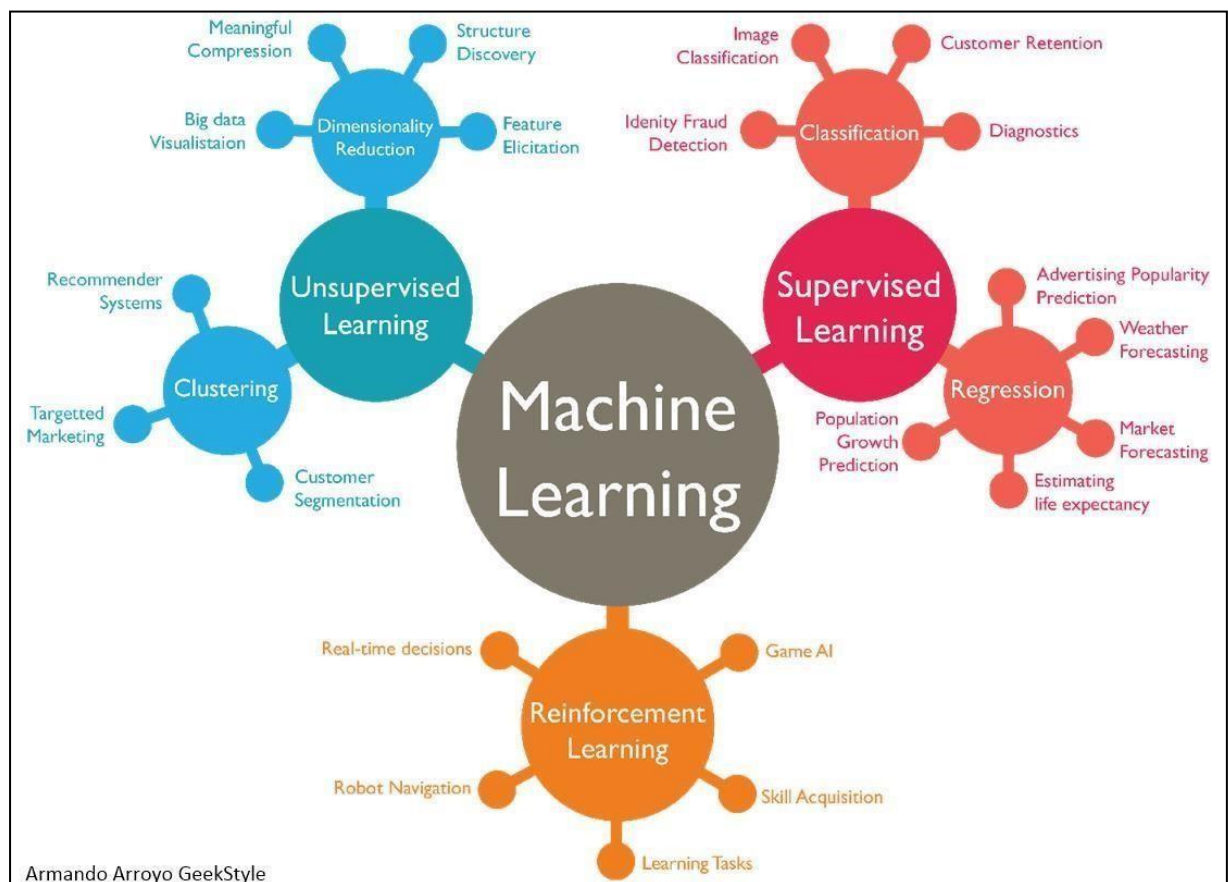


Fig 2.2 Places ML is used

## 2.2 EXISTING SYSTEM

Traditionally, the diagnosis of breast cancer and the classification of the cancer as malignant or benign was done by various medical procedures like:

- **Breast exam** – The doctor would check the breasts and lymph nodes in the armpits to check if there are any lumps or abnormalities.
- **Mammogram** – These are like X-ray of the breast. They are used to check whether there is breast cancer or not. If any issues are found, the doctor may ask the patient to take a diagnostic mammogram to check for further abnormalities.
- **Breast ultrasound** - Ultrasound uses sound waves to produce images to determine whether a new breast lump is a solid mass or a fluid-filled cyst.
- **Removing a sample of breast cells for testing (biopsy)** – This is probably the only definite way of checking if a patient has breast cancer. The doctor uses a specialized needle device guided by the X-ray or any other test to take samples of tissues from the area to be checked.
- **MRI of the breasts** - An MRI machine uses a magnet and radio waves to create pictures to see the interiors of the breast tissues.

Blood tests, CT scans and PET scans are also done to check for breast cancer.

### **Disadvantages:**

- Time consuming.
- Not completely accurate.
- Very expensive.



## 2.3 LITERATURE REVIEW

The work in this paper is focusing on various models for predicting the time of breast cancer tumor recurrence. Methods include screening the database from GLOBOCAN, CDC, and WHO health repository highlights the lethality of breast cancer, taking thousands of lives each year. However, a timely prediction of cancer can help patients to consult the doctor on time. In the past, various studies have successfully predicted the nature of the tumor to be benign or malignant and if the breast cancer tumor will reoccur or not but, no time-based models have been studied. With the help of Machine Learning, this study shows various prediction models that can be used to predict tumor reoccurrence time as accurately as 1 year. Among the 198 patients analyzed, 40% of the total patients were predicted to have breast cancer tumors reoccurring within 1st year of the diagnosis. The proposed machine learning techniques use various classification models such as Spectral clustering, DBSCAN, and k-means along with prediction models like Support Vector Machines (SVM), Decision trees, and Random Forest. The results demonstrate the ability of the model to predict the time taken by the tumor to reoccur or the time taken by the patient for full recovery with the best accuracy of 78.7% using SVM. This population-based study performed on multivariate real attributed characteristics data can therefore provide the patients a reasonable estimate about their recovery time or the time before which they should consult the doctor.[1]

Priyanka Khanna, Mridu Sahu, Bikesh Kumar Singh, Vikrant Bhatia proposed a model integrating pre-trained Convolutional neural network (CNN) with machine learning for prognosticating pathologic complete response (PCR) using breast cancer dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) prior to commencement of neoadjuvant chemotherapy (NACT). For predicting pathologic complete response (PCR) using dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) for breast cancer prior to the start of neoadjuvant chemotherapy, they presented a hybrid model integrating a pre-trained Convolutional neural network (CNN) with machine learning (NACT). In this retrospective study, 64 patients receiving NACT for invasive breast cancer are examined. Deep learning-based pre-trained CNN models ResNet-50 and ResNet-18 were used to extract features from patient visit 1 MRI images (before the initiation of NACT). Mann-Whitney U tests is used to assess features and their relevance (significance level  $p < 0.05$  and confidence interval is 95%). Furthermore, features extracted and features selected were independently given as an input to different machine learning classifiers for the prediction of response of NACT. Classification performance was assessed under different data division

protocols using accuracy, specificity, sensitivity, and area under the receiver operating characteristic curve (AUROC). The proposed model employing DCE-MRI images acquired before starting chemotherapy has considerable accuracy in classifying PCR and non-PCR patients. The efficacy of the prediction model can improve considerably on the back of a larger dataset.[2]

The work on this paper proposes and suggests a unique feature selection method based on the Eagle Strategy (ESO) Optimization, Gravitational Search Optimization (GSO) algorithm, and their hybrid algorithm. They chose this infection as their subject of investigation since the number of women with breast cancer is increasing rapidly on a global scale. After lung cancer, which affects more women than any other kind of cancer, breast cancer is the second leading cause of cancer mortality. The goal of this study is to categorize breast cancer into two groups using the benchmark feature set (Wisconsin Diagnostic Breast Cancer (WDBC)) and to choose the fewest features (feature selection) to achieve maximum accuracy. This work also provides a hybrid technique for finding important features that combines two algorithms, ESO and the GSO algorithm, while reducing insignificant characteristics (features) and complexity. Thus, this work presented a new approach for classifying breast cancer tumors. In this research, they coupled soft computing methodologies—our implemented algorithms are applied for the first time to this problem—with artificial intelligence-based machine learning strategies to create a prediction model. The efficacy of suggested technique was evaluated using WDBC breast cancer data sets, and the findings show that our proposed hybrid algorithm performs very well in breast cancer classification. They have been able to attain astonishing results with accuracy up to 98.9578%, sensitivity up to 0.9705, specificity up to 1.000, precision up to 1.000, F1-score up to 0.9696, and an AUC up to 0.9980 (close to maximum, i.e., 1.0000). The study's goal is to incorporate our findings into a valid clinical prediction system, allowing visual science specialists to make more accurate and effective judgments in the future. Furthermore, the suggested technology might be used to detect a wide range of diseases.[3]

In this study Mohammed Amine Naji, Sanaa El Filali, Kawtar Aarika, EL Habib Benlahmar, Rachida Ait Abdelouhahid, Olivier Debauche applied five machine learning algorithms: Support Vector Machine (SVM), Random Forest, Logistic Regression, Decision tree (C4.5) and K-Nearest Neighbours (KNN) on the Breast Cancer Wisconsin Diagnostic dataset, after obtaining the results, a performance evaluation and comparison is carried out between these different classifiers. The main objective of this research paper is to predict and diagnosis breast cancer, using machine-

learning algorithms, and find out the most effective with respect to confusion matrix, accuracy and precision. It is observed that Support vector Machine outperformed all other classifiers and achieved the highest accuracy (97.2%). All the work is done in the Anaconda environment based on python programming language and Scikit-learn library.[4]

Breast cancer is one of the regularly found cancer in India. In this paper the data mining techniques are used to provide the analysis for the classification and prediction algorithms. The algorithms used here are kNN classification algorithm and kNN prediction algorithm. The algorithms are used to find whether the tumour is either benign or malignant. Data sets are taken from the Wisconsin University database to find the success rate and the error rate. This is used to compare with the original success rate and the error rate.[5]

Information in digital mammogram images has been shown to be associated with the risk of developing breast cancer. Longitudinal breast cancer screening mammogram examinations may carry spatiotemporal information that can enhance breast cancer risk prediction. No deep learning models have been designed to capture such spatiotemporal information over multiple examinations to predict the risk. Saba Dadsetan, Dooman Arefan, Wendie A. Berg, Margarita L. Zuley, Jules H. Sumkin, Shandong Wu proposed a novel deep learning structure, LRP-NET, to capture the spatiotemporal changes of breast tissue over multiple negative/benign screening mammogram examinations to predict near-term breast cancer risk in a case-control setting. Specifically, LRP-NET is designed based on clinical knowledge to capture the imaging changes of bilateral breast tissue over four sequential mammogram examinations. They evaluated the proposed model with two ablation studies and compared it to three models/settings, including 1) a “loose” model without explicitly capturing the spatiotemporal changes over longitudinal examinations, LRP-NET but using a varying number of sequential examinations, and a previous model that uses only a single mammogram examination. On a case-control cohort of 200 patients, each with four examinations, experiments on a total of 3200 images show that the LRP-NET model outperforms the compared models/settings.[6]

Generative adversarial network (GAN) is able to learn from a set of training data and generate new data with the same characteristics as the training data. Based on the characteristics of GAN, this paper developed its capability as a tool of disease prognosis prediction, and proposed a prognostic model PregGAN based on conditional generative adversarial network (CGAN). The idea of

PregGAN is to generate the prognosis prediction results based on the clinical data of patients. PregGAN added the clinical data as conditions to the training process. Conditions were used as the input to the generator along with noises. The generator synthesized new samples using the noises vectors and the conditions. In order to solve the mode collapse problem during PregGAN training, Wasserstein distance and gradient penalty strategy were used to make the training process more stable. Results: In the prognosis prediction experiments using the METABRIC breast cancer dataset, PregGAN achieved good results, with the average accurate (ACC) of 90.6% and the average AUC (area under curve) of 0.946. Conclusions: Experimental results show that PregGAN is a reliable prognosis predictive model for breast cancer. Due to the strong ability of probability distribution learning, PregGAN can also be used for the prognosis prediction of other diseases.[7]

Osayba Al-Azzam, Ibrahim Shatnawi used nine machine learning classification algorithms for supervised and semi-supervised learning 1) Logistic regression; 2) Gaussian Naive Bayes; 3) Linear Support vector machine; 4) RBF Support vector machine; 5) Decision Tree; 6) Random Forest; 7) Xgboost; 8) Gradient Boosting; 9) KNN. The Wisconsin Diagnosis Cancer dataset was used to train and test these models. To ensure the robustness of the model, they have applied K-fold cross-validation and optimized hyperparameters. They have evaluated and compared the models using accuracy, precision, recall, F1-score, and ROC curves to compare and evaluate the performance and accuracy of the key supervised and semi-supervised machine learning algorithms for breast cancer prediction. The SSL has high accuracy (90%–98%) with just half of the training data. The KNN model for the SL and logistic regression for the SSL achieved the highest accuracy of 98%. Using a small sample of labeled and low computational power, the SSL is fully capable of replacing SL algorithms in diagnosing tumor type.[8]

Computer-assisted pathology analysis is an emerging field in health informatics and extremely important for effective treatment. Herein, Shallu Sharma, Sumit Kumar demonstrated the ability of the pre-trained Xception model for magnification-dependent breast cancer histopathological image classification in contrast to handcrafted approaches. The Xception model and SVM classifier with the ‘radial basis function’ kernel has achieved the best and consistent performance with the accuracy of 96.25%, 96.25%, 95.74%, and 94.11% for 40X, 100X, 200X and 400X level of magnification, respectively. A comparison with existing state-of-the-art techniques has been conducted based on accuracy, recall, precision, F1 score, area under ROC and precision–recall curve evaluation metrics.[9]

Breast cancer (BC) is the most commonly found disease among women all over the world. The early diagnosis of breast cancer can potentially reduce the mortality rate and increase the chances of a successful treatment. The paper which is purposed by V.Nanda Gopal, Fadi Al-Turjman, R. Kumar, L. Anand, M. Rajesh, focuses on proposing a methodology to conduct early diagnosis of breast cancer using the Internet of Things and Machine Learning. The main objective of this paper is to explore the machine learning techniques in predicting breast cancer with IoT devices. Proposed classifier resulted in 98%, 97%, 96% and 98% of precision, recall, F\_Measure and accuracy, respectively. The minimum error rate for the classifier have also been determined and found to be 34.21%, 45.82%, 64.47% of Mean Absolute Error (MAR), Root Mean Square Error (RMSE) and Relative Absolute Error (RAE), respectively. It was evident through the obtained results that the MLP classifier yields a higher accuracy with a minimum error rate when compared to LR and RF. [10]

Feature selection help select an optimal subset of features from a large feature space to achieve better classification performance. The performance of KNN classifier can be improved significantly using an appropriate subset of features from a large feature space. Recent development in General Purpose Graphics Processing Units (GPGPU) has provided us a low cost yet high-performance computing support for wide range of applications. The paper purposed by Shashank Shekhar, Nazrul Hoque, Dhruba K. Bhattacharyya presented a parallel KNN classifier powered by a mutual information based feature selection called PKNN-MIFS for effective classification of real life data. It selects an optimal subset of features from the original feature set by exploiting the mutual information concept for the estimation of feature-class and feature-feature relevance. It selects non-redundant feature by giving higher priority on feature-class relevance. The performance of the proposed PKNN-MIFS has been evaluated over several datasets and has been found to be superior to its closed counterpart.[11]

The study proposed by Abdul Majid, Safdar Ali, Mubashar Iqbal, Nabeela Kausar presents a novel prediction approach for human breast and colon cancers using different feature spaces. The proposed scheme consists of two stages: the preprocessor and the predictor. In the preprocessor stage, the mega-trend diffusion (MTD) technique is employed to increase the samples of the minority class, thereby balancing the dataset. In the predictor stage, machine-learning approaches of K-nearest neighbor (KNN) and support vector machines (SVM) are used to develop hybrid MTD-SVM and MTD-KNN prediction models. MTD-SVM model has provided the best values of

accuracy, G-mean and Matthew's correlation coefficient of 96.71%, 96.70% and 71.98% for cancer/non-cancer dataset, breast/non-breast cancer dataset and colon/non-colon cancer dataset, respectively. They found that hybrid MTD-SVM is the best with respect to prediction performance and computational cost. MTD-KNN model has achieved moderately better prediction as compared to hybrid MTD-NB (Naïve Bayes) but at the expense of higher computing cost. MTD-KNN model is faster than MTD-RF (random forest) but its prediction is not better than MTD-RF. To the best of our knowledge, the reported results are the best results, so far, for these datasets. The proposed scheme indicates that the developed models can be used as a tool for the prediction of cancer. This scheme may be useful for study of any sequential information such as protein sequence or any nucleic acid sequence.[12]

Predictive models for the occurrence of cancer symptoms by using machine learning (ML) algorithms could be used to aid clinical decision-making in order to enhance the quality of cancer care. This study aimed to develop and validate a selection of classification models that used ML algorithms to predict the occurrence of breast cancer-related lymphedema (BCRL) among Chinese women. This was a retrospective cohort study of consecutive cases that had been diagnosed with breast cancer, stages I-IV. Forty-eight variables were grouped into five feature sets. Five classification models with ML algorithms were developed, and the models' performance and the variables' relative importance were assessed accordingly. Of 370 eligible female participants, 91 had BCRL (24.6%). The mean age of this study sample was 49.89 (SD = 7.45). All participants had had breast cancer surgery, and more than half of them had had a modified radical mastectomy (n = 206, 55.5%). The mean follow-up time after breast cancer surgery was 28.73 months (SD = 11.71). Most of the tumors were either stage I (n = 49, 31.2%) or stage II (n = 252, 68.1%). More than half of the sample had had postoperative chemotherapy (n = 227, 61.4%). Overall, the logistic regression model achieved the best performance in terms of accuracy (91.6%), precision (82.1%), and recall (91.4%) for BCRL. Although this study included 48 predicting variables, we found that the five models required only 22 variables to achieve predictive performance. The most important variable was the number of positive lymph nodes, followed in descending order by the BCRL occurring on the same side as the surgery, a history of sentinel lymph node biopsy, a dietary preference for meat and fried food, and an exercise frequency of less than three times per week. These factors were the most influential predictors for enhancing the ML models' performance. This study found that in the ML training dataset, the multilayer perceptron model and the logistic regression model were the best discrimination models for predicting the outcome of BCRL, and the

k-nearest neighbors and support vector machine models demonstrated good calibration performance in the ML validation dataset. Future research will need to use large-sample datasets to establish a more robust ML model for predicting BCRL deeply and reliably.[13]

The KNN classification algorithm is one of the most commonly used algorithm in the AI field. This paper proposes two improved algorithms, namely KNNTS, and KNNTS-PK+. The two improved algorithms are based on KNNPK+ algorithm, which uses PK-Means ++ algorithm to select the center of the spherical region, and sets the radius of the region to form a sphere to divide the data set in the space. The KNNPK+ algorithm improves the classification accuracy on the premise of stabilizing the classification efficiency of KNN classification algorithm. In order to improve the classification efficiency of KNN algorithm on the premise that the accuracy of KNN classification algorithm remains unchanged, KNNTS algorithm is proposed. It uses tabu search algorithm to select the radius of spherical region, and uses spherical region division method with equal radius to divide the data set in space. On the basis of the first two improved algorithms, KNNTS-PK+ algorithm combines them to divide the data sets in space. Experiments are carried out on the new data set and the classification results were obtained. Results revealed show that the two improved algorithms can effectively improve the classification accuracy and efficiency after the data samples are cut reasonably.[14]

Breast cancer is the second prevalent type of cancer among women. Breast Ultrasound (BUS) imaging is one of the most frequently used diagnostic tools to detect and classify abnormalities in the breast. To improve the diagnostic accuracy, Computer Aided Diagnosis (CAD) system is helpful for breast cancer detection and classification. Normally, a CAD system consists of four stages: pre-processing, segmentation, feature extraction, and classification. In this paper, the pre-processing step includes speckle noise removal using Speckle Reducing Anisotropic Diffusion (SRAD). The goal of segmentation is to locate the Region of Interest (ROI) and Active contour-based segmentation is used in this work. The texture features are extracted and fed to a classifier to categorize the images as Normal, Benign and Malignant. In this work three classifiers namely K-Nearest Neighbors (KNN) algorithm, Decision tree algorithm and Random Forest classifier are used and the performance is compared based on the accuracy of classification.[15]

## 2.4 PROPOSED SYSTEM

In the proposed system we plan on using existing data of breast cancer patients which has been collected for a number of years and run different machine learning algorithms on them. These algorithms will analyze the data from the datasets to predict whether the patient has breast cancer or not and it will also tell us if the cancer is malignant or benign.

It is done by taking the patient's data and mapping it with the dataset and checking whether there are any patterns found with the data. If a patient has breast cancer, then instead of taking more tests to check whether the cancer is malignant or benign, ML can be used to predict the case based on the huge amount of data on breast cancer. This proposed system helps the patients as it reduces the amount of money they need to spend just for the diagnosis.

Also, if the tumor is benign, then it is not cancerous, and the patient doesn't need to go through any of the other tests. This saves a lot of time as well.

### **Advantages:**

- Reduces costs for medical tests.
- Does not take huge amount of time.
- Accurate.
- Intelligent way of using available data.



## MODULE DESCRIPTION

- **DATASET**

The **Wisconsin Diagnostic Breast Cancer (WDBC) dataset** which can be found in University of California, Irvin's Machine Learning Dataset Repository will be used for this project. All ML algorithms will be performed on this dataset. **The features which are in the dataset were computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. These features describe different characteristics of the cell nuclei found in the image. There are 569 data points in the dataset: 212 for Malignant and 357 for Benign.**

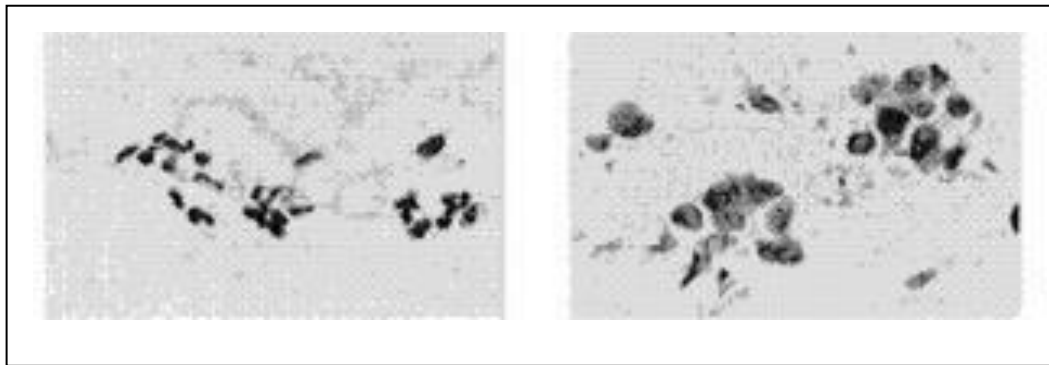


Fig 2.3 Digitized images of FNA (a) Benign (b)

MalignantThe dataset contains these features:

- **radius**
- **texture**
- **perimeter**
- **area**
- **smoothness**
- **compactness**
- **concavity**
- **concave points**
- **symmetry**

- fractal dimension

Each of these features have information on mean, standard error, and “worst” or largest (mean of the three largest values) computed. Hence, the dataset has a total of 30 features.

Radius	Mean of distances from center to points on the perimeter
Texture	Standard deviation of gray-scale values
Perimeter	The total distance between the snake points constitutes the nuclear perimeter.
Area	Number of pixel on the interior of the snake and adding one-half of the pixel in the perimeter
Smoothness	Local variation in radius length, quantified by measuring the difference between the length of a radial line and the mean length of lines surrounding it.
Compactness	$\text{Perimeter}^2 / \text{area}$
Concavity	Severity of concave portions of the contour
Concave points	Number of concave portions of the contour
Symmetry	The length difference between lines perpendicular to the major axis to the cell boundary in both directions.
Fractal dimension	Coastline approximation. A higher value corresponds to a less regular contour and thus to a higher probability of malignancy

Table 2.1 Description of features used in the dataset

## ARTIFICIAL NEURAL NETWORKS

Artificial neural networks is a very important tool used in machine learning. This technique, as the name suggests, is inspired by the brain and its activities. They were designed in a way to replicate the way humans learn. Neural networks normally consist of input as well as output layers, and sometimes a hidden layer consisting of units that change the input into something that the output layer can use. These tools help in finding different patterns which are too hard for a human to find himself. A programmer programs the machine to recognize it.

- **ML ALGORITHM**

### K-NEAREST NEIGHBOUR

This algorithm is one of the simplest Machine learning techniques. It is a lazy learning algorithm used for regression and classification. It classifies the objects using their “k” nearest neighbors. k-NN only considers the neighbors around the object, not the underlying data distribution. If  $k = 1$ , it basically assigns the unknown to the class of the nearest neighbor. If  $k > 1$ , the classification is decided by majority vote based on the k nearest neighbor prediction result.

Both for classification and regression, a useful technique can be used to assign weight to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of  $1/d$ , where  $d$  is the distance to the neighbor.

When KNN is used for classification, the output can be calculated as the class with the highest frequency from the K-most similar instances. Each instance in essence votes for their class and the class with the most votes is taken as the prediction.

If you are using K and you have an even number of classes (e.g. 2) it is a good idea to choose a K value with an odd number to avoid a tie. And the inverse, use an even number for K when you have an odd number of classes.

Ties can be broken consistently by expanding K by 1 and looking at the class of the next most similar instance in the training dataset.

Here are some things to keep in mind:

- As we decrease the value of  $K$  to 1, our predictions become less stable.
- Inversely, as we increase the value of  $K$ , our predictions become more stable due to majority voting / averaging, and thus, more likely to make more accurate predictions (upto a certain point). Eventually, we begin to witness an increasing number of errors. It is at this point we know we have pushed the value of  $K$  too far.
- In cases where we are taking a majority vote (e.g. picking the mode in a classification problem) among labels, we usually make  $K$  an odd number to have a tiebreaker.

Advantages:

- The algorithm is simple and easy to implement.
- There's no need to build a model, tune several parameters, or make additional assumptions.
- The algorithm is versatile. It can be used for classification, regression, and search (as we will see in the next section).

Disadvantage

- The algorithm gets significantly slower as the number of examples and/or predictors/independent variables increase.

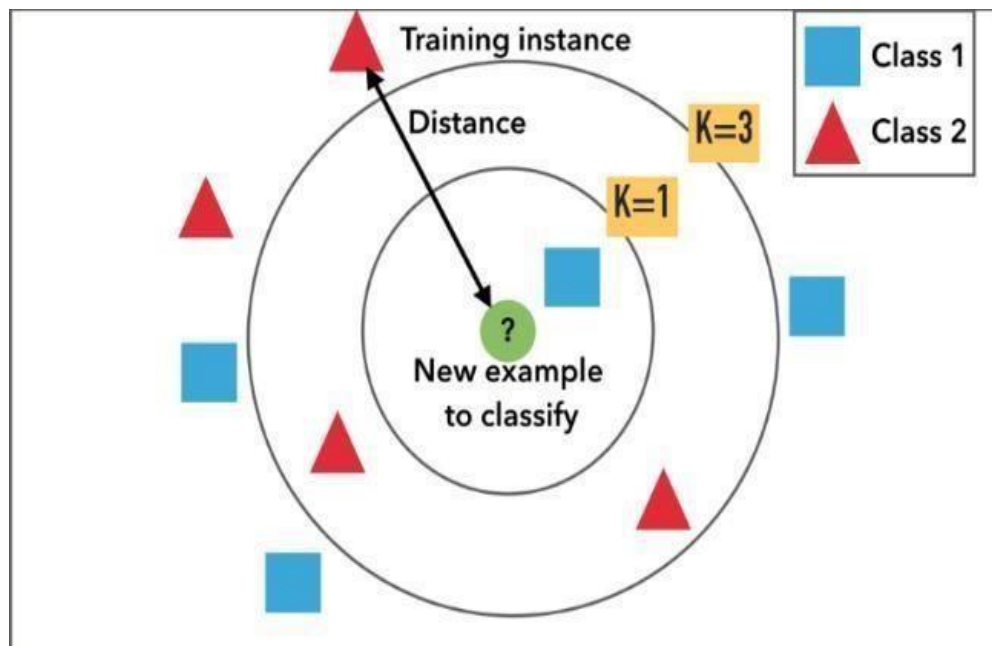


Fig 2.5 K-NN

## ➤ SOFTWARE DESCRIPTION

### • PYTHON

Python is a translated, question situated, abnormal state programming dialect with dynamic semantics. Its abnormal state worked in information structures, joined with dynamic composing and dynamic authoritative, make it exceptionally appealing for Rapid Application Development, and for use as a scripting or paste dialect to interface existing parts together. Python's straightforward, simple to learn sentence structure accentuates intelligibility and subsequently decreases the expense of program upkeep. Python underpins modules and bundles, which energizes program seclusion and code reuse. The Python mediator and the broad standard library are accessible in source or parallel frame without charge for every single significant stage and can be openly disseminated.

### • SCIKIT-LEARN

Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithm including support vector machines, random forests, gradient boosting,  $k$ -means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy. The scikit-learn project started as scikits.learn, a Google Summer of Code project by David Cournapeau. Its name stems from the notion that it is a "SciKit" (SciPy Toolkit), a separately-developed and

distributed third-party extension to SciPy. The original codebase was later rewritten by other developers. Of the various scikits, scikit-learn as well as scikit-image were described as "well-maintained and popular" in November 2012. As of 2018, scikit-learn is under active development.

## **CHAPTER 3**

### **REQUIREMENT ANALYSIS**

#### **3.1 METHODOLOGY FOLLOWED**

The steps followed to do this project are:

1. Collection of dataset.
2. Understanding features of dataset.
3. Pre-processing the data.
4. Split data into training dataset and testing dataset.
5. Apply ML algorithm to dataset to predict the breast cancer.
6. Improving results

#### **3.2 FUNCTIONAL AND NON FUNCTIONAL REQUIREMENTS**

##### **FUNCTIONAL:**

The functional requirement define the system or the components of the system. A function is basically inputs, behaviors and outputs. Stuff that can be called functional requirements are: calculations, technical details, data manipulation and processing. It tells us what a system is supposed to do.

Here, the system has to perform the following tasks:

- Understand all the features as well as the data provided in the dataset.

- Map the data in the dataset with the given input data. Find patterns, if any, with both the dataset as well as input data.
- Check whether the input data of a patient will result in the diagnosis of breast cancer or not.
- If breast cancer is diagnosed, provide information on the type of breast cancer, i.e, benign or malignant.
- Provide the percentage accuracy of the proposed prediction.

## **NON FUNCTIONAL:**

A non-functional requirement is a requirement gives the criteria that can be used to judge how well a system can function. It comes under system/requirements engineering. It gives a judgement on the overall unlike functional requirements which define specific behavior or functions. Functional requirements are implemented by using the system design whereas system architecture is what is used for implementing the non-functional requirements.

Non-functional requirements are also called constraints.

Some of the quality attributes are as follows:

### **3.2.1 ACCESSIBILITY:**

Accessibility is a term that is used to describe if a product or software is accessible to the public and how easily can it be accessed.

It is easy to access as the dataset is open source and can be found on the University of California, Irvin's ML dataset repository. Unlike breast cancer diagnosis tests in hospitals which cost a lot, anyone can access this dataset for free.

### **3.2.2 MAINTAINABILITY:**

Maintainability tells us how easily a software or tool or system can be modified in order to:

- Correct defect



- Meet new requirements

Different programming languages can be used to make the predictive model based on the programmer's wishes. The datasets can also be modified and new data can be added as and when the data is updated by doctors. Different ML algorithms can also be used to check which algorithm will give the best result.

As python programming language that can adapt to new changes easily, it is easy to maintain this type of system.

### **3.2.3 SCALABILITY:**

The system can work normally under situations such as low bandwidth and huge datasets. The R studio as well as Excel can take care of these data and can perform the algorithms with ease.

### **3.2.4 PORTABILITY:**

Portability is a feature which tells us about the ease at which we can reuse an existing piece of code when we move from one location or environment to some other.

This system uses python and R programming languages and they can be executed under different operation conditions provided it meet its minimum configurations. Only system files and dependent assemblies would have to be configured in such case.

## **3.3 HARDWARE REQUIREMENTS**

Processor	: Any Processor above 500 MHz
RAM	: 512Mb
Hard Disk	: 10 GB
Input device	: Standard Keyboard and Mouse

Output device : VGA and High Resolution Monitor

### **3.4 SOFTWARE REQUIREMENTS**

- Operating System : Linux ,Windows (8,7,10,11) [Anyone]
- Application Required : Pycharm
- Technology : Python language
- Library Used : NumPy, Pandas, SkLearn
- Framework : Flask
- Documentation : MS Word

## **3.5 Feasibility Study**

### **Feasibility Study**

**Feasibility Study** is a study to evaluate feasibility of proposed project or system.

As name suggests feasibility study is the feasibility analysis or it is a measure of the software product in terms of how much beneficial product development will be for the organization in a practical point of view. Feasibility study is carried out based on many purposes to analyze whether software product will be right in terms of development, implantation, contribution of project to the organization etc.

#### **3.5.1 Technical feasibility**

This project scheme was developed to reduce some amount of work for the physicians and other doctors so that they don't have to conduct many tests on the patients. It also helps minimize the amount of time and money spent by the patients undergo these tests. As everything is digitalized and based on data analysis, it takes less amount of time to get results. Based on the results, further action can be taken. It also helps researchers in the medical as well as IT sector to understand how different algorithms can predict different outcomes.

#### **3.5.2 Operational Feasibility**

This prediction can help doctors prescribe different medical examinations for the patients based on the cancer type. This helps save a lot of time as well as money for the patient. These tools can help discover and identify patterns and relationships between the cancer data, from huge datasets, while they are able to effectively predict future outcomes of a cancer type. Patients have to spend a lot of money on different tests and treatments to check whether they have breast cancer or not. These tests can take a long time and the results can be delayed. Also, after confirmation that the patient has cancer, more tests need to be done to check whether the cancer is benign or malignant.

#### **3.5.3 Behavioral Feasibility**

This scheme normally requires a huge amount of data about different patient history and the cancer

details. This data has been collected by many doctors for a long period of time and will be used to do the analysis. This reduces the computational time required to gather all the data necessary.

# **CHAPTER 4**

## **DESIGN**

### **4.1 DESIGN GOALS**

Under our model, the goal of our project is to create a design to achieve the following:

#### **4.1.1 ACCURACY**

Only accurate outcomes can help make this model a good one. It can be reliable only when all the outcomes are correct and can be trusted. As this data is required for healthcare purposes, it is important that no errors occur.

#### **4.1.2 EFFICIENCY**

The model should be efficient as there is no requirement of manual data entry work or any work by doctors. It takes less time to predict outcomes after all the ML algorithms has been used on the data.

### **4.2 SYSTEM ARCHITECTURE**

This project has UI and the architecture is basically the dataset and the features of the dataset. It is trying to understand the dataset and try making the system as simple and easy as possible.

The dataset is first split into training and testing set. The training set is first exposed to the machine learning algorithms so that the system understands what data gives what type of outcome.

After the system is trained, the testing data is used to test whether the system can correctly predict the class of the data. It checks the percentage accuracy of the model.

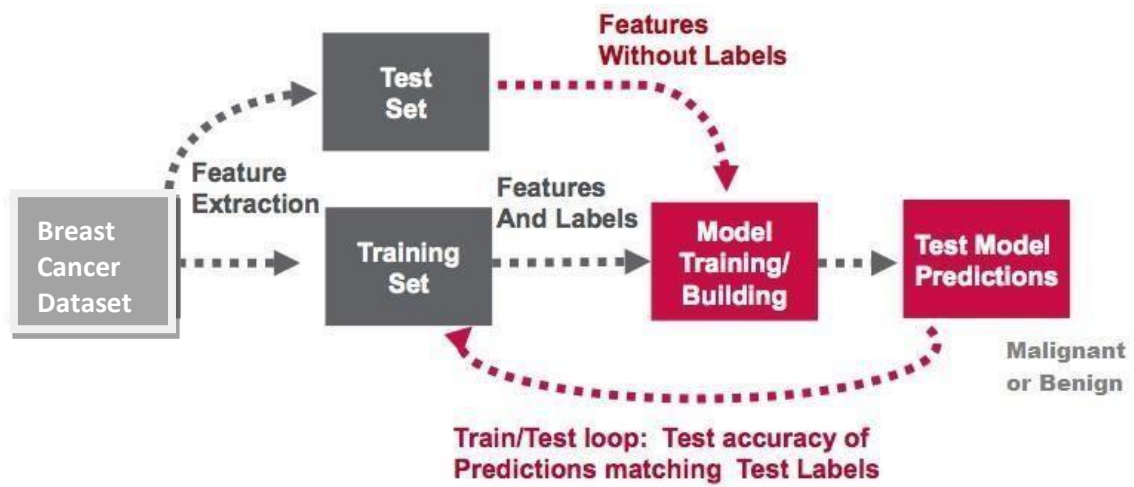


Fig 4.1 System Architecture

## 4.3 DATA FLOW DIAGRAM

The dataflow diagram shows the way in which the data from the dataset moves.

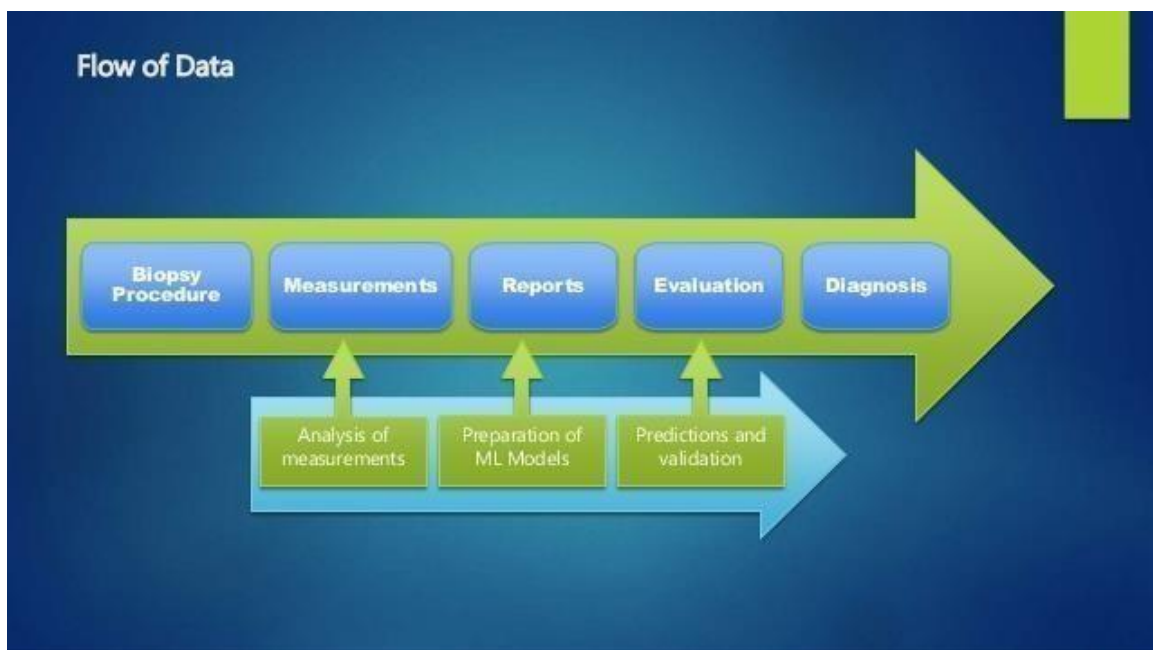


Fig 4.2 Data Flow Diagram

## 4.4 EXPECTED OUTCOME

The outcome of this model is to correctly check and predict whether a patient has breast cancer or not. If yes, the model should also be able to tell if the patient has malignant or benign type of cancer.

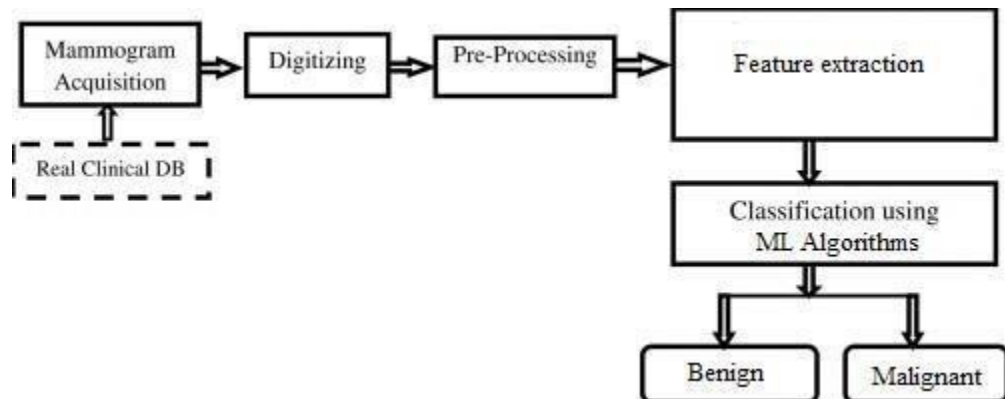


Fig 4.3 Expected Outcome

### 4.1 Input / Output Form (Screenshot)

**Breast Cancer Prediction Form**

K-Nearest Neighbor Model is developed based on 10 features that classify whether the breast cancer is benign or malignant.

Texture Mean	<input type="text" value="14.36"/>
Area Mean	<input type="text" value="143"/>
Concavity Mean	<input type="text" value="0.00"/>
Area SE	<input type="text" value="6.88"/>
Concavity SE	<input type="text" value="0.40"/>
Fractal Dimension SE	<input type="text" value="0.03"/>
Smoothness Worst	<input type="text" value="0.22"/>
Concavity Worst	<input type="text" value="1.25"/>
Symmetry Worst	<input type="text" value="0.66"/>
Fractal Dimension Worst	<input type="text" value="0.21"/>
<input type="button" value="Predict"/>	

#result.html

**Naïve Bayes classification technique is based on the Bayes' Theorem. It is named 'naïve' because this algorithm assumes each input variable to be independent.**

Breast Cancer Prediction Form

K-Nearest Neighbor Model is developed based on 10 features that classify whether the breast cancer is benign or malignant.

Texture Mean

9.71 to 39.28

Area Mean

143.00 to 2501.99

Concavity Mean

0.00 to 0.43

Area SE

6.88 to 542.20

Concavity SE

0.00 to 0.40

Fractal Dimension SE

0.00 to 0.03

Smoothness Worst

0.07 to 0.22

Concavity Worst

0.00 to 1.25

Symmetry Worst

0.16 to 0.66

Fractal Dimension Worst

0.06 to 0.21

Predict

#result.html

As a result, when new data appears, the K-NN algorithm can be easily used to classify it into an appropriate category.

The patient is more likely to have BENIGN Cancer



## CHAPTER 5

### CODING

Module wise code

#### **Prediction form(html)**

```
<!DOCTYPE html>
<html>
<body>
  <style>
    * {
      box-sizing: border-box;
    }

    input[type=text], select, textarea {
      width: 100%;
      padding: 12px;
      border: 1px solid #ccc;
      border-radius: 4px;
      resize: vertical;
    }

    label {
      padding: 12px 12px 12px 0;
      display: inline-block;
    }

    input[type=submit] {
      background-color: #04AA6D;
      color: white;
      padding: 12px 20px;
      border: none;
      border-radius: 4px;
      cursor: pointer;
      float: right;
    }

    input[type=submit]:hover {
      background-color: #45a049;
    }

    .container {
      border-radius: 5px;
      background-color: #f2f2f2;
      padding: 20px;
    }

    .col-25 {
```

```

float: left;
width: 25%;
margin-top: 6px;
}

.col-75 {
float: left;
width: 75%;
margin-top: 6px;
}

/* Clear floats after the columns */
.row:after {
content: "";
display: table;
clear: both;
}

/* Responsive layout - when the screen is less than 600px wide, make the two columns stack on top of each
other instead of next to each other */
@media screen and (max-width: 600px) {
.col-25, .col-75, input[type=submit] {
width: 100%;
margin-top: 0;
}
}
</style>
<h1><center>Breast Cancer Prediction Form</center></h1>
<p><center><b>K-Nearest Neighbor Model is developed based on 10 features that classify whether the breast
cancer is benign or malignant.</b></center></p>

<div class="container">
<form action="/result" method="POST">
<div class="row">
<div class="col-25">
<label for="t_mean">Texture Mean</label>
</div>
<div class="col-75">
<input type="number" step="any" name="texture_mean" min="9.71" max="39.28"
value="{ { request.form['texture_mean'] } }" placeholder="9.71 to 39.28">
</div>
</div>

<div class="row">
<div class="col-25">
<label for="a_mean">Area Mean</label>
</div>
<div class="col-75">
<input type="number" step="any" name="area_mean" min="143.00" max="2501.99"
value="{ { request.form['area_mean'] } }" placeholder="143.00 to 2501.99">
</div>
</div>

```

```

<div class="row">
  <div class="col-25">
    <label for="c_mean">Concavity Mean</label>
  </div>
  <div class="col-75">
    <input type="number" step="any" name="concavity_mean" min="0.00" max="0.43"
    value="{{ request.form['concavity_mean'] }}" placeholder="0.00 to 0.43">
  </div>
</div>

<div class="row">
  <div class="col-25">
    <label for="area_se">Area SE</label>
  </div>
  <div class="col-75">
    <input type="number" step="any" name="area_se" min="6.80" max="542.20"
    value="{{ request.form['area_se'] }}" placeholder="6.88 to 542.20">
  </div>
</div>

<div class="row">
  <div class="col-25">
    <label for="c_se">Concavity SE</label>
  </div>
  <div class="col-75">
    <input type="number" step="any" name="concavity_se" min="0.00" max="0.40"
    value="{{ request.form['concavity_se'] }}" placeholder="0.00 to 0.40">
  </div>
</div>

<div class="row">
  <div class="col-25">
    <label for="fd_se">Fractal Dimension SE</label>
  </div>
  <div class="col-75">
    <input type="number" step="any" name="fractal_dimension_se" min="0.00" max="0.03"
    value="{{ request.form['fractal_dimension_se'] }}" placeholder="0.00 to 0.03">
  </div>
</div>

<div class="row">
  <div class="col-25">
    <label for="s_worst">Smoothness Worst</label>
  </div>
  <div class="col-75">
    <input type="number" step="any" name="smoothness_worst" min="0.07" max="0.22"
    value="{{ request.form['smoothness_worst'] }}" placeholder="0.07 to 0.22">
  </div>
</div>

<div class="row">

```

```

    <div class="col-25">
      <label for="c_worst">Concavity Worst</label>
    </div>
    <div class="col-75">
      <input type="number" step="any" name="concavity_worst" min="0.00" max="1.25"
        value="{{ request.form['concavity_worst'] }}" placeholder="0.00 to 1.25">
    </div>
  </div>

  <div class="row">
    <div class="col-25">
      <label for="s_worst">Symmetry Worst</label>
    </div>
    <div class="col-75">
      <input type="number" step="any" name="symmetry_worst" min="0.16" max="0.66"
        value="{{ request.form['symmetry_worst'] }}" placeholder="0.16 to 0.66">
    </div>
  </div>

  <div class="row">
    <div class="col-25">
      <label for="fdw">Fractal Dimension Worst</label>
    </div>
    <div class="col-75">
      <input type="number" step="any" name="fractal_dimension_worst" min="0.06" max="0.21"
        value="{{ request.form['fractal_dimension_worst'] }}" placeholder="0.06 to 0.21">
    </div>
  </div>

  <button type="submit">Predict</button>

</form>
</div>
</body>
</html>

#result.html
<!DOCTYPE html>
<html lang="en">
<head>
  <meta charset="UTF-8">
  <title>Result</title>
</head>
<body style="font-size:20px;">

  {{ prediction_text }}

</body>
</html>

```

## Deploying Breast Cancer Prediction Model Using Flask APIs

In deploying this prediction model into production, a web application framework called Flask API is used.

**model.py**

Create model.py and run for saving and loading model

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import io
import os
import pickle
from flask import render_template
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.neighbors import KNeighborsClassifier
from sklearn.datasets import load_breast_cancer

df = pd.read_csv('BCPD.csv')
print(df.head())
#print("target name:", df[""])
# select dependent and independent variable
x = df[["texture_mean", "area_mean", "concavity_mean", "area_se", "concavity_se", "fractal_dimension_se",
        "smoothness_worst", "concavity_worst", "symmetry_worst", "fractal_dimension_worst"]]
#x = df[["radius_mean", "perimeter_mean", "area_mean", "symmetry_mean", "compactness_mean", "concave
points_mean"]]
y = df[["diagnosis"]]

# split the data into train and test

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.33, random_state=42)

# feature scaling
sc = StandardScaler()
Fit=sc.fit(x_train)
x_train = Fit.transform(x_train)
x_test = Fit.transform(x_test)

# instantiate model
classifier = KNeighborsClassifier()

# fit the model
classifier.fit(x_train, y_train)
```

```
# make pickle file of our model
pickle.dump(classifier, open("model.pkl", "wb"))
```

Create an API app.py and run the flask server for getting input from the user and computes a predicted value.

### **App.py(flask)**

```
import os
import numpy as np
from flask import Flask, request, jsonify, render_template
import pickle
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
import pandas as pd

# create flask app
app = Flask(__name__)
df = pd.read_csv('BCPD.csv')
x = df[["texture_mean", "area_mean", "concavity_mean", "area_se", "concavity_se", 'fractal_dimension_se',
        "smoothness_worst", "concavity_worst", "symmetry_worst", "fractal_dimension_worst"]]
y = df[["diagnosis"]]
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.33, random_state=42)

# load pickle model
model = pickle.load(open("model.pkl", "rb"))

@app.route("/")
def home():
    return render_template("prediction_form.html")
    #return render_template("newprediction.html")

@app.route("/result", methods=["GET", "POST"])
def predict():
    l = []
    form_value = 1
    if request.method == "POST":
        print(request.values)
    imd = request.form
    imd.to_dict(flat=False)
    print(imd)
    for k,v in imd.items():
        form_value.append(v)
```

```

print(type(request.form))
print(request.form)
float_features = [float(x) for x in form_value]
features = np.array([np.array(float_features)])
sc=StandardScaler()
Fit=sc.fit(x_train)
features=Fit.transform(features)
prediction = model.predict(features)
if prediction[0] == 1:
    print("Malignant")
    return render_template("result.html", prediction_text=" MALIGNANT Cancer")
else:
    print("Benign")
    return render_template("result.html", prediction_text=" BENIGN Cancer")

if __name__ == "__main__":
    HOST = os.environ.get('SERVER_HOST', 'localhost')
    try:
        PORT = int(os.environ.get('SERVER_PORT', '5555'))
    except ValueError:
        PORT = 5555
    app.run(HOST, PORT)

```

This will return an output that shows that your app is running. Just copy the URL and paste it into your browser to test the app.

### **result.html**

```

<!DOCTYPE html>
<html lang="en">
<head>
    <meta charset="UTF-8">
    <title>Result</title>
<style>
    body {
        margin: 8px 16px;
        background-color: LightPink;
        padding: 12px 24px;
        border: 1px solid black;
        border-radius: 4px;
    }
</style>

</head>

<body style="font-size:20"; background-color="#212F3C"; color=" #FFFFF0";>

```

```
< 1>The patient is more likely to have {{ prediction_text }}</h1>
</body>
</html>
```

## STEPS:

- Remove the unnecessary columns or columns with nominal data.
- Normalize the data as all data use different scales.
- Split data into training and testing data (70-30).
- Calculate K value.
- Using the training data and KNN algorithm, get the predicted values for the testing data.
- Compare the predicted values to actual values and calculate accuracy of model.

**The data has 569 diagnosis, 357 malignant and 212 benign.**

**Accuracy: 92.11%**

**Execution time: 0.020711 seconds**



## CHAPTER 6

### CONCLUSION

In this paper we have worked to collect the suitable dataset needed to help in this predictive analysis. This dataset is then processed to remove all the junk data. The predictive analysis method is being used in many different fields and is slowly picking up pace. It is helping us by using smarter ways to solve or predict a problem's outcome. Our scheme was developed to **reduce the time and cost factors of the patients as well as to minimize the work of a doctor**. We have tried to use a very simple and understandable model to do this job. Next, machine learning algorithms should be used on the training data and the testing data should be used to check if the outcomes are accurate enough.

In the future, we can also use other machine algorithms for prediction and dataset to predict the **re-occurrence of breast cancer after a surgery or chemotherapy session. Artificial Neural Networks can be applied to make the prediction better and smarter.** Accuracy can be increased by selecting better features.

## REFERENCES

- [1] Gupta(2022) 'Prediction time of breast cancer prediction'. *Cancer Treatment and Research Communication*.  
<https://www.sciencedirect.com/science/article/pii/S2468294222000922>
- [2] Priyanka Khanna, Mridu Sahu, Bikesh Kumar Singh, Vikrant Bhateja, 'Early Prediction of Pathological Complete Response to neoadjuvant Chemotherapy in breast cancer MRI images using combined Pre-trained Convolutional neural network and machine learning'.  
*Measurement*(2022) <https://www.sciencedirect.com/science/article/pii/S0263224122014658>
- [3] Law Kumar Singh, Munish Khanna, Rekha Singh, 'Artificial intelligence based medical decision support system for early and accurate breast cancer prediction', *Advances in Engineering Software*, Volume 175 (2023),  
<https://www.sciencedirect.com/science/article/pii/S0965997822002393>
- [4] Mohammed Amine Naji, Sanaa El Filali, Kawtar Aarika, EL Habib Benlahmar, Rachida Ait Abdelouahid, Olivier Debauche, 'Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis', *Procedia Computer Science*, Volume 191, (2021), <https://www.sciencedirect.com/science/article/pii/S1877050921014629>
- [5] Rashmi, G.D., Lekha, A., Bawane, N. (2016). 'Analysis of Efficiency of Classification and Prediction Algorithms (kNN) for Breast Cancer Dataset'. [https://link.springer.com/chapter/10.1007/978-81-322-2752-6\\_18#citeas](https://link.springer.com/chapter/10.1007/978-81-322-2752-6_18#citeas)
- [6] Saba Dadsetan, Dooman Arefan, Wendie A. Berg, Margarita L. Zuley, Jules H. Sumkin, Shandong Wu, 'Deep learning of longitudinal mammogram examinations for breast cancer risk prediction', *Pattern Recognition*, Volume 132, (2022), <https://www.sciencedirect.com/science/article/pii/S0031320322004009>
- [7] Fan Zhang, Yingqi Zhang, Xiaoke Zhu, Xiaopan Chen, Haishun Du, Xinhong Zhang, 'PregGAN: A prognosis prediction model for breast cancer based on conditional generative adversarial networks', *Computer Methods and Programs in Biomedicine*, Volume 224, (2022)  
<https://www.sciencedirect.com/science/article/pii/S0169260722004084>
- [8] osayba Al-Azzam, Ibrahim Shatnawi, 'Comparing supervised and semi-supervised Machine Learning Models on Diagnosing Breast Cancer', *Annals of Medicine and Surgery*, Volume 62, (2021) <https://www.sciencedirect.com/science/article/pii/S2049080120305604>
- [9] Shallu Sharma, Sumit Kumar, 'The Xception model: A potential feature extractor in breast cancer histology images classification', Volume 8, Issue1, (2022) <https://www.sciencedirect.com/science/article/pii/S240595952100165X>
- [10] V.Nanda Gopal, Fadi Al-Turjman, R. Kumar, L. Anand, M. Rajesh, 'Feature

selection and classification in breast cancer prediction using IoT and machine learning', *Measurement*, Volume 178, (2021) <https://www.sciencedirect.com/science/article/pii/S0263224121004310>

[11] Shashank Shekhar, Nazrul Hoque, Dhruba K. Bhattacharyya, 'PKNN-MIFS: A Parallel KNN Classifier over an Optimal Subset of Features', *Intelligent Systems with Applications*, Volume 14, (2022) <https://www.sciencedirect.com/science/article/pii/S266730532200014X>

[12] Abdul Majid, Safdar Ali, Mubashar Iqbal, Nabeela Kausar, 'Prediction of human breast and colon cancers from imbalanced data using nearest neighbor and support vector machines', *Computer Methods and Programs in Biomedicine*, Volume 113, Issue 3, (2014), <https://www.sciencedirect.com/science/article/pii/S0169260714000029>

[13] Xiumei Wu, Qiongyao Guan, Andy S.K. Cheng, Changhe Guan, Yan Su, Jingchi Jiang, Yingchun Zeng, Linghui Zeng, Boran Wang, 'Comparison of machine learning models for predicting the risk of breast cancer-related lymphedema in Chinese women', *Asia-Pacific Journal of Oncology Nursing*, Volume 9, Issue 12, (2022), <https://www.sciencedirect.com/science/article/pii/S2347562522001597>

[14] Haiyan Wang, Peidi Xu, Jinghua Zhao, 'Improved KNN algorithms of spherical regions based on clustering and region division', *Alexandria Engineering Journal*, Volume 61, Issue 5, (2022), <https://www.sciencedirect.com/science/article/pii/S1110016821006062>

[15] S. Pavithra, R. Vanithamani, J. Justin, 'Computer aided breast cancer detection using ultrasound images', *Materials Today: Proceedings*, (2020), <https://www.sciencedirect.com/science/article/pii/S2214785320362532>

[16] Wolberg, Street and Mangasarian, "Wisconsin Diagnostic Breast Cancer Dataset" <http://archive.ics.uci.edu/ml>

[17] Mengjie Yu, "Breast Cancer Prediction Using Machine Learning Algorithm", The University of Texas at Austin, 2017.

[18] Wenbin Yue, Zidong Wang, Hongwei Chen, Annette Payne and Xiaohui Liu, "Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis", 2018.

[19] S. Palaniappan and T. Pushparaj, "A Novel Prediction on Breast Cancer from the Basis of Association rules and Neural Network", 2013.

[20] Joseph A. Cruz, David S. Wishart, "Applications of Machine Learning in Cancer Prediction and Prognosis"