Breast Cancer Prediction Using KNN-Classification

Sirichandana Bomma
Department of Information Science
University of North Texas
Denton, Texas, USA
Email: SirichandanaBomma@my.unt.edu

***Abstract*—Breast cancer is one of the foremost prevailing malignancies in women. Early identification of carcinoma is crucial, so we can reduce the number of fatalities by the first identification. Breast cancer is split into sorts benign and malignant. Benign is non-cancerous, whereas, malignant is cancerous. To determine if the tumor is malignant or benign, we develop a predictive model using machine learning methods. The early identification of this fatal disease can save the lives of many women. Radiologists use mammography images to determine the presence and absence of ovarian cancer. In this project, the detection of breast cancer is done by the application of machine learning techniques in the field of bioinformatics. Our goal is to create an advanced, automated diagnostic tool that can correctly and repeatedly determine whether the tumor is benign or malignant and that can say whether the tumor is death-causing or not. We have applied the K-Nearest Neighbors (KNN) algorithm to the practice of error analysis and distance measurement for various values of K. We classify the collected dataset(Breast-Cancer.csv) to different groups i.e., Manhattan, and Euclidean distances. Then, we generate a classification report and calculate the best values of Precision, recall, f1-score, and support K. We predict the feature 'STAGE', which let us know the stage of breast cancer by which we can say whether it is death-causing or not. We examine every variant for accuracy and consider the most accurate forecast for the outcome. We will then implement KNN by using the Minkowski distance metric at some values of 'K', which will result in the accuracy of the KNN. The results we got clearly show how many patients are suffering from death-causing breast cancer and how many are not. We calculated accuracy and error at every K value and got the best accuracy and least error at K = 1 79%, 21% respectively which tells us breast cancer disease was classified successfully using K-Nearest Neighbors algorithm.**

***Index Terms*—Breast Cancer, Mammography, K Nearest Neighbor (KNN) Algorithm, Minkowski Distance, Manhattan Distance, Euclidean Distance, death-causing, Accuracy.**

I. INTRODUCTION

On a global scale, cancer is the second biggest killer. Breast cancer, which is among the most common form of cancer, is the second most prevalent. Breast cancer can afflict both males and females, however, it is noted that women tend to be the disease's primary victims. It's malignant cancer that develops from the stromal and infundibulum cells of the breast. A cancerous tumor is a collection of tumor cells that have spread to other organs and affected surrounding tissues [1]–[15]. The most significant increases in fatty food intake, later ages at childbirth, fewer children, shorter length of breastfeeding, dread of self-examination, fear of chemotherapy, and fewer children are the primary risk factors for breast cancer [5]. Breast cancer symptoms lead to thyroid, lymphoma, and melanoma. Early diagnosis is the best method to prevent and fight against breast cancer [2]. Early identification of breast cancer is good for prevention. For early identification, medical experts need to classify the symptoms and predict whether the patients have a chance of getting breast cancer or not which requires mammogram tests [1]–[15].

Currently, a biopsy is the only surefire approach to rule out breast cancer when a mammography test indicates abnormal findings. A breast biopsy is a procedure where fluid or tissue is taken from a questionable location. It gives a sample of tissue that enables medical professionals to recognize and treat anomalies in breast lump-causing cells, other odd breast alterations, or troubling mammography or ultrasound results. The cells are taken out, analyzed under a microscope, and tested for breast cancer afterward. The three types of biopsies that are typically conducted are surgical, core needle, and fine-needle aspiration. According to the location, size, form, patient medical history, or other characteristics of the anomaly, one type of biopsy will probably be suggested over another [1]–[15]. Breast cancer screening by biopsy has several drawbacks due to some side effects like Breast Sagging, Breast burning, infections, the appearance of removed tissue, and soreness [1].

While anomalies in breast tissue can be detected by mammography, it takes more time and may not get 100% results without a biopsy. This is because a cell analysis is required to determine the type of cells involved in melanoma in addition to the degree, or score, of the specific type. Furthermore, it is thought to be good approach to know for certain if you have cancer before starting a treatment protocol. The number of annual biopsies will substantially decrease if there is a different way to anticipate breast cancer [1]–[15].

V. Rodriguez, K. Sharma, and D. Walker carried out research on the mammographic dataset using KNN Classification. Their goals were to retrieve the precision, recall, specificity, and F1 Score and compare results of 6 distance metrics such as Euclidean, Manhattan, Minkowski, Chebyshev, cosine distances, and cosine similarity of the model for different K values and show which distance gives the best accuracy at all K values. Prior to the algorithm, the authors don't perform any data cleaning methods for the missing attributes. The results showed that Manhattan distance is the distance metric that results from the highest accuracy at every K value and

the average of each accuracy is 80.75% [1].

B. S. Kumar, T. Daniya, and J. Ajayan have conducted an experiment on a small breast cancer dataset with 20 records for which they have used KNN Classification. Their goal is to compare Euclidean distance, and Manhattan distance and calculate the accuracy of the dataset. Prior to the comparison they used mean imputation for data pre-processing i.e., replaced missing data with the mean of the non-missing data. It resulted in an accuracy of 83.33%. But it was a disadvantage that they have not calculated the accuracy for different K values and does not show the comparison result [2].

Sasmitha Kularathne did work on breast cancer detection and visualization using the KNN classification technique. The article describes how to use the control of the K-Nearest Neighbor classification model to display the data with data exploration and review the outcomes of the KNN model to determine which characteristics are most competent of occurring as a risk of breast cancer that uses the data set. Based on information prediction and analysis, the author was using essential methods to evaluate the diagnostic accuracy of risk of breast cancer. The classification study and confusion matrix in the analysis chapter clearly demonstrated the anticipated model's accuracy score of 96.4% [13].

M.D. Baktha Vachalam Dr. S. Albert Antony Raj carried out research on the observation of breast cancer using K-Nearest Neighbor with Distinct Distance Measures and Classification Techniques Using Machine Learning. By employing the two Master Methodologies as Naive Bayesian Classifier and K-Nearest Neighbors Method, Results are compared. "The Comparison Study shows KNN has the greatest Accuracy score of 97.51% and NB has a decent accuracy of 96.19%. In high accuracy and processing time, the KNN classifier comes out on top [14].

The KNN Classification for the Wisconsin breast cancer dataset was executed by Zohaib Mushtaqa, Akbari Yaqubb, Shaima Sanic, and Adnan Khalidd. The primary objective of this paper is to investigate KNN effectiveness using distance measures and k-values in order to find the most fruitful KNN. The authors utilized two datasets in this article: Wisconsin breast cancer (WBC) and Wisconsin diagnostic breast cancer (WDBC). The authors used three feature models: one without feature selection, the second with L1-norm selection from the approach, and finally with Chi-square feature selection. For both datasets, the author conducted three of those features, and the Chi-square-based feature selection reached the best accuracy with the Canberra or Manhattan distance functions [4].

Dr.G. Wiselin Jiji and Jini.R. Marsilin applied the CBIR approach to breast cancer prediction. The article proposes a method for retrieving mammogram images using a pattern resemblance scheme. The authors had done recovery based on feature extraction, KNN classification, pattern instantiation, and computation of pattern similarity, with each part considered as a pattern. Finally, for image retrieval based on similarity to the search query, pattern similarity is approximated [5].

Breast Tissue Characterization Using Combined KNN Classifier was developed by K. Vaidehi and T.S. Subashin. Feature extraction is performed in this work, but statistical characteristics such as Standard Deviation (SD), Mean, Kurtosis, and Skewness are used. The authors used three distinct distance measures, with the majority of the two distances regarded as the end outcome. The proposed method includes Feature Selection and classification for this investigation, 322 mammograms Were used from the Mini-Mias database. "In comparison, K-NNOP with City block performed better for Both dense i.e 94.1% and Fatty i.e 91.72%. Overall K-NN accuracy was attained is 91.72% [15].

The authors (Meriem AMRANE, Saliha OUKID, Ikram GAGAOUA, and Tolga ENSARO) collaborated on the Classification Of Breast cancer using the Wisconsin breast cancer database. Machine Learning is being used to determine whether a patient's tumor is benign or malignant. In this experiment, researchers used two machine learning classifiers: the Nave Bayesian Classifier and the k-nearest neighbor. After comparing the two algorithms thoroughly, the studies found that KNN has a higher efficiency of 97.51%, while NB has a slightly lower accuracy of 96.19% [7]. To forecast breast cancer by combining IoT and machine learning with the Wisconsin breast cancer database, the authors (V. Nanda Gopal, Fadi Al-Turjman, R. Kumar, L. Anand, and M. Rajesh) undertook feature selection and classification. The coefficient of the correlation function and various filtration approaches were used to pick the features for effective feature extraction. Machine learning methods require classification. In this proposed method, three different categorization analysis kinds were used. Models like Multi-Layer Perception Classifier, Random Forest, and Logistic Regression, as well as 10 cross-fold techniques, were created for classification analysis. According to the findings, the Multi-Layer Perception classifier performs better than the other models at identifying breast cancer cells. The construction of the correlation matrix utilizing Principal Component Analysis in feature extraction has an effect on the F-Measure and accuracy measures [8].

Breast cancer diagnosis based on K-Nearest Neighbors: A Review by Khorshid, Shler & Mohsin Abdulazeez, Adnan and Mohsin, Adnan The KNN algorithm's function as a classifier for mammography images and how precise it is at detecting dangerous breast lesions were the authors' main areas of interest. In this research, researchers looked for the most recent publications examining K-effectiveness NNś as a machine learning algorithm for detecting breast cancer. They also suggested that KNN can be easily implemented. KNN provided the most accurate prediction (99.12%) in terms of accuracy [6].

A study was conducted for Wisconsin Breast Cancer Diagnostic dataset where the goal is to compare Min-Max Normalization and Z-Score normalization and calculate accuracy. Prior to that, the data results got classification problems while in the data preparation stage which is cleaned

by normalization. The comparison results in the best accuracy for min-max Normalization technique [12].
Machine learning research was conducted on breast cancer risk variables, and prediction models were put into practice to assess the likelihood that breast cancer patients will survive. The classification of normal and pathological cells was accomplished by utilizing three models: NB, RBF network, and J48. [9]

In this paper, we are going to use the KNN algorithm to classify breast cancer whether it is dangerous are not. We classify into groups and calculate Euclidean Distance and Manhattan distance. We are not using any Normalization techniques for checking the accuracy of our model. We are using only two distance metrics to check the accuracy for every K value. We are focusing on the following question "How can we say whether a cancer tissue is death-causing or not?"

II. THE METHODOLOGY

Dataset Description: A dataset is the collection of data in a meaningful format. The dataset we use for this project is the Wisconsin Breast Cancer dataset which is created by Dr. William H. Wolberg. We have collected the data from Kaggle where we have free access to datasets. The dataset contains information about patients who have symptoms of breast cancer. These features are calculated from a mammogram image of a Fine Needle Aspirate (FNA) of a breast tumor. They characterize the traits of the visible cell nuclei in the picture.
The dataset we collected had 569 instances and 33 attributes. By analyzing the data, we can observe that there is only one instance Unnamed that contains 569 missing values. The attributes in the dataset include ID, Diagnosis, radius_mean, texture_mean, radius_se, radius_worst, Unnamed, and means, standard error, and worst of all the symptoms where ID acts as the key attribute i.e. when a person wants to know the information about a patient that person should only need the id by that they will get all the information of the patient. Whereas Diagnosis is the attribute that explains whether the patient has breast cancer or not. From the dataset, out of 569 instances of Class 357 are Benign which is 62.74% of the total data and 212 are Malignant which is 37.26% of the total data.
While using a model for the project, it is necessary to train the model. So, we are reserving 75% of the dataset for training and 25% of the dataset for testing. Here, we will use the training set to train the model and the testing set for applying the trained model on it and testing for further execution. As the class feature contains information about the diagnosis of the patient whether the tissue is cancerous or not, we are taking it as the dependent variable the model will predict whether it is malignant or benign based on all other features which show the symptoms of the breast tissue of a particular person.
We may get biased if the training set is poor, or model performance mismatches. In our project, we are trying to recognize whether a person has cancerous tissue in her breast or not, but it may lead to a mismatch due to high-resolution images and make incorrect decisions while predicting. We are going to avoid bias by choosing the best Machine Learning Model which is K Nearest Neighbor (KNN) Classification, using the right training dataset by applying a random state while splitting the data and making the dataset balanced, and performing data pre-processing. As our data have an attribute with all values missing, it is nowhere to reflect the other instances. So, we have removed the complete column to make the dataset clean.
As per the dataset description, we have 33 attributes out of those one is dependent which we are going to predict and the other 31 are independent attributes and one attribute has all missing values. So, we can say that there will be a relationship between every attribute then only decision-making is possible. In our project, we are predicting the diagnosis attribute based on other attributes. The diagnosis attribute has two values Benign and Malignant. Benign means the tumor is non-cancerous, and Malignant means the tumor is cancerous. The sample dataset and count of benign and malignant are shown below with sample images.

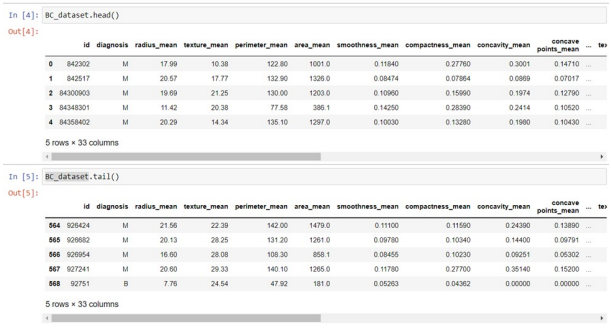

Fig. 1. sample dataset

```
B    357
M    212
Name: target, dtype: int64
```

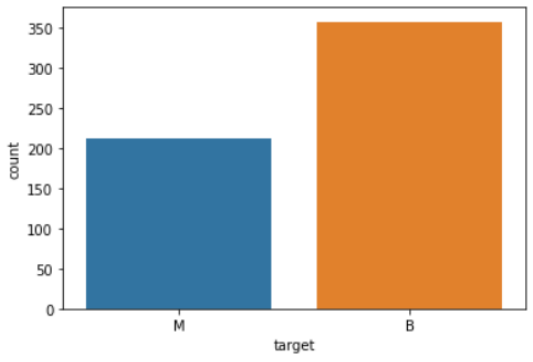

Fig. 2. count

As the dataset is prepared based on the breast tissue of a patient it has taken radius (how thick the tissue is), texture, perimeter, area, Smoothness, compactness, Concavity, concave points, Symmetry, and fractal dimension. Based on all these features of the tissue, the Machine Learning model can predict the tumor. For the prediction, it is calculated four means, standard error, & worst (shows how bad the situation of breast tissue is) of all these features and included in the dataset. The radius is the distance from the center point to the tumor which shows the thickness of the tissue. The perimeter shows the size of the tumor. These two are calculated based on the mammogram images and all the other attributes are calculated based on the radius, perimeter, and texture. Based on the original values, the mean is calculated, and based on the mean standard error is calculated and these two are applied to calculate the worst. The below images show the relationship between some attributes.

Fig. 3. Confusion Matrix to show the relationship between some attributes

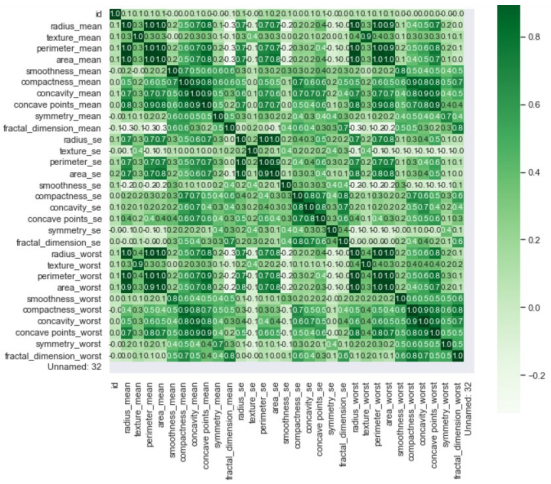

Fig. 4. Heatmap to show the relationship between some attributes

A dataset is said to be imbalanced when the target variable has uneven distributions i.e., a class with a high number of observations and another class with a low number of observations. An imbalanced dataset is not always bad for real-time datasets, because real-time datasets always have some degree of imbalance. The developer can make the dataset balanced by using evaluation metrics such as Precision, Recall, and Confusion Matrix. When a target variable reaches a 2:1 ratio, the dataset is said to be imbalanced. A balanced dataset always has a target variable 1:1 ratio and doesn't reach a 2:1 ratio. From Fig.2, we can observe that the target variable has two values B and M and there is no high difference between the classes inside it. It doesn't reach 2:1. As the data is already balanced, we are not performing any evaluation metrics.

We can predict breast cancer by using different Machine Learning algorithms like Support Vector Machine (SVM), Naive Bayes (NB), Decision Trees using Random Forest Algorithm, and KNN Classification.

Data Pre-processing: When the data is collected from different sources, it will be collected in the raw format, which is not potential for analysis. So, pre-processing data acts as the first and most important step in Machine learning, Data pre-processing includes data cleaning, dimensionality reduction, and data transformation. Whenever the data is in raw format, it should be converted into understandable ways like images, videos, and tables. The tabular data is much better than others. The data will have errors when collected from different external sources, so data cleaning is important to make the data error-free and clean. Data processing makes the data more meaningful, clean, and impactful for further decision-making. As data cleaning makes the data accurate, the data which is applied for data pre-processing will give better results than other data.

We can clean the data by using the following techniques.

1) Removing the Missing values and replacing them with meaningful data.
2) Ignoring the noisy values by replacing them with a random keyword.
3) Finding the Mean value (Average value within the column) of a column and replacing the missing with the Mean value.
4) Finding the Median value of a column and replacing the missing rows with the Median value.

The Dataset we are using for this project has 33 attributes where the values in the 33rd attribute "Unnamed" are all missing. To make the given dataset more meaningful, we are removing this attribute as it has all missing values/ null values which are nowhere related to any other attribute. We can't even consider it as an independent attribute or the dependent variable as it is related to any of the columns in the given dataset.

METHODS:

Support Vector Machine: support Vector Machine is a supervised machine learning algorithm that is used to classify the data, and identify the errors or outliers in the data based

on the feature of the data and decision boundary.

Naive Bayes: Naive Bayes is used to solving Classification problems based on two parameters i.e., Maximum Likelihood Error (MLE) and Maximum Posterior (MAP) Once the input data seems to be very dimensional, the Naive Bayes Classifier is used. This approach is incredibly helpful in computer vision tasks.

Decision Tree: In supervised machine learning, which trains models using labeled input and output datasets, decision trees are a method employed. The method is mostly used to address classification issues, which include categorizing or classifying an object using a model.

Random Forest: Random Forest is a method used by the decision tree algorithm that is used for decision-making by building the tree. It used the K-cross fold technique for dataset validation calculating the accuracy of the model. From the literature survey, after going through those papers, we got that KNN Classification produces better results with breast cancer prediction. So we are using KNN Classification for our dataset.

Implemented Algorithm:

KNN Classification:

KNN Classification is a supervised Machine Learning algorithm that is used to solve regression and classification problems based on its nearest neighbors. It uses different distance metrics for classification. While splitting the data, it will remember the training data which is fitted to the model and apply it for predicting the target variable or decision-making.

KNN works with the below steps:

1) Select K number of Neighbors
2) Calculate the Distance metric applied
3) Take the K values as per the calculated distance.
4) Count the data points for each K value.
5) Assign new data points to categories with maximum K value.
6) Predict the target variable.

For our model, we have taken K as 10 and worked with two distance metrics i.e., Euclidean distance and Manhattan Distance, for which we have followed the following process.

1) Loaded the Dataset
2) Calculated the Distance
3) Apply KNN Algorithm
4) Calculated the Accuracy

Euclidean Distance:

Euclidean distance is the distance between two points in a straight line.

$$euclidean\ dist = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

Manhattan Distance:

Manhattan distance is the distance between real vectors in a graph.

$$manhattan\ dist = \sum_{i=1}^{n}|p_i - q_i|$$

After Calculating the distances, we applied the KNN Classification algorithm for calculating the accuracy at different K values, then take the results in a table.

III. THE RESULTS

In this paper, we have applied the K-Nearest Neighbors algorithm to classify the breast cancer diagnosis dataset focusing on whether breast cancer is death-causing or not. The below figure shows the sample of our dataset.

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	...
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	...
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	...
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	...
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	...

5 rows × 33 columns

Fig. 5. sample dataset

We had several modifications for the given dataset as follows: First, we checked for null values in the dataset and imputed the data in a way that it does not affect nor depend on the other columns as it contains null values. Figure 6 shows the sample Evaluation Metric of the Model. In the Machine Learning KNN algorithm, we can use evaluation metrics to calculate accuracy.

As per our core question, we have done feature engineering by that we have added two extra features called “STAGE, and Death Causing”. In the STAGE column, we have provided conditions based on radius_worst and area_worst as these two are the main features to know what Stage of breast cancer is. After adding the feature Stage, we have included Death Causing which focuses on our core idea, that we have given the condition based on Stage in a way that if Stage = 0,1,2 it is not death-causing and can be controlled with medication, and if Stage = 3,4 it is Death causing. Figure 7 shows the count plot for the Death Causing column.

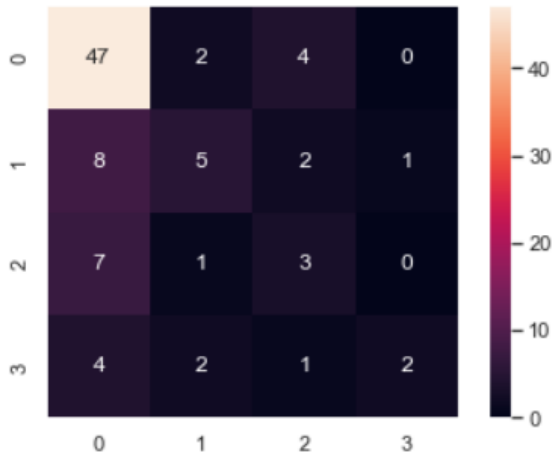

Fig. 6. confusion Matrix

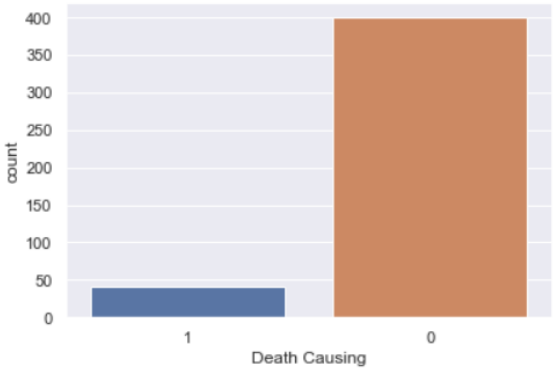

Fig. 7. count

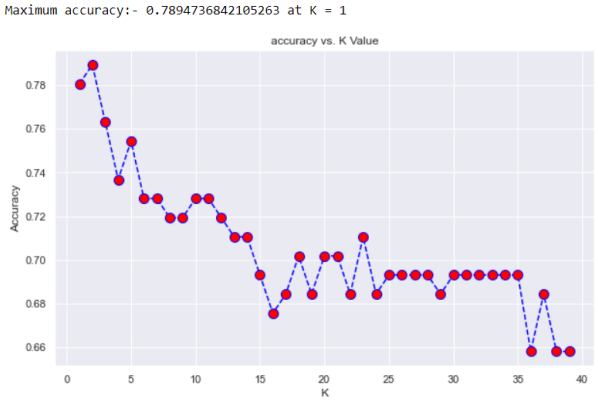

Fig. 8. Maximum Accuracy

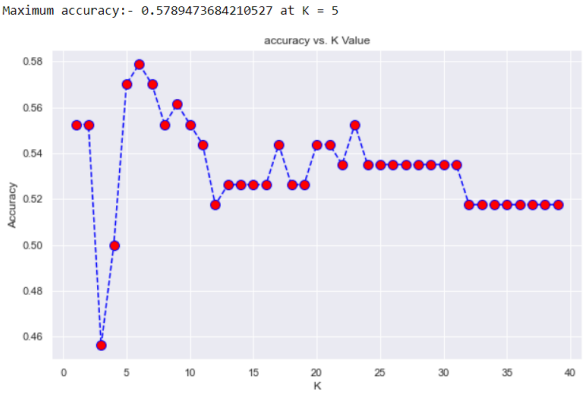

Fig. 9. Maximum Accuracy

Taking K as 10, we have generated Minkowski distance which is a generalized distance metric for Manhattan and Euclidean distance, i.e., if p = 1, Minkowski distance equals Manhattan, and if p = 2, Minkowski distance equals Euclidean distance. Figures 8 and 9 show the best accuracy and K value. The best accuracy is at K = 1 with accuracy = 79% and the least error at K = 1 with error = 21% for the original dataset. And for the dataset after feature engineering best accuracy is 58% at K = 5 and the error same as the above. We then visualized the model by constructing a confusion matrix and classification report.

Comparision with Previous Studies:

When compared [9], the Wisconsin Breast Cancer dataset is trained with Java 48, Sequential Minimal Optimization, and Naive Bayes and the best accuracy is for J48 with 75.52%. The authors in [8] says that the Error rate is less for Multilayer Perception. The researchers in [7] used the Nave Bayesian Classifier and the k-nearest neighbor to predict whether a patient's tumor is malignant or benign and discovered that KNN has a better efficiency of 97.51 percent and NB has a slightly lower accuracy of 96.19 percent. The authors in [1] calculated different distance measurements at different K values and discovered Manhattan distance gives better results at K = 7 with a result of 81.8%. In [2], the authors calculated Mean Square Error (MSE), and Root Mean Square Error (RMSE) at 2-fold, 5-fold, and 10-fold cross-validation techniques, then attained minimum MSE at K = 2, and minimum RMSE at K = 3, k = 5, and K = 6.

In our model, we used only two distance metrics and got an accuracy as 79%, We plotted a graph to show the accuracy and error rate at every K value in a single graph. We have also added an extra feature called "STAGE" to say whether breast cancer is death-causing or not based on all other features. We got 91 variables as death-causing and 478 variables as not, and we got the accuracy as 58% at K = 5 after adding the feature. The percentage of the ROC curve is 78, which says that there is a 78% of probability for ordering the target value and other values.

IV. CONCLUSION

In this paper, we have utilized K Nearest Neighbors Classification to predict breast cancer and whether the disease is death-causing or not. We have calculated two distance metrics Euclidean, and Manhattan combinedly using Minkowski distance at different K values from 1 to 10. This paper focuses on "How can we say whether a cancer tissue is death-causing or not?", To answer the question in a logical way, we have first added column "STAGE" based on which we have added "Death Causing". As Breast Cancer is of 5 stages i.e., Stage 0 to Stage 4, Stage 0 says no presence of cancer tissue, Stage 1 says the size of the tumor is 2cm, and Stage 2 says the size of the tumor is between 2 to 5cm. These three stages are of less size, they are not death-causing. Whereas, Stage 3 says the tumor size is greater than 5cm and the area spread almost 100m in the body, and Stage 4 says the same but in Stage 4 the tumor spread to multiple areas of the body. So both Stage 3 and Stage 4 are death-causing. We can cure the death-causing breast cancer by chemotherapy and there is a 99% chance of patient cure in Stage 3, and only 5% of cures for Stage 4. A patient with Stage 4 breast cancer can live up to 5 years. This approach involved feature engineering for adding extra features to answer the core focus of this paper.

REFERENCES

[1] Victoria Rodriguez, Karan Sharma, and Dana Walker. Breast cancer prediction with k-nearest neighbor algorithm using different distance measurements. *Research Gate*, 2018.

[2] B Santhosh Kumar, T Daniya, and J Ajayan. Breast cancer prediction using machine learning algorithms. *International Journal of Advanced Science and Technology*, 29(3), 2020.

[3] Can Eyupoglu. Breast cancer classification using k-nearest neighbors algorithm. *The Online Journal of Science and Technology*, 8(3):29–34, 2018.

[4] Zohaib Mushtaq, Akbari Yaqub, Shaima Sani, and Adnan Khalid. Effective k-nearest neighbor classifications for wisconsin breast cancer data sets. *Journal of the Chinese Institute of Engineers*, 43(1):80–92, 2020.

[5] Jini R Marsilin and G Wiselin Jiji. An efficient cbir approach for diagnosing the stages of breast cancer using knn classifier. *Bonfring International Journal of Advances in Image Processing*, 2(1):01–05, 2012.

[6] Shler Farhad Khorshid and Adnan Mohsin Abdulazeez. breast cancer diagnosis based on k-nearest neighbors: A review. *PalArch's Journal of Archaeology of Egypt/Egyptology*, 18(4):1927–1951, 2021.

[7] Meriem Amrane, Saliha Oukid, Ikram Gagaoua, and Tolga Ensari. Breast cancer classification using machine learning. *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*, pages 1–4, 2018.

[8] V Nanda Gopal, Fadi Al-Turjman, R Kumar, L Anand, and M Rajesh. Feature selection and classification in breast cancer prediction using iot and machine learning. *Measurement*, 178:109442, 2021.

[9] Siham A Mohammed, Sadeq Darrab, Salah A Noaman, and Gunter Saake. Analysis of breast cancer detection using different machine learning techniques. In *International Conference on Data Mining and Big Data*, pages 108–117. Springer, 2020.

[10] Tsehay Admassu Assegie. An optimized k-nearest neighbor based breast cancer detection. *Journal of Robotics and Control (JRC)*, 2(3):115–118, 2021.

[11] Walid Cherif. Optimization of k-nn algorithm by clustering and reliability coefficients: application to breast-cancer diagnosis. *Procedia Computer Science*, 127:293–299, 2018.

[12] Henderi Henderi, Tri Wahyuningsih, and Efana Rahwanto. Comparison of min-max normalization and z-score normalization in the k-nearest neighbor (knn) algorithm to test the accuracy of types of breast cancer. *International Journal of Informatics and Information Systems*, 4(1):13–20, 2021.

[13] Sasmitha Kularathne. Prediction and data visualization of breast cancer using k-nearest neighbor (knn) classifier algorithm. *Analytics Vidhya*, 2020.

[14] MD Bakthavachalam and S Albert Antony Raj. A study of breast cancer analysis using k-nearest neighbor with different distance measures and classification rules using machine learning. *European Journal of Molecular & Clinical Medicine*, 7(03):2020, 2020.

[15] K Vaidehi and TS Subashini. Breast tissue characterization using combined k-nn classifier. *Indian Journal of Science and Technology*, 8(1):23, 2015.