Breast Cancer Detection Using KNN and SVM

CSC-6740/4740 Data Mining

Project Proposal – Team # 1

Chaitanya Sai Linga
clinga@student.gsu.edu
Durga Bhavani Pothineni
dpothineni1@student.gsu.edu

Naima Mohamed
nmohamed6@student.gsu.edu
Deon Winston Europe
dwinstoneurope1@dtudent.gsu.edu

Introduction:

The topic that we have decided to conduct our project on is Breast Cancer. According to the World Health Organization (WHO), breast cancer is the most frequent cancer among women. Approximately 1 in 8 women in the United States (over 200,000) will be diagnosed with breast cancer in their lifetime. Breast cancer is also the second leading cause of death for women in the United States. Early detection is essential as treatment in earlier stages generally results in a greater prognosis.

According to Susan G. Komen, detection and treatment in stage I results in a 5-year survival rate of 98% - however treatment beginning in stage III decreases tis rate to 66-99%. Similarly, the National Cancer Institute's (NCI) Surveillance, Epidemiology, and End Results (SEER) database provides statistics on survival rates based on local, regional, and distant stages. When the cancer is in its distant stage (metastasized to distant parts of the body such as the lungs, liver, bones), the survival rate is only 28%.

Motivation:

We chose to solve this problem with the hopes of increasing the effectiveness of early detection methods. A false-positive mammogram, defined as a positive result on a screening test for breast cancer that is subsequently recognized not to be cancer. Breast cancer over/misdiagnosis can be defined as a positive result on a screening test for breast cancer that is subsequently recognized not to be cancer. The research on breast cancer in terms of detecting the difference between benign and malignant tumors is still lacking and the likelihood of being misdiagnosed with breast cancer by way of a false positive is rather high. Komen.org states, "After 10 yearly mammograms, the chance of having at least one false-positive result is about 50-60 percent." This statistic is especially troubling as women who are misdiagnosed could potentially endure unnecessary additional diagnostic workup, treatment along with the accompanying risks, time, and psychological distress.

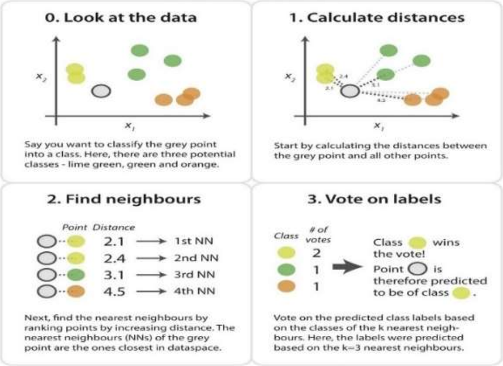
There is also a sizeable financial disadvantage to this trend. According to expenditure data from a major US healthcare plan for more than 700,000 women, the average cost for each false-positive mammogram is $852. Combined with non-essential invasive breast cancer removal ($51,837) and ductal carcinoma in situ ($12,369), this translates to a national plan cost of $4 billion each year. A great portion of this cost is passed down to the patient via increased premiums and deductibles.

Problem Statement:

Although primarily affecting women, breast cancer is the most widely diagnosed and 4th deadliest cancer in the United States. Early diagnostics significantly increase the chances of effective treatment and survival, but this process can be tedious and often leads to disagreement between pathologists. So, we are using the idea of Computer-Aided diagnosis systems which showed the potential for improving diagnostic accuracy. So, we feel that Early detection and prevention can significantly reduce the chances of death.

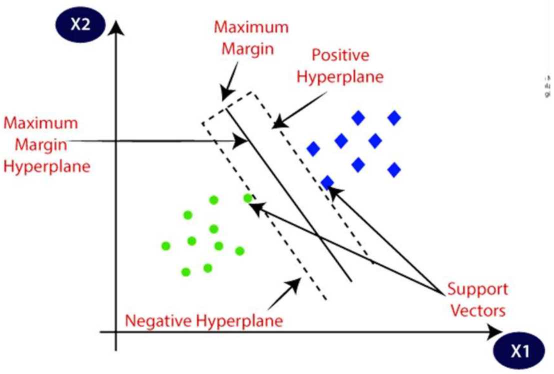
In terms of research, early detection of breast cancer is still lacking. Our solution is intended to use data to help determine whether a tumor is malignant or benign. The models that will plan to use to classify the tumors are the K-Nearest Neighbors (KNN) and Support Vector Model (SVM). We chose to use KNN model because of its robustness to noisy training data and its effectiveness in large training datasets.

KNN Algorithm:


0. Look at the data
x_2
x_1
Say you want to classify the grey point into a class. Here, there are three potential classes - lime green, green and orange.
1. Calculate distances
x_2
x_1
2.4
2.1
3.1
4.5
Start by calculating the distances between the grey point and all other points.
2. Find neighbours
Point Distance
2.1 → 1st NN
2.4 → 2nd NN
3.1 → 3rd NN
4.5 → 4th NN
Next, find the nearest neighbours by ranking points by increasing distance. The nearest neighbours (NNs) of the grey point are the ones closest in dataspace.
3. Vote on labels
Class
# of votes
2
1
1
Class wins the vote!
Point is therefore predicted to be of class .
Vote on the predicted class labels based on the classes of the k nearest neighbours. Here, the labels were predicted based on the k=3 nearest neighbours.

The Different steps involved in KNN Algorithm are Calculate the distances using Manhattan Distance(L1-norm), Finding the nearest neighbors and creating labels.

SVM Algorithm:



A support vector machine is a linear data science model for classification in regression models. It imposes a hyperplane (which in two dimensions is simply a line) that best separates the data tags. This line is the decision boundary. It acts as a discriminative classifier; it means it discriminates between classes. Unoptimized decision boundaries could result in greater misclassification of new data. This implies that only the support vector is important whereas other training examples are ignorable. It maximizes the space between that line and both of these classes. The advantages of using SVM are Effective in high dimensional space and Different kernel functions for various decision functions.

This project focuses on investigating the probability of predicting the chances of breast cancer from the given characteristics of breast mass computed from the dataset. We will be examining the data available and predicting the possibility of breast cancer.

We will use the following steps to reach our solution:

- Data collection: from the Breast Cancer Wisconsin (Diagnostic) Data Set from the UCI Machine Learning Repositor. Link to the dataset is provided in the References [3]
- Familiarize with the data by looking at data and the relationship between the variables, correlation and various other attributes.
- Preprocess the data if needed.
- Splitting the data into testing and training samples.
- Employing the various classifiers (KNN, SVM….) to predict the data with the different types of training samples.
- Identify the best prediction model and reduce the size of the training set to find the limit to best predict data.
- Comparing the best identified classifier with the evaluation metric stated at the beginning of the project.
- Publish conclusions in final report.

Attributes available in the dataset include ID, radius, texture, area, smoothness mean, compactness, symmetry mean.

Project Plan:

As a team we plan to meet 1-2 times per week after class. Below are our approximate dates to accomplish tasks. We are planning for a quick daily stand-up call in the last two weeks of the project deadline to wrap up the work and to be on the same page.

Task	Schedule
Submit Proposal	21 Feb
Deep Dive in ML Algorithms	22 Feb - 10 Mar
Data preparation	10 Mar – 22 Mar
Modelling	22 Mar – 10 Apr
Evaluation of Model	10 Apr – 15 Apr
Accuracies and Testing	15 Apr – 18 Apr
Documentation	18 Apr – 24 Apr
Presentation	18 Apr – 24 Apr
Presentation	18 Apr – 24 Apr
Submit Final Report	25 Apr
Submit Code/Demo Video	25 Apr

References:

1. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4127616/
2. https://seer.cancer.gov/statfacts/html/common.html
3. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29
4. https://www.komen.org/breast-cancer/screening/mammography/accuracy/#:~:text=False%20positive%20results&text=After%2010%20yearly%20mammograms%2C%20the,with%20dense%20breasts%20%5B36%5D
5. https://www.acpjournals.org/doi/full/10.7326/0003-4819-151-10-200911170-00010
6. https://www.komen.org/breast-cancer/facts-statistics/breast-cancer-statistics/survival-rates
7. https://pubmed.ncbi.nlm.nih.gov/20706161/
8. https://pubmed.ncbi.nlm.nih.gov/25847639/

Peer Evaluation:

	Chaitanya Sai. L	**Durga Bhavani. P**	**Naima Mohamed**	**Deon Winston. E**
Chaitanya Sai. L	**X**	**10**	**10**	**10**
Durga Bhavani. P	**10**	**X**	**10**	**10**
Naima Mohamed	**10**	**10**	**X**	**10**
Deon Winston E	**10**	**10**	**10**	**X**