

# **Detecting The Severity of Breast Cancer by Mammography**

Using K Nearest Neighbours Classifier

## **Submitted To**

**Dr. Mohammad Shoyaib**

Professor

IIT, University of Dhaka

&

**Kishan Kumar Ganguly**

Lecturer

IIT, University of Dhaka

## **Submitted By**

**Abdullah-Al-Jahid**

**BSSE 1030**

**Exam Roll: 1022**

**Submission Date:** 5 September 2021



## Table of Contents

<b>1. Introduction</b>	<b>1</b>
<b>2. Objective and Scope of the Project</b>	<b>1</b>
<b>3. Methodology</b>	<b>2</b>
3.1 Dataset Description	2
3.2 Naive KNN	3
3.3 Euclidean Distance Based Weighted KNN	3
3.4 Evaluation Process Used	3
<b>4. Results</b>	<b>5</b>
4.1 Output	5
4.2 Performance Measures	6
4.2.1 Using Naive KNN	6
4.2.2 Using Euclidean distance-based weighted KNN	7
4.3 Single Datapoint Testing	7
<b>5. Tools and Technology</b>	<b>8</b>
<b>6. Conclusion</b>	<b>8</b>
<b>7. Github Link</b>	<b>9</b>
<b>8. References</b>	<b>9</b>

## 1. Introduction

Mammography is the most effective method for breast cancer screening available today. The prediction of breast cancer biopsy outcomes using mammography emphasizes an intelligible decision process. To reduce the high number of unnecessary breast biopsies, several computer-aided diagnosis (CAD) systems have been proposed in the last years. These systems help physicians in their decision to perform a breast biopsy on a suspicious lesion seen in a mammogram or to perform a short-term follow-up examination instead. Using Mammography is a much easier, faster way to predict breast cancer biopsies. By using the mammographic masses data set we can predict the severity (benign or malignant) of a mammographic mass lesion from BI-RADS attributes and the patient's age. The low positive predictive value of breast biopsy resulting from mammogram interpretation leads to approximately 70% unnecessary biopsies with benign outcomes. To reduce these unnecessary biopsies, I build a Machine Learning based tool using the K Nearest Neighbors (KNN) classifier to detect the severity of breast cancer. I have used a dataset for this purpose that has been used widely by researchers working in this field.

## 2. Objective and Scope of the Project

The aim of this project is to develop a tool that can detect the severity of breast cancer using the K Nearest Neighbor (KNN) classifier. It will help the doctors to make their decision whether to perform a breast biopsy or not. I trained the tool only using one dataset located in src/mammographic\_masses.data. Within the scope, classification using KNN only on the following parameters: BI-RADS assessment, patient's age, mass shape, mass margin and mass density.

### 3. Methodology

The program uses the K nearest neighbor algorithm with  $k = 10$  to classify the data points. I have run n-fold ( $n=10$ ) cross-validation using different weight functions. These different weight functions are discussed below.

#### 3.1 Dataset Description

This dataset contains a BI-RADS assessment, the patient's age and three BI-RADS attributes together with the ground truth (the severity field) for 452 benign and 429 malignant masses that have been identified on full-field digital mammograms collected at the Institute of Radiology of the University Erlangen-Nuremberg between 2003 and 2006. Each instance has an associated BI-RADS assessment ranging from 1 (definitely benign) to 5 (highly suggestive of malignancy). In Mammography, a mass is defined as a space-occupying lesion, visible in two different projections.

Number of Instances: 881

Number of Attributes: 6

Attribute Description:

- BI-RADS Assessment: Value of BI-RADS assessment (between 1 and 5)
- Age: Patient's age in years (integer)
- Shape: Mass shape
  - Value 1: Round
  - Value 2: Oval
  - Value 3: Lobular
  - Value 4: Irregular
- Margin: Mass margin
  - Value 1: Circumscribed
  - Value 2: Micro-lobulated
  - Value 3: Obscured
  - Value 4: Ill-defined
  - Value 5: Spiculated
- Density: Mass density

- Value 1: High
- Value 2: ISO
- Value 3: Low
- Value 4: Fat-containing
- Severity: Benign(0) or Malignant(1)

Class Distribution:

0 : 452

1: 429

## 3.2 Naive KNN

Here each point amongst the k nearest neighbors is considered equal. Each of them has a weight of 1. The side with larger weight “wins”. For example, out of 10 nearest neighbors, let's say 7 is benign and 3 is malignant. Then the predicted result is benign.

## 3.3 Euclidean Distance Based Weighted KNN

Here the k nearest neighbors are given a weight equal to the inverse of their distance from the point in the test dataset. The closer the point, the greater its influence on the prediction.

```
if (resultPoints.get(i).getClassName() == 0)
    benign +=1.0/(resultPoints.get(i).getDistance()+1);
```

Figure 1: weighted knn

Note: here 1 is added with the distance because the distance can be 0.

## 3.4 Evaluation Process Used

I used the confusion matrix, accuracy, precision, recall, and F1 Score to evaluate the quality of the machine learning model.

		Predicted value	
		P	N
True value	P	TP	FN
	N	FP	TN

Here, P = positive, N = negative, TP = true positive, FP = false positive, FN = false negative and TN = true negative.

We used an accuracy score to evaluate how well the model is performing. The equation for accuracy is -

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Equations for other measures are:

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{F1 Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

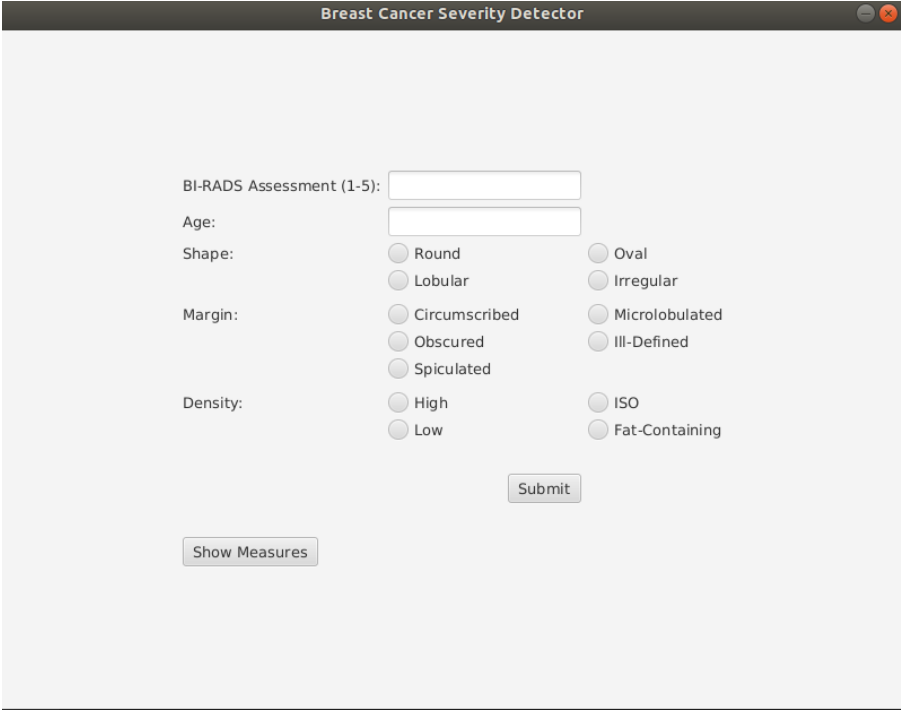
In addition to that, the program can classify a single unknown point. The inputs for a point are BI-RADS assessment, age, shape, margin and density. It predicts the severity of a given unknown point. It uses KNN with a euclidean distance-based weighted function.

The whole program has been coded in Java using the basic utils that it provides. No library for machine learning or data analysis has been used. JavaFX has been used for creating the user interface. The entire program has been coded from scratch.

## 4. Results

### 4.1 Output

Whenever we run the program a user interface like below will be shown.



The screenshot shows a window titled "Breast Cancer Severity Detector". Inside the window, there are several input fields and radio button options. The "BI-RADS Assessment (1-5)" field is a text box. The "Age:" field is also a text box. The "Shape:" label is followed by two columns of radio button options: "Round", "Lobular", "Oval", and "Irregular". The "Margin:" label is followed by two columns of radio button options: "Circumscribed", "Obscured", "Spiculated", "Microlobulated", and "Ill-Defined". The "Density:" label is followed by two columns of radio button options: "High", "Low", "ISO", and "Fat-Containing". There are two buttons: "Submit" and "Show Measures".

Figure 2: User Interface

If a user clicks on the “Show Measures” button an alert prompt will appear. It will contain different types of measures such as accuracy, precision, recall, f1 score.

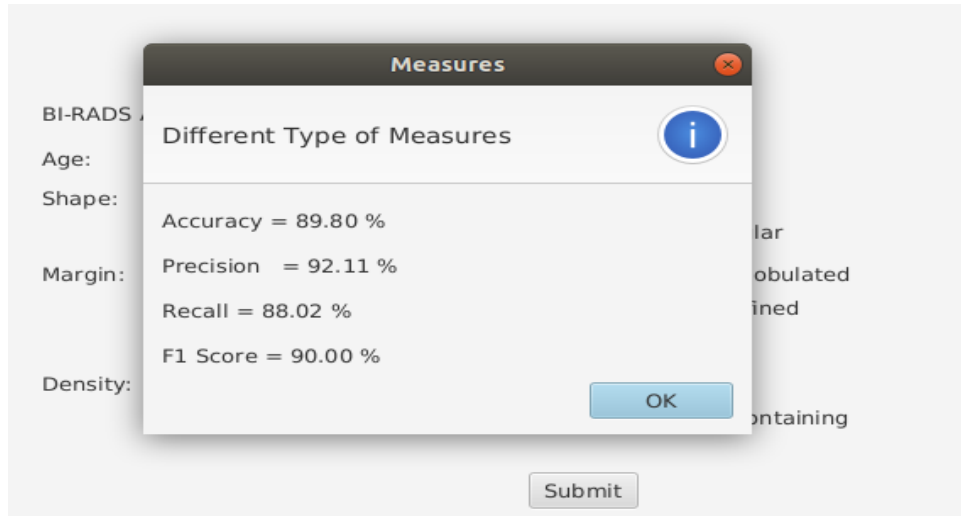


Figure 3: Different types of Measures

The measures level in 10 fold cross-validation always seems to stay between the following values.

## 4.2 Performance Measures

### 4.2.1 Using Naive KNN

Metric	Value
Accuracy	83%-89 %
Precision	82%-90%
Recall	81%-88%
F1 Score	81%-89%



#### 4.2.2 Using Euclidean distance-based weighted KNN

Metric	Value
Accuracy	86%-92 %
Precision	87%-92%
Recall	86%-90%
F1 Score	87%-91%

### 4.3 Single Datapoint Testing

Users can give input in UI and whenever they press the submit button a dialog prompt will be displayed containing the result.

The screenshot shows a web application for breast cancer assessment. The main form includes the following fields and options:

- BI-RADS Assessment (1-5):** A dropdown menu with the value '5' selected.
- Age:** A text input field containing '34'.
- Shape:** Radio button options: ☒ Round, ☐ Lobular, ☐ Oval, ☐ Irregular.
- Margin:** Radio button options: ☐ Circumscribed, ☐ Obscured, ☐ Spiculated, ☒ Microlobulated, ☐ Ill-Defined.
- Density:** Radio button options: ☐ High, ☒ Low, ☐ ISO, ☐ Fat-Containing.
- Submit:** A button at the bottom of the form.

A modal dialog box titled "Severity Level" is open, displaying the result:

- Severity Level:** Benign
- Message:** Relax! It's in benign stage. No need to biopsy.
- Action:** An "OK" button.

Figure 4: Test Single Point

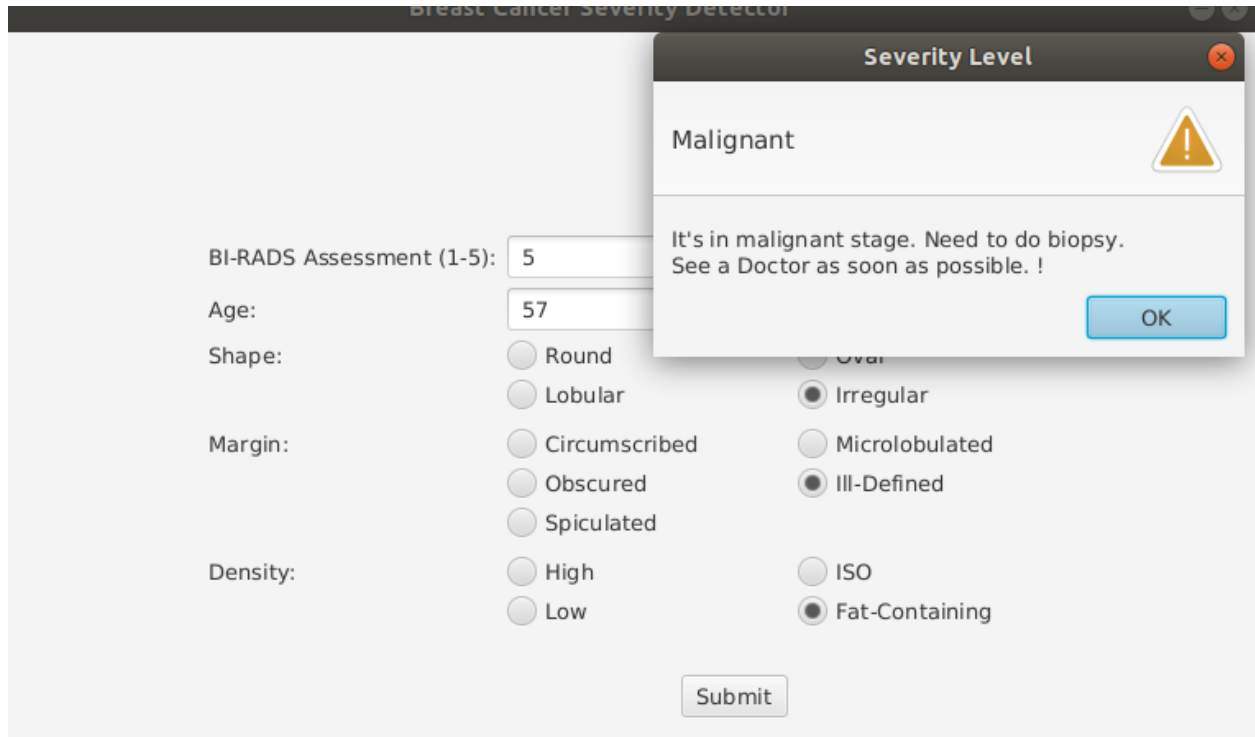


Figure 5: Test Single point

## 5. Tools and Technology

- Java
- JavaFX
- IntelliJ IDE

## 6. Conclusion

In this project, I intended to build a useful tool for detecting the severity of breast cancer to reduce the high number of unnecessary breast biopsies. I used two different techniques to improve accuracy. Hence finished obtaining an average accuracy of 89%, which is very promising.

## 7. Github Link

<https://github.com/Jahid1999/breast-cancer-severity-detector>

## 8. References

1. Dataset: <http://archive.ics.uci.edu/ml/datasets/mammographic+mass>