

CancerEase: A New Approach Towards Early Prediction of Breast Cancer Using MRI Scan Data.

Jash Nanda

Singapore International School
jashnanda1234@gmail.com

Reetu Jain

On My Own Technology Pvt. Ltd.
reetu.jain@onmyowntechnology.com

Abstract: Breast Cancer is the second leading cause of death among women. 1 in every 8 women has a chance of getting breast cancer during their lifetime. This magnitude of cancer patients is overwhelming for our current healthcare systems and is costing a lot of lives every year due to late diagnosis of the disease. To aid our healthcare system during this era of the exponential rise in cancer cases, we have come up with CancerEase: a machine learning software that can diagnose a patient's cancer as malignant or benign using MRI scan data. CancerEase uses 3 different models: Neural Networks, Logistic Regression, and K-Nearest Neighbors Classifier. It then chooses the most effective model by comparing their accuracy scores and uses it to diagnose patients. CancerEase uses MRI data from the Kaggle website to train and test its Machine Learning models. This data has been contributed by multiple professionals, making it fit to be used for the Machine Learning Models. After comparing the accuracy score of the 3 models, the Neural Network Model came out as the most effective model in predicting the diagnosis with an accuracy score of approximately 92.98%. This is a very promising result, making CancerEase an effective model for tackling the world breast cancer problem.

Keywords: Neural networks, Breast cancer detection, KNN Classification, MRI Scan data for breast cancer.
Early prediction

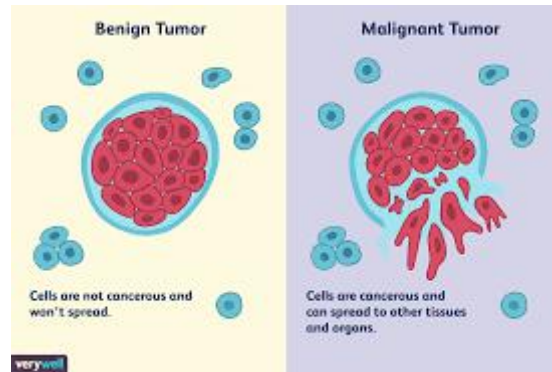
Introduction:

What is Meant by Cancer:

Cancer is a disease in which some of the body's cells grow uncontrollably and spread to the body's other parts. Cancer can occur in almost any part of the body. However, breast cancer is the most common type. Breast Cancer is the second leading cause of death among women worldwide. According to a study, there is a 1 in 8 (13%) chance that a woman will develop breast cancer during her lifetime, and a 1 in 39 (2.5%) chance she will die due to it. Good medical care is not enough to defeat the disease. Breast Cancer is on the rise in both the urban and rural populations. In 2018, 1,62,468 new cases, and 87,090 deaths were reported in India.

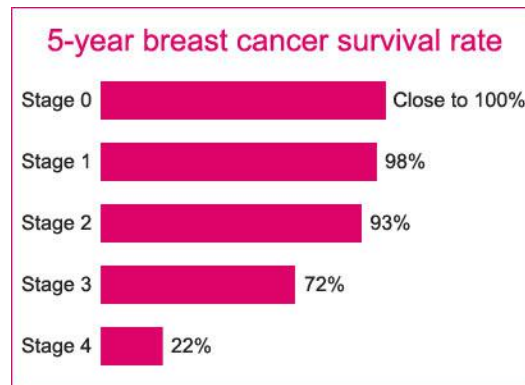
Types of Cancer:

The uncontrollable growth of cells may cause tumors, which are lumps of tissue. These lumps are of 2 kinds: Malignant (Cancerous), Benign (Non-Cancerous). Benign tumors are easy to treat as once they are surgically removed, they do not grow back. The Malignant cancer cells however multiply rapidly and have a chance of emerging back stronger. There is a 15% chance that breast cancer may reappear after it is removed. This attack is stronger and may also be fatal.



Current Treatment:

Currently, the only solution to treating cancer is to diagnose and remove it in its early stage. Breast Cancer is treatable, and the chances of survival are high if detected early. Modern treatments such as Chemotherapy, Hormonal Therapy, Biological Therapy, and Radiation Therapy can treat cancer in the early stages without the need for operating. The graph below shows the same.



Problems With the Current Method of Diagnosis:

However, the current diagnosis method (Mammography) is not efficient enough to diagnose the growing population. In Mammography, X-Ray images are manually searched for signs of Cancer which leads to a lot of wastage of time. Every year, 2.3 million people get Breast Cancer, and early detection through an alternative, time-efficient diagnosis method can reduce the mortality rate of the disease. This project has been designed to deal with this very problem that costs thousands of lives every year.

Why Artificial Intelligence May Be Our Best Option:

Significant advancements made in deep convolutional neural networks (often referred to as deep learning techniques) and artificial intelligence (AI) are lowering the performance gap between humans and computers in many medical imaging applications, including breast cancer detection. This project has the potential to improve the performance of breast cancer screening programs and may eventually be improved thanks to this new generation of deep learning-based CAD systems. Aside from advancing AI algorithms, screening can be made better with the help of the AI system. According to prior research, utilizing CAD simultaneously as a decision support tool is more beneficial to radiologists than employing prompts to assess soft-tissue lesions. AI models can diagnose patients in a matter of seconds. With high enough accuracy, Artificial Intelligence may be our best option to treat cancer.

Literature Review:

1. Early Diagnosis of Breast Cancer:

Numerous breast diagnostic methods, such as mammography, magnetic resonance imaging, ultrasound, computed tomography, positron emission tomography, and biopsy, have been examined by researchers. These methods do have some drawbacks, though, such as the fact that they are pricey, time-consuming, and unsuitable for young girls. It is necessary to create a quick and highly sensitive early-stage breast cancer diagnosis tool. Researchers have focused on creating biosensors in recent years that can identify breast cancer using various indicators. Microwave imaging techniques have also been extensively researched as a possible diagnostic tool for quick and affordable early-stage breast cancer diagnosis, in addition to biosensors and biomarkers. This study seeks to present a summary of recent significant advancements in breast screening techniques (especially microwave imaging), breast biomarkers, and biosensors for quickly diagnosing breast cancer.

2. Clinical Diagnosis and Management of Breast Cancer:

They discuss modern methods for diagnosing and treating breast cancer in this article. These methods include advice on screening, pathologic evaluation to identify the severity of the disease, diagnostic imaging, surgery, and radiation therapy, as well as a variety of systemic treatments such as chemotherapy, endocrine therapy, and targeted medicines. They also take into account how functional imaging can help usher in a time of individualized, tumor-specific treatment.

3. Breast Cancer Diagnosis and Prognosis Via Linear Programming:

In this study, two linear programming applications in medicine are discussed. Machine learning techniques based on linear programming are utilized to improve the precision and objectivity of breast cancer diagnosis and prognosis. Using features of individual cells isolated from a fine needle aspiration, the first application to diagnose breast cancer separates benign from malignant breast masses. This eliminates the need for a surgical biopsy and enables an accurate diagnosis. The diagnostic system now in use at the University of Wisconsin Hospitals was trained on samples from 569 patients and has made 131 additional diagnoses with 100% chronological accuracy. The second application, which has just entered clinical use, uses a technique to create a surface that forecasts when patients' breast cancer is likely to return.

4. Breast cancer diagnosis: Imaging techniques and biochemical markers:

They outlined the numerous imaging methods and biochemical biomarkers that may be used to diagnose breast cancer patients. Additionally, we identified exosomes and microRNAs as fresh diagnostic and therapeutic biomarkers for keeping tabs on breast cancer patients.

5. Improving breast cancer diagnosis with computer-aided diagnosis:

Eight computer-extracted features from standard-view mammograms were employed in this study's computer classification approach to evaluate the likelihood of malignancy for clustered microcalcifications. In a series of nearly consecutive biopsies, 104 cases of microcalcifications with histological verification (46 malignant, 58 benign) were employed in this investigation. 10 radiologists who read the original standard and magnification-view mammograms had their performance as observers evaluated. A percentage assessment of the chance of malignancy was offered by the computer aid. Using receiver operating characteristic (ROC) analysis and comparing biopsy suggestions, the performance of computer-aided and unaided (regular clinical) groups was compared. The average ROC curve area (A_z) increased from 0.61 without aid to 0.75 with the computer aid ($P < .0001$). On average, with the computer

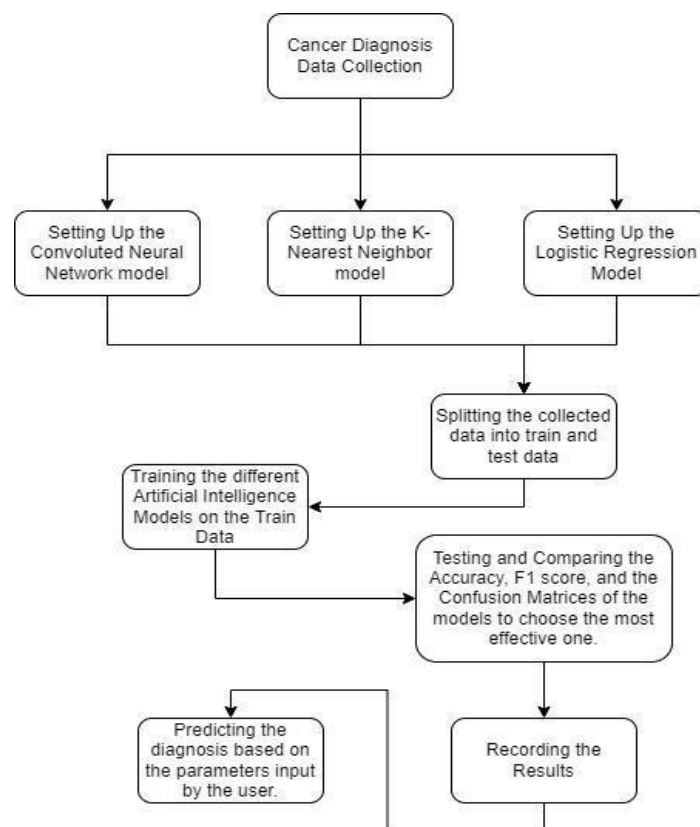
aid, each observer recommended 6.4 additional biopsies for cases with malignant lesions ($P = .0006$) and 6.0 fewer biopsies for cases with benign lesions ($P = .003$). This improvement corresponded to increases in sensitivity (from 73.5% to 87.4), specificity (from 31.6% to 41.9%), and hypothetical positive biopsy yield (from 46% to 55%). In this research work, they have concluded that CAD can be used to improve radiologists' performance in breast cancer diagnosis.

6. MicroRNAs as biomarkers for early breast cancer diagnosis, prognosis, and therapy prediction:

MicroRNAs (miRNAs) are dysregulated in all stages of breast cancer, according to a large body of research. Small non-coding RNA molecules known as miRNAs can regulate gene expression and are readily available and quantifiable. This review emphasizes miRNA as therapy-predictive diagnostic, prognostic, and diagnostic biomarkers for early breast cancer. It also looks at the difficulties they will face as biomarkers. It is noteworthy that this analysis focuses on miRNAs identified in early breast cancer patients before chemo, radiation, surgery, or distant metastasis (unless indicated otherwise). In this setting, the miRNAs discussed in this review were found to be significantly altered by at least two studies and/or many statistical tests. From sample collection to data analysis, a systematic process is suggested for miRNA assessment, ensuring comparative data analysis and repeatability of results.

Our Solution:

CancerEase is designed to perform accurate diagnoses, quickly. To do this, the solution uses three different Artificial Intelligence models including the Neural Network, the K-Nearest Neighbor Classifier, and the Logistic Regression. The system uses data from the Kaggle website. After the system imports the data, it divides it into training and testing data. The training data is then used to train the three models. Then, the accuracy, and f1 score are calculated and the confusion matrices are compared using the test data to find the most effective model. This model is selected to be used to predict the diagnosis based on the parameters input by the user.



Methodology:

CancerEase uses the Sci-kit Learn Libraries to deploy the Machine Learning models in our solution. Before training the model, however, we took several different steps which are explained below.

- 1) The cancer detection dataset is imported from Kaggle.
- 2) The dataset is modified to numeric data.
- 3) This data is further divided into train and test data in a ratio of 4:1
- 4) The dataset is divided into the input data and diagnosis data.
- 5) The Logistic Regression, KNN Classification, and Neural Network models are declared.
- 6) The models are trained on the training data.
- 7) The accuracy and F1 score of the 3 models are compared based on their test data predictions.
- 8) The optimal model is chosen
- 9) Patient cellular data is input from the user
- 10) This data is used by the optimal model to predict whether the patient has malignant or benign cancer.

The Dataset Record:

Diagnosis	M	Radius_Se	1.095	Radius Worst	25.38
Radius	17.99	Texture_Se	0.9053	Texture Worst	17.33
Texture	10.38	Perimeter_Se	8.589	Perimeter Worst	184.
Perimeter	122.8	Area_Se	153.4	Area Worst	2019
Area	1001	Smoothness_Se	0.005399	Smoothness Worst	0.1500
Smoothness	0.1184	Compactness_Se	0.04904	Compactness Worst	0.5575
Compactness	0.278	Concavity_Se	0.05373	Concavity Worst	0.7119
Concavity	0.3001	Concave Points_Se	0.01587	Concave Points Worst	0.2554
Concave Points	0.1471	Symmetry_Se	0.03003	Symmetry Worst	0.4701
Symmetry	0.2419	Fractal Dimensions_Se	0.005193	Fractal Dimensions Worst	0.1189
Fractal Dimensions	0.07871				

Importing Different Libraries:

In this part, we are importing libraries for the Neural Network, performance metrics, and data splitter from sci-kit learn. We are also importing data libraries like pandas and NumPy to allow us to operate on the dataset. Data plotting libraries such as matplotlib and seaborn are also imported to give us a better understanding of the dataset.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.neural_network import MLPClassifier
from sklearn.metrics import accuracy_score
```

Uploading the Dataset to the notebook:

In order for us to operate on the dataset in the notebook, the CSV file has to be first uploaded onto google classroom by importing files from google.colab and uploading the downloaded dataset from the computer.

```
from google.colab import files
uploaded=files.upload()
```

Converting the dataset into a Pandas dataframe:

Before the model is executed on the dataset, certain functions must be done on the dataset. It also must be plotted to give us a better view. Thus, the dataset in the CSV file is stored in a pandas dataframe called df which provides various tools to manipulate the data.

```
df=pd.read_csv('data (1).csv')
```

Converting the diagnosis into a numeric value:

Since the Machine Learning models can only process numeric data, the 'M', and 'B' values of the diagnosis, representing malignant and benign respectively, are converted to 1 and 0 where 1 is malignant and 0 is benign.

```
for i in range(569):
    if df['diagnosis'][i]=='M':
        df['diagnosis'][i]=1
    else:
        df['diagnosis'][i]=0
```

Removing Unnecessary Columns:

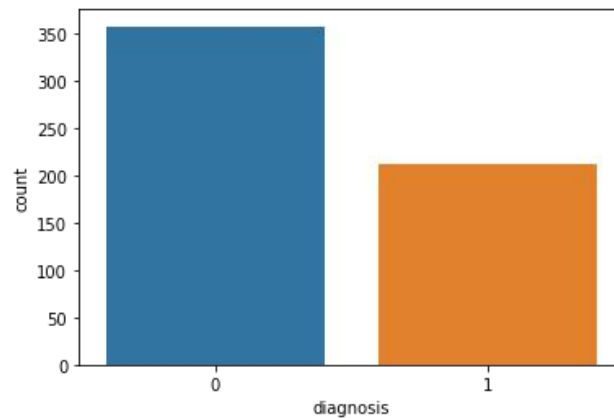
The entire dataset will not be used in the training of the model. Attributes such as id of the cancer cell, and other unnecessary values which do not contribute to the model are removed from the dataframe.

```
df=df.drop('id',axis=1)
df=df.drop('Unnamed: 32', axis=1)
```

Viewing the dataset distribution:

Before we start designing our model, we must view the distribution of data among the different diagnoses. This is done so that we know whether the distribution of data is not extreme in the 2 classes.

```
sns.countplot(df['diagnosis'], label="Count")
```



Splitting the dataset into train and test data:

A machine learning model learns how to predict the diagnosis by observing the data called train data. It is then again fed new data called test data to check its performance. However, the test and the train data cannot be the same as the test would then not be a true measure of how the model would perform in the real world as the model should have the capability to generalize its predictions since the real world data will not be the same as the train data. In this code, 80% of the dataset is assigned to the training data and the remaining 20% is assigned to testing data.

```
y=df['diagnosis']  
X=df.drop('diagnosis', axis=1)  
y=y.astype('int')  
xTrain,xTest,yTrain,yTest=sklearn.model_selection.train_test_split(X,y,test_size=0.2,random_state=5)
```

Implementing the Neural Network:

It is now time to prepare our Neural Network and test its accuracy. For this, the code uses the MLP Classifier model from sci-kit learn. The model has 5 hidden layers with each layer containing 10 neurons. The model is then trained using the fit() function and its predictions of the testing data are recorded. These predictions are compared with the actual diagnoses and the accuracy of the model is printed. This model gives us correct predictions approximately 93% of the time.

```
mlp=MLPClassifier(hidden_layer_sizes=(5,10))  
mlp.fit(xTrain,yTrain)  
pred_mlp=mlp.predict(xTest)  
print(accuracy_score(yTest,pred_mlp)*100)
```

Output:

```
92.98245614035088
```

Implementing the Logistic Regression Model:

The logistic regression model is implemented similarly to the neural network except there are no hyperparameters. So first the logistic regression model is declared and trained on the training data. It is made to make predictions on the testing data and the accuracy score is measured.

```
log=LogisticRegression()
log.fit(xTrain,yTrain)
yPred=log.predict(xTest)
print(accuracy_score(yTest,yPred)*100)
Output:
84.8598948945899
```

Implementing the K Nearest Neighbors Model:

The k value is taken to be 7 as it provided the highest accuracy score. It is calculated by taking the square root of the total number of training data records. The model is trained on the same training data as the previous 2 models. The predictions are then calculated, and the accuracy score is measured.

```
knn=KNeighborsClassifier(7)
knn.fit(xTrain,yTrain)
y_pred_knn=knn.predict(xTest)
print(accuracy_score(yTest,y_pred_knn)*100)
Output:
89.93334555432900
```

Result:

The Neural Network model can predict cancer cell diagnosis with an accuracy of approximately 92.98%.

The Logistic Regression model can predict cancer cell diagnosis with an accuracy of approximately 84.90%.

The KNN model can predict cancer cell diagnosis with an accuracy of approximately 89.93%.

Conclusion:

We have collected the breast cancer database from Kaggle. We have trained a really accurate model with an accuracy of 92.98%. The neural network algorithm can be effective in the precise diagnosis of breast cancer. During the process of making it, we tried 3 different models, Logistic Regression, KNN Classification, and neural network. We have found that the neural network has the highest accuracy. As of now, we have used data sets from the internet, but its model can be improved to a great extent when we will use the data that we will be collecting from the care centers and the hospitals. So, to conclude we can say that my solution can be used as a rapid AI tool to diagnose breast cancer in women.

Future Scope:

There are lots of improvements that can be done to make our solution better. Right now, the dataset used is from Kaggle. However, to practically implement the model in hospitals, we would have to train the model on actual hospital records. This, since it is the actual data, would conclusively prove whether the system is as effective as the current method of diagnosis (mammography). Furthermore, to further improve the model's accuracy, experts from the field of Machine Learning can be involved in the development and more complex models such as convoluted neural networks can be implemented. To implement this software on a practical scale, the Machine Learning Model will have to be deployed along with other methods of diagnosis to make reliable predictions. After the model has made its predictions, Mammography will have to be used to conclusively prove that the

patient has cancer. So, the system would work as a filter to relieve the stress from existing methods of diagnosis. Only the severe cases detected in the software would go on to be tested using mammography to improve the overall efficiency of the medical system.

References:

1. Wang, L. Early Diagnosis of Breast Cancer. *Sensors* **2017**, *17*, 1572. <https://doi.org/10.3390/s17071572>
2. Clinical Diagnosis and Management of Breast Cancer, Elizabeth S. McDonald, Amy S. Clark, Julia Tchou, Paul Zhang, Gary M. Freedman, *Journal of Nuclear Medicine* Feb 2016, *57* (Supplement 1) 9S-16S; **DOI:** 10.2967/jnumed.115.157834
3. Jafari, SH, Saadatpour, Z, Salmaninejad, A, et al. Breast cancer diagnosis: Imaging techniques and biochemical markers. *J Cell Physiol.* 2018; *233*: 5200– 5213. <https://doi.org/10.1002/jcp.26379>
4. Breast Cancer Diagnosis and Prognosis Via Linear Programming, Olvi L. Mangasarian, W. Nick Street, and William H. Wolberg, *Operations Research* 1995 *43*:4, 570-577, <https://doi.org/10.1287/opre.43.4.570>, <https://pubsonline.informs.org/doi/abs/10.1287/opre.43.4.570>
5. Yulei Jiang, Robert M. Nishikawa, Robert A. Schmidt, Charles E. Metz, Maryellen L. Giger, Kunio Doi, Improving breast cancer diagnosis with computer-aided diagnosis, *Academic Radiology*, Volume 6, Issue 1, 1999, Pages 22-33, ISSN 1076-6332, [https://doi.org/10.1016/S1076-6332\(99\)80058-0](https://doi.org/10.1016/S1076-6332(99)80058-0). (<https://www.sciencedirect.com/science/article/pii/S1076633299800580>)