Breast cancer prediction using KNN approach

SHADHIN.MD.YOUSUF ALI
dept. Computer Science
AIUB
Dhaka, Bangladesh
20-42783-1@student.aiub.edu

TAHSIN FERDOUS
dept. Computer Science
AIUB
Dhaka, Bangladesh
20-43413-1@student.aiub.edu

SYEDA TASNIM CHOWDHURY
dept. Computer Science
AIUB
Dhaka, Bangladesh
19-41161-2@student.aiub.edu

TAMALIKA BOWMICK
dept. Computer Science
AIUB
Dhaka, Bangladesh
19-41120-2@student.aiub.edu

Abstract*—Breast cancer is one of the most serious illnesses affecting women worldwide. However, when this deadly illness can be found in its early stages, it can save many people's lives. A mass of aberrant tissue is called a tumor. Breast cancer tumors can be classified as either "benign" or "malignant," depending on whether they are cancerous or not. Mammography images are used by radiologists to determine whether breast cancer is present or not. For the purpose of diagnosing breast cancer in particular, the science of bioinformatics makes use of machine learning techniques. The most widely used supervised machine learning algorithms—K-Nearest Neighbors, SVM, and Logistic Regression—are experimented with in this study. KNN was developed for various k-folds in order to classify the disease of breast cancer. K values with cross-validation. The classification accuracy results were then compared to those using logistic regression. Using data from the UCI Machine Learning Repository's Breast Cancer Data Set (BCD), this study makes predictions about breast cancer. The KNN method was used in the proposed study to attain a highest accuracy of 93.68% and a lowest accuracy of 90.35%.

Keywords: Image Processing, Benign, Malignant, Mammography, Machine Learning, K-nearest neighbor.

I. INTRODUCTION

Early detection of cancer is crucial for a quick response and a better chance of recovery. Unfortunately, because there are no symptoms at first, early detection of cancer is frequently challenging. As a result, many researchers continue to focus on the study of cancer in an effort to produce data that will advance diagnosis, therapy, and prevention.

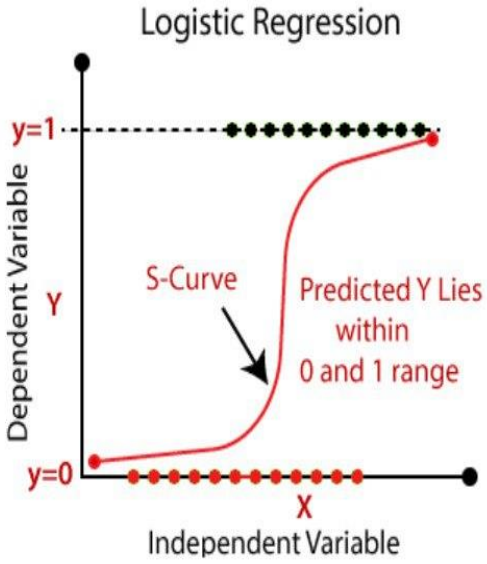
Research in this field is a search for knowledge through surveys,
studies and experiments conducted with applications to discover and interpret new knowledge in order to prevent and minimize the risk of adverse consequences. To better understand this issue, tools are still needed to help oncologists select the treatment needed to cure or prevent recurrence by reducing the harmful effects of certain treatments and their cost.

A. LOGISTIC REGRESSION

A supervised machine learning method used for classification tasks is logistic regression (for making predictions based on training data). Similar to linear regression, logistic regression also employs an equation, but the outcome is a categorical variable as opposed to a value in other regression models. Binary outcomes can be predicted from the independent variables. The result of the dependent variable is discrete. Logistic regression uses a simple equation that shows the linear relationship between the independent variables. These independent variables along with their coefficients are linearly combined to form a linear equation that is used to predict the output [8]. The equation used by the basic logistic model. The logistic function is used here to suppress the result value between 0 and 1. The logistic function can also be called a sigmoid function or a cost function. The logistic function is a shaped curve that takes the input (numeric

value) and changes it to a value between 0 and 1[9].

B. Fig

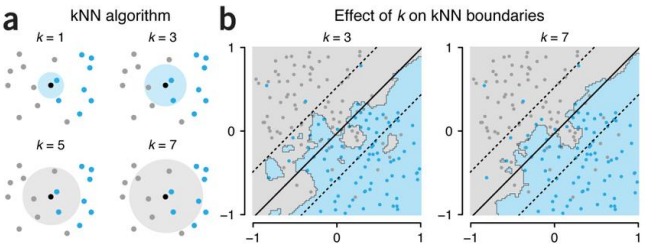


B. SUPPORT VECTOR MECHINE

A supervised machine learning approach called Support Vector Machine (SVM) is used for both classification and regression. Although we also mention regression issues, classification is the best application. The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points.

C. K-NEAREST NEIGHBORS METHOD

Because of the labeled data that is provided to K-Nearest Neighbor, it is a supervised machine learning algorithm. It is a non-parametric method since the closest training data points are used to classify test data points rather than the dataset's dimensions (or parameters). It helps with classification and regression problems. Using the k closest training examples in the feature space, the classification algorithm categorizes the objects. The underlying theory of KNN makes the assumption that the same data points are present in the same setting. It lessens the work required to create a model, modify a set of parameters, or add extra assumptions. Based on a mathematical calculation called Euclidean distance, it captures the notion of proximity.



1. THE CHOICE OF THE PARAMETER K (THE NUMBER OF NEAREST NEIGHBORS)

An object that needs to be classed is put in the class that best represents its closest neighbors. The data point is included in the category with just one nearest neighbor if k equals 1. The distances between each new input data point and every other point in the training dataset are calculated. The training set data points that are closer to the test data point in terms of distance are regarded as our test data's closest neighbors. Finally, the test data point is placed into one of its nearest neighbor classes. Therefore, the classification of the test data point depends on the classification of its nearest neighbors [8]. The most important step in putting the KNN algorithm into practice is selecting the value of K. Depending on the type of record, each record's K value is different and not constant. The stability of the forecast is poorer when K is smaller. In the same way, if we raise its value, the ambiguity will be diminished, resulting in more stable borders. When using KNN, the Ks value alone determines which category a new data point is assigned to. K is the number of nearby training data points that are the closest to a given test data point, and the test data point is then allocated to the class with the greatest number of nearby training data points (i.e., high frequency class).

II. PROPOSED BREAST CANCER DIAGNOSIS MODEL

The Wisconsin Breast Cancer datasets from the UCI Machine Learning Repository is used, to distinguish malignant (cancerous) from benign (non-cancerous) samples. website where the dataset was collected in python’s format. Dataset file was in csv format. There are 8 groups and 699 information into the chronological grouping of the data.

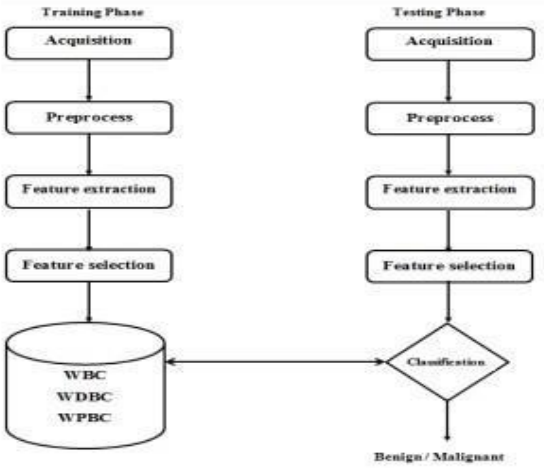


Fig. 03. Proposed Model

III. DATASET DESCRIPTION

As Dr. Walberg recounts his clinical cases, samples are periodically delivered. Therefore, this chronological grouping of the data is reflected in the database. The grouping information can be seen just below.

Dataset	No. of Attributes	No. of Instances	No. of Classes
Wisconsin Breast Cancer (Original)	11	699	2

Table. 01. Description of Breast Cancer Dataset

having been eliminated from the original data. KNN was used to categorize breast cancer disease and was implemented for various k values and k-fold cross-validation. The classification accuracy results were then compared to those using logistic regression.

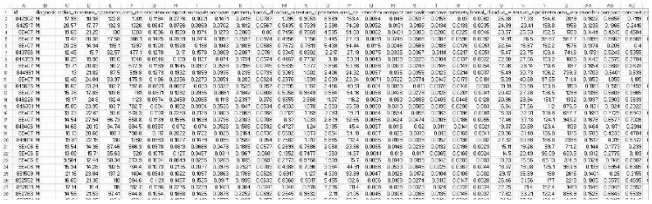


Fig. 04. Wisconsin Dataset

Total amount of malignant and benign data in the Wisconsin dataset is shown in Figure 2.

IV. FEATURE EXTRACTION AND SELECTION

An important step in the breast cancer diagnostic model is feature extraction. The optimal feature set should have effective and distinctive features while mainly reducing the redundancy of the feature space to avoid the problem of the curse of dimensionality. The curse of dimensionality suggests that the sample density of the training data is too low to promise a meaningful estimate of a high-dimensional classification function with the available finite number of training data.

A. Wisconsin Breast Cancer Dataset [9]

The details of the attributes found in this dataset listed in table

In clump thickness, benign cells tend to be grouped in monolayers while cancer cells are often grouped in multilayers. Whereas in the uniformity of cell size/shape, the

	Attribute	Domain
1	Sample code number	id number
2	Clump Thickness	1 - 10
3	Uniformity of Cell Size	1 - 10
4	Uniformity of Cell Shape	1 - 10
5	Marginal Adhesion	1 – 10
6	Single Epithelial Cell Size	1 – 10
7	Bare Nuclei	1 – 10
8	Bland Chromatin	1 – 10
9	Normal Nucleoli	1 – 10
10	Mitoses	1 - 10
11	Class	2 for benign, 4 for malignant

Table. 02. Wisconsin Breast Cancer Dataset Attributes

cancer cells tend to vary in size and shape. Because of this, these parameters are valuable in determining whether the cells are cancerous or not. In the case of marginal adhesion, the normal cells tend to stick together while cancer cells tend to lose this ability. Loss of adhesion is thus a sign of malignancy. In the case of the single epithelial cell size, the size is related to the uniformity mentioned above. epithelial cells that are significant enlarged may be a malignant cell. The Bare nuclei is a term used for nuclei that is not surrounded by cytoplasm (the rest of the cell). Those are typically seen in benign tumors.

Total amount of malignant and benign data in the Wisconsin dataset is shown in Figure 5.

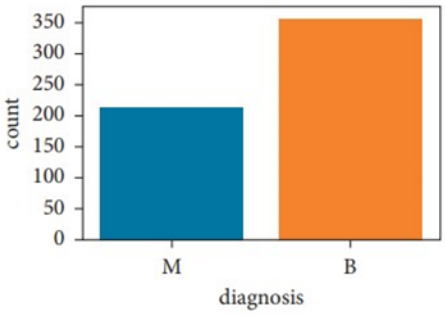


Fig. 05. Total number of malignant and benign data

The Bland Chromatin describes a uniform "texture" of the nucleus seen in benign cells. In cancer cells the chromatin tends to be coarser. The Normal nucleoli are small structures seen in the nucleus. In normal cells the nucleolus is usually very small if visible. In cancer cells the nucleoli become more prominent, and sometimes there are more of them. Finally, Mitoses is nuclear division plus cytokines and produce two identical daughter cells during prophase. It is the process in which the cell divides and replicates. Pathologists can determine the grade of cancer by counting the number of mitoses.

B. Wisconsin Diagnosis Breast Cancer (WDBC) [14]

The details of the attributes found in WDBC dataset: ID number, Diagnosis (M = malignant, B = benign) and ten real valued features are computed for each cell nucleus: Radius, Texture, Perimeter, Area, Smoothness, Compactness, Concavity, Concave points, Symmetry and Fractal dimension [15]. These features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe

characteristics of the cell nuclei present in the image [16]. When the radius of an individual nucleus is measured by averaging the length of the radial line segments defined by the centroid of the snake and the individual snake points. The total distance between consecutive snake points constitutes the nuclear perimeter. The total distance between consecutive snake points constitutes the nuclear perimeter. The area is measured by counting the number of pixels on the interior of the snake and adding one-half of the pixels on the perimeter. The perimeter and area are combined to give a measure of the compactness of the cell nuclei using the formula. Smoothness is quantified by measuring the difference between the length of a radial line and the mean length of the lines surrounding it. This is similar to the curvature energy computation in the snakes. Concavity captured by measuring the size of the indentation (concavities) in the boundary of the cell nucleus. Chords between nonadjacent snake points are drawn and measure the extent to which the actual boundary of the nucleus lies on the inside of each chord. Concave Points: This feature is Similar to concavity but counted only the number of boundary point lying on the concave regions of the boundary. In order to measure symmetry, the major axis, or longest chord through the center, is found. Then the length difference between lines perpendicular to the major axis to the nuclear boundary in both directions is measured. The fractal dimension of a nuclear boundary is approximated using the "coastline approximation" described by Mandelbrot. The perimeter of the nucleus is measured using increasingly larger "rulers". As the ruler size increases, decreasing the precision of the measurement, the observed perimeter decreases. Plotting log of observed perimeter against log of ruler size and measuring the downward slope gives (the negative of) an approximation to the fractal dimension. With all the shape features, a higher value corresponds to a less regular contour and thus to a higher probability of malignancy. The texture of the cell nucleus is measured by finding the variance of the gray scale intensities in the component pixels.

V. DATA UNDERSTANDING AND DATA SELECTION

WBCD consisted of 699 instances and 11 features. These 11 features provide precise information pertaining to the occurrence of breast cancer. Moreover, the dataset was scrutinized for unknown values, inconsistency and erroneous data. Unknown values can have a consequential effect on the interpretations that can be derived from the data. 16 instances with missing values are present which are denoted by"?" in the Breast cancer dataset.

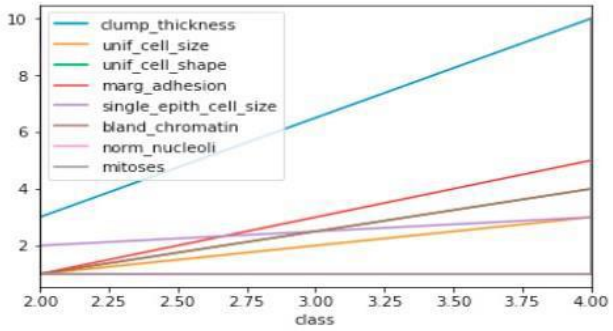


Fig. 06. Collective data variability with respect to class

Every missing value is substituted with a constant random number and ignored throughout analysis. Understanding the distribution of data across datasets is a highly important task that must be completed successfully. Data distribution analysis reveals a variety of intriguing relationships and insights that can be helpful in choosing the optimal predictive feature.

VI. TRAINING AND CLASSIFICATION

Data sets are categorized according to specified characteristics that sample variables must have in order to be able to categorize them, and each sample variable is given a malignant or benign class. Predictions based on known sample data that have been learned from training data are the main method of classification. The algorithm is first trained on the class labels for the existing collection of data samples, and it then applies this learning to predict the class labels for the incoming, unknown set of data samples. The classification goal for this project is to identify the best classifier for diabetes classification by enhancing accuracy utilizing LR, SVM, and KNN classifiers.

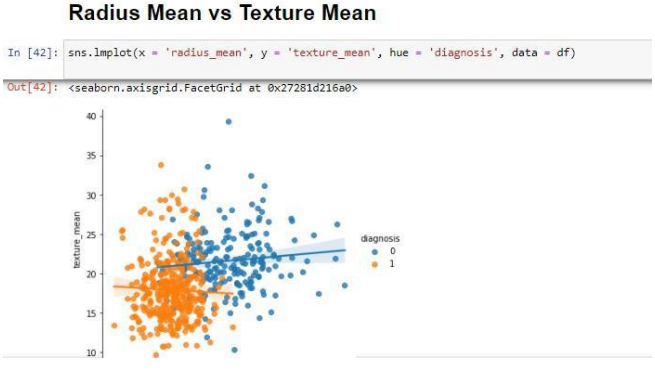
The test dataset, which contains the unknown sample required to predict the class label, is used to evaluate the classifier's performance after it has been trained using known sample data from a training dataset. K Neighbors Classifier is an instance-based, supervised classifier that gains knowledge from samples of labeled data. Algorithm 1 provides the KNN classifier's pseudo code. For training data, the K folds cross validation procedure is employed. This method divides the original sample into k equivalent size subsamples, one of which is used to validate the model while the remaining k-1 subsamples serve as training data. Following that, this cross-validation procedure is repeated k times (referred to as the folds), with each of the k subsamples being utilized as the validation data just once. It operates in a loop. The value of k was fixed at 10 for this study.

VII. RESULTS AND DATA ANALYSIS

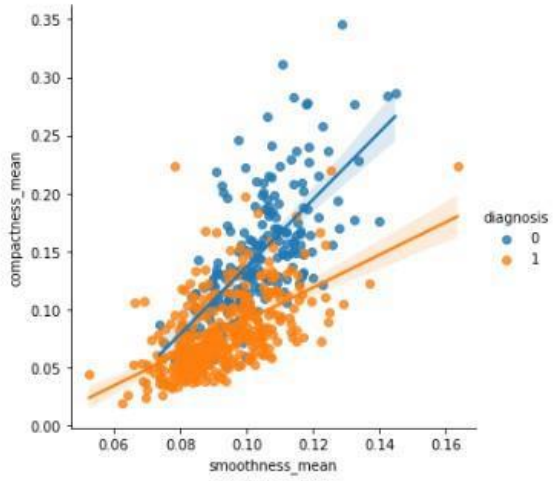
This section covers all methods and materials, as well as the representation of the dataset, the block graph, the stream chart, and the evaluation grids.

A. RESULTS

1. Radius Mean vs Texture Mean

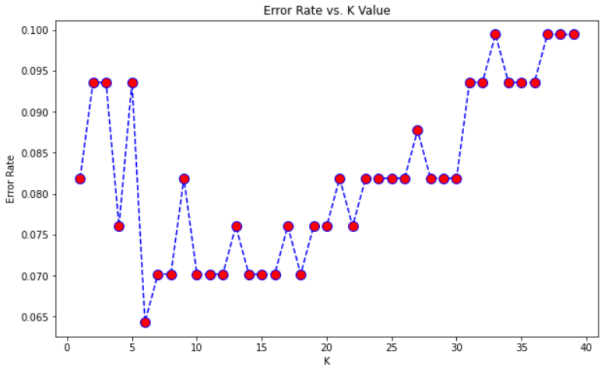


2. Smoothness Mean vs Compactness Mean

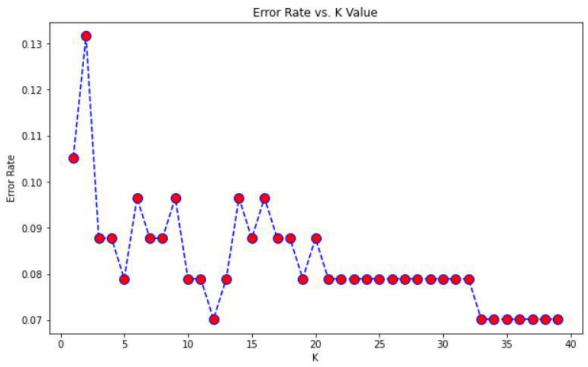


3. Error Rate vs K value

For test size 30%

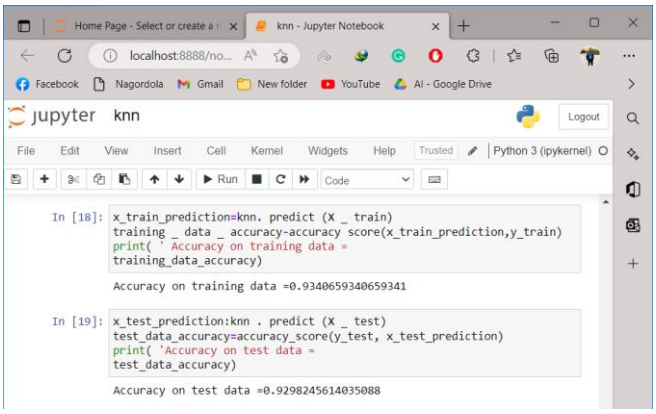


For test size 20%

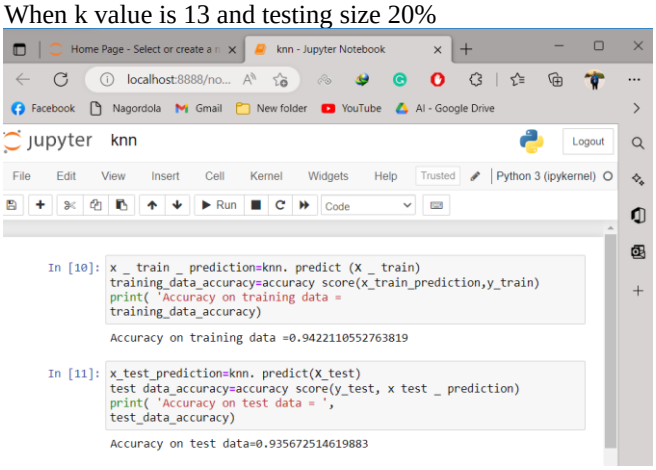


1. Training and Testing

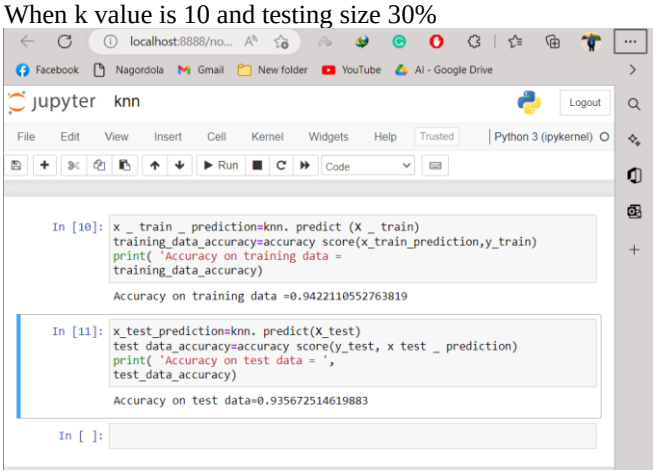
When k value is 10 and testing size 20%



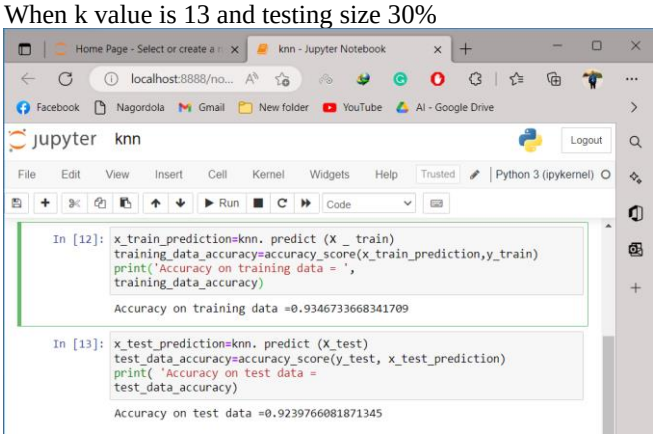
When k value is 13 and testing size 20%



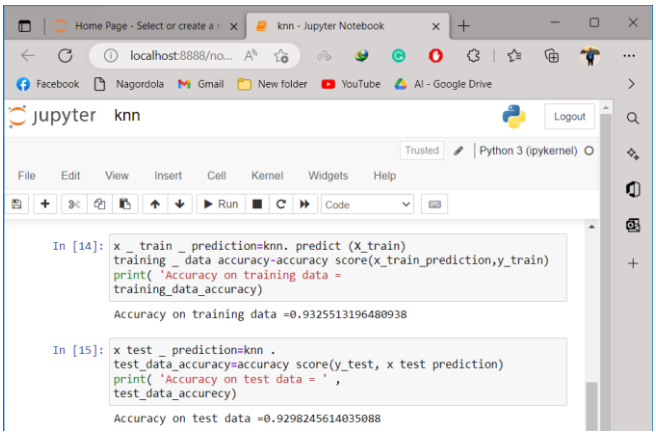
When k value is 10 and testing size 30%



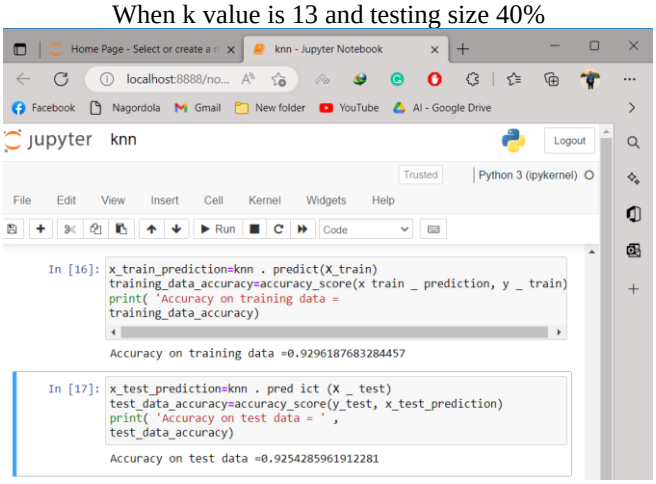
When k value is 13 and testing size 30%



When k value is 10 and testing size 40%



When k value is 13 and testing size 40%



B. Data Analysis

In order to create the training set, we selected the data at random. We have experimented with different values of K from K = 1 to 30 while adjusting the training and testing size. We have produced a graph to analyze the error rate versus k-value. The test set's classification result using the KNN algorithm varies between 89.12% and 95.02%. When K is 13 years old, the performance is at its peak.

VIII. DISCUSSION

The reason behind the conduct of this study is to predict breast cancer using the KNN approach. We used Kaggle's Breast Cancer Wisconsin (Diagnostic) Data Set consists of 699 instances in chronological format. Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. Compared to the methods reported in [12], [14] (11), [13], the advantages of the KNN algorithm are that the algorithm is very simple, and its implementation is very easy. Since there is no need of any training session, there is no convergence problem. In contrast, the other approaches employing neural networks may face the convergence problem, and may need long training time. New training data can also be added to the KNN algorithm without any retraining. But for the other techniques, adding new training data needs retraining because the new training data disturb the structure of the existing training set, and all the parametric or semiparametric classifiers critically depend on this structure.

IX. CONCLUSION

This paper treats the Wisconsin-Madison Breast Cancer diagnosis problem as a pattern classification problem. The nonparametric classifier is based on the KNN method. Based on the class label that the majority of the K-closest training data have, the KNN algorithm assigns the class label of the new datum. The KNN algorithm produces the best classification results so far for this issue. Conclusion: The KNN algorithm's successful performance does not guarantee that it will always work well for all diagnosis issues. In actuality, no single method is found to be effective for all diagnosing issues. In the future, we'd like to look into more sophisticated algorithms in relation to the Breast Cancer problem and other pertinent issues.

X. ACKNOWLEDGMENT

Thank are due to Dr. William H. Walberg at the University of Wisconsin for supporting us with the breast cancer dataset which we have used in our experiments

REFERENCES

[1]HTTPS://WWW.NATIONALBREASTCANCER.ORG/BREAST-TUMORS/
[2]HTTPS://WWW.EJMCM.COM/ARTICLE_5172_967A0564C5EFCD9CA4D9CA59189E807A.PDF
[3] J. S. Snchez, R. A. Mollineda, and J. M. Sotoca. An analysis of how training data complexity affects the nearest neighbor classifiers. Pattern Analysis and Applications, 10(3), 2007.K. Elissa, "Title of paper if known," unpublished.
[4] M.Raniszewski. Sequential reduction algorithm for nearest neighbor rule. Computer Vision and Graphics, 6375, 2010.
[5]D.CoomansandD.L.Massart.Alternativek-nearest neighbour rules in supervised pattern recognition. Analytica Chimica Acta, 136, 1982.
[6] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.
[7] Onel Harrison," Machine Learning Basics with the KNearestNeighbors Algorithm "
[8] Mohammad Bolandraftar and SadeghBafandehImandoust - "Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background"- International Journal of Engineering Research and Applications Vol. 3, Issue 5, Sep-Oct 2013
[9] Frank, A. & Asuncion, A. (2010). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science
[10] Angeline Christobel. Y, Dr. Sivaprakasam (2011). An Empirical Comparison of Data Mining Classification Methods. International Journal of Computer Information Systems,Vol. 3, No. 2, 2011.
[11] International Conference on Computational Intelligence and Data Science (ICCIDS 2018) Breast Cancer Prediction system Madhu Kumaria, Vijendra Singhb b a Department of Computer Science and Engineering, The NorthCap University, Sector 23A,Gurugram, Haryana, 122017, India

[12] 4] Yeh WC, Chang WW, Chung YY. A new hybrid approach for mining breast cancer pattern using discrete particle swarm optimization
[13] and statistical method. Expert Systems with Applications. 2009 May 1;36(4):8204-11.
[14] [5] Marcano-Cedeño A, Quintanilla-Domínguez J, Andina D. WBCD breast cancer database classification applying artificial
[15] metaplasticity neural network. Expert Systems with Applications. 2011 Aug 1;38(8):9573-9.
[16] [6] Kaya Y, Uyar M. A hybrid decision support system based on rough set and extreme learning machine for diagnosis of hepatitis
[17] disease. Applied Soft Computing. 2013 Aug 1;13(8):3429-38.
[18] [7] Nahato KB, Harichandran KN, Arputharaj K. Knowledge mining from clinical datasets using rough sets and backpropagation neural
[19] network. Computational and mathematical methods in medicine. 2015;2015.
[20] [8] Liu L, Deng M. An evolutionary artificial neural network approach for breast cancer diagnosis. InKnowledge Discovery and Data
[21] Mining, 2010. WKDD'10. Third International Conference on 2010 Jan 9 (pp. 593-596). IEEE.
[22] [9] Chen HL, Yang B, Liu J, Liu DY. A support vector machine classifier with rough set-based feature selection for breast cancer
[23] diagnosis. Expert Systems with Applications. 2011 Jul 1;38(7):9014-22.
[24] Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers Gouda I. Salama1 , M.B.Abdelhalim2 , and Magdy Abd-elghany Zeid3 1, 2, 3Computer Science, Arab Academy for Science Technology & Maritime Transport, Cairo, Egypt dr_gouda80@yahoo.com1 , mbakr@ieee.org2 and Magdy_zeid83@yahoo.com3