Breast Cancer Detection Using K-Nearest Neighbour Algorithm

Shagun Chawla[a], Rajat Kumar[a], Ekansh Aggarwal[a], Sarthak Swain[a]

[a]G.B Pant Engineering College, New Delhi, Delhi 110020, India

Abstract:
Breast cancer is one of the common occurring cancer in women across the globe, affecting about significant percentage of women at some point in their life. Even with the development of new technologies in the field of medicine and research, the accurate diagnosis of this fatal disease outcome is one of the most important tasks needed to be done till date. Our objective is to develop a sophisticated and automated diagnostic system that yields accurate and reproducible results for predicting whether a breast cancer tumour is benign (non-cancerous) or malignant (cancerous). We have implemented K-Nearest Neighbour Algorithm using various normalization techniques and distance functions at different values of K. A comparative study using various normalization techniques, i.e., Min-Max normalization, Z-Score normalization and Decimal Scaling normalization, and different distance metrics, i.e., Manhattan distance, Euclidean distance, Chebyshev distance and Cosine distance has been done. The accuracy of each variation is tested and the maximum accurate prediction is considered for the result. Highest accuracy of 98.24% is achieved, with KNN implementation using Manhattan distance metric, at K=14, along with Decimal scale normalization.

Keywords: K-Nearest Neighbour, Normalization, Manhattan distance, Euclidean distance, Chebyshev distance, Cosine distance Machine learning and Decimal Scaling

1. Introduction

Breast cancer is the second most common type of cancer in women and one of the leading causes of cancer related deaths. As per statistics (Berry, 2017), the breast cancer occurs more in western countries when compared to developing countries. On the contrary, the death rates due to breast cancer, from the developing countries, disease are higher as compared to the death rates, due to the same, from developed countries. It also states that the rate of breast cancer per 100,000 women is higher in the U.S., Canada, and Europe. Breast cancer occurs when an infected tissue (tumour) begins to spread quickly. These cancerous cells can move anywhere within the body causing further damage.
There are two types of breast cancer tumours:

1. Non-cancerous or 'benign'
2. Cancerous or 'malignant'

Timely prediction requires a precise and authentic methodology to distinguish between benign breast tumors from malignant ones. Nowadays, diagnostic tests such as Surgical Biopsy have been replaced by Fine Needle Aspiration (FNA).

However, the accurate prediction of fatal disease outcome is still one of the most challenging tasks needed to be done till date and various Machine Learning techniques have become a popular tool to resolve this problem. Machine learning predictive analytics and pattern recognition have achieved 89% accuracy rate (Yun Liu, 2017). That's quite a bit ahead of an average score of 73% for the existing system of surgical biopsy. Also, using machine learning techniques, there is relative reduction in cost as less human effort is required. Furthermore, the fast speed at which machine learning consumes data allows the system to produce real-time data and predictions in a shorter duration of time.

The present paper measures the performance of K-Nearest Neighbour algorithm under various normalization techniques and distance metrics. Different distance metrics considered for the same include Manhattan distance, Euclidean distance, Chebyshev distance and Cosine distance and the different normalization techniques include Min-Max, Z-Score and

Decimal Scaling. Our aim is to check the efficiency of these approaches at different values of 'K'. The accuracy of each normalization technique is tested for every distance metric and the maximum accurate prediction is considered for the final output. The paper is organized as follows. In section 2, the related work has been explored. Context of the experiment is proposed in section 3. Whereas section 4, elaborates the selected inputs and the methods undertaken. In section 5, experimental outcomes are discussed. Finally, section 6 concludes the paper.

2. Previous works

Authors in (Seyyid Ahmed Medjahed, 2013), compared the accuracy of k-NN using several distances and different normalization techniques. Various distances included were Euclidean distance, City Block Distance, Cosine distance and Correlation distance. The k-parameter in the algorithm had a range from 1 to 50. Highest accuracy of 98.70% was obtained when the k-parameter was taken as 1 and Euclidean distance was chosen as the distance metric.

The accuracy of Naïve Bayes, SVM and Ensemble Algorithm were analyzed by Animesh et. al. in (Hazra, Mandal & Gupta, 2016). Using feature selection, it was found that Naïve Bayes gave maximum accuracy of 97.3978% by selecting only 5 dominant features.

In (Janghel, 2010), authors' implemented four models of neural networks namely Back Propagation Algorithm, Radial Basis Function Networks, Learning vector Quantization and Competitive Learning Network. In the best configuration, it was observed that Learning Vector Quantization, Competitive Learning and Multi-Layer Perceptron algorithms had testing accuracy of 95.82, 74.48 and 51.88% respectively.

In (Hiba, Hajar & Hassan, 2016), Hiba Asri, HajarMousannif, Hassan Al Moatassime and Thomas Noel have compared efficiencies among Support Vector Machine (SVM), Naïve Bayes (NB) and k Nearest Neighbours (k-NN) on the breast cancer data set with accuracy of 97.13, 95.99and 95.27 respectively.

Othman and Thomashave analyzed the breast cancer dataset using WEKA in (Bin & Yau, 2017), by applying Bayes Network, Radial Basis Function, Pruned Tree, Single Conjunctive Rule Learner and Nearest Neighbour Algorithm. The highest accuracy of 89.71% was achieved by Naïve Bayes algorithm followed by Radial Basis Function with an accuracy of 87.43%. while the Nearest neighbours algorithm achieved the accuracy of 84.57%.

In (Rana, 2015) Mandeep Rana et. al. compared the accuracy of Support Vector Machine, Logistic Regression, KNN and Naive Bayes. It was observed that the highest testing accuracy of 95.68% was achieved by k-NN with Euclidean distance.

Above discussed work had taken the Wisconsin Breast Cancer Dataset (Street, Wolberg, Mangasarian & Goldgof, 1993) for the reference.

3. Experiment environment, materials and methods

The various materials that have been used in the paper include: Python for coding purposes and Wisconsin (Diagnostic) Breast Cancer Database (Street,
Wolberg, Mangasarian & Goldgof, 1993) (WBCD) that has been taken from the University of Wisconsin. The technique used is K-Nearest Neighbour.

3.1 Experiment Environment

The experiments that have been discussed in this research paper have been done with the help of Python. Python contains a lot of libraries which help in classification, prediction, regression and various other machine learning techniques which help in simplifying the code and reduce the human labour put in towards the code while generating the most efficient and accurate results.

3.2 Breast Cancer Dataset

This paper uses the Wisconsin (Diagnostic) Breast Cancer Database (WBCD)(Street, Wolberg, Mangasarian & Goldgof, 1993) obtained from the University of Wisconsin. It comprises of 569 instances out of which 212 cases are malignant and 357 cases are benign. Features in this data set are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass which describes characteristic of the cell nuclei present in the image.

Ten real-valued feature (Hamou & Mohamed, 2018), are figured for every cell core:

1. Radius
This is calculated by taking the mean of all the distances from the center of the cell nuclei to the points on its perimeter.

2. Texture
The texture of a cell nucleus (Street, Wolberg, Mangasarian & Goldgof, 1993) is calculated by taking the standard deviation of the grayscale values.

3. Perimeter (the sum of the distances between consecutive boundary points)
The total perimeter is calculated by summing up the distances between consecutive boundary points of the cell nuclei.

4. Area
The nuclear area (Street, Wolberg, Mangasarian & Goldgof, 1993) is calculated simply by counting the number of pixels on the interior of the cell and adding one-half of the pixels in the perimeter.

5. Smoothness
The smoothness of a nuclear contour (Street, Wolberg, Mangasarian & Goldgof, 1993) is quantified by measuring the distance between the length of the radial line and the mean length of the lines surrounding it.

6. Compactness
Compactness is used to calculate how compact a cells is. It is calculated by squaring the perimeter of a cell per its unit area.

7. Concavity
It is the estimation of the severity of concave portions of the contour.

8. Concave points
It is the measure of number of concave portions of the contour.

9. Symmetry
It is evaluated by taking the relative contrast between pairs of lines perpendicular to the axis of the cell.

10. Fractal dimension
The fractal dimension is approximated using the "Coastline Approximation" described by Mandelbrot (Mandelbrot, 1977).

The mean value, worst case and standard error value are computed for each of the features, thus making it a total of 30 attributes. Apart from these, the other 2 attributes in the data set are id and diagnosis result.

3.3 K-Nearest Neighbour (K-NN)

In K-NN, an object is classified by the majority of its neighbours. The value of K determines the number of neighbours to be considered for the classification. Figure 1 depicts the flow chart of K-NN. In the test data, we select the distance metric and k value to be used. Here k value represents the number of neighbours to be considered for prediction. After based on distance selected, difference of test data from each training sample is computed. Then k minimum distances are selected, based on the majority of their category that class is assigned to the test data.

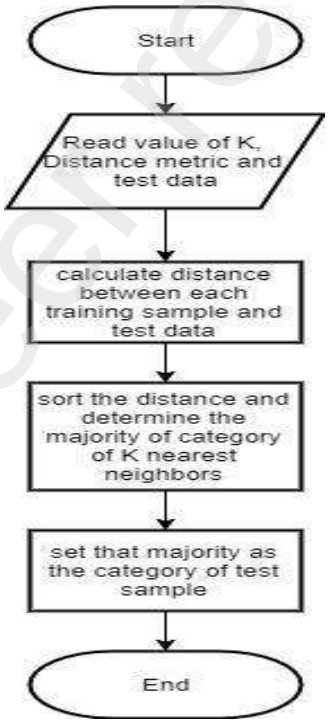

Fig. 1 - Flow chart of K Nearest Neighbours.

3.3.1 Normalization techniques used

Normalization is a scaling technique which is used to bring all the features of a dataset to an equal weight. Normalization can be termed as a pre-processing stage (Shalabi, Shaaban & Kasasbeh, 2006). It is usually done to make sure that all the features, especially the ones with the least values, do not get ignored during the computation process. The dataset considered for this paper contained attributes of varying scales. Therefore, we have used the following normalization techniques in order to scale the features to a particular range, so that each feature contributes effectively in deciding the algorithm's output.

3.3.1.1 Min-Max

The normalized value of a member of the set of observed values of *x* (Dodge, 2003) is given by:

$$Z = \frac{x - min(x)}{max(x) - min(x)} \quad (1)$$

where min and max are the minimum and maximum values of x in given range.

3.3.1.2 Z-Score

Z-score is a number of standard deviations a data point is away from the population mean. Mathematically, Z-score can be defined as

$$Z = \frac{x - \mu}{\sigma} \quad (2)(\text{Katayama \& Satoh, 2001})$$

3.3.1.3 Decimal Scaling

It is another normalization technique, in this; we move the decimal point of values of the attribute. This movement of decimal points totally depend on the maximum value among all values in the attribute (Dodge, 2003).

$$v' = \frac{v}{10^j} \quad (3)$$

Where,
v=observed value,
j=number of digits in maximum value and
v'=normalized value

These normalization techniques have been used to represent the data within the following range, as shown in Table 1.

Table 1 - Feature scaling range

Normalization Technique	Range scaled
Min-Max	0 to +1
Z-score	-2 to +2
Decimal scaling	-1 to +1

3.3.2 Distance functions used

The significance of picking the correct distance function winds up basic when we are managing high-dimensional information. The quality of how well a certain distance function fits a given problem might be firmly influenced by sparseness of data (Katayama & Satoh, 2001). Thus, we can say that selecting a 'better' distance function would play a crucial role in improving the performance of algorithms, particularly, distance-based algorithms such as KNN. For a KNN classifier, just the distances between each point and its nearby neighbourhood require be 'good' for it to perform well. Thus, undertaking the appropriate distance function used for choosing the closest neighbours can lead to a better classification of identifier to which family a particular point belongs, which can further improve the classifier's performance.

The above mentioned normalization techniques have been implemented for four distance functions. The chosen distance functions, where p = (p_1, p_2,...,p_n) and q = (q_1, q_2,..., q_n) are two points from i=1 to n, are as follows

3.3.2.1 Euclidean distance

The Euclidean distance (Anton, 1994), between points p and q is the length of the line connecting them. It is the "ordinary" straight-line distance between two points in Euclidean space (Deza & Deza, 2009).

$$D = \sqrt{\sum (q_i - p_i)^2} \quad (4)$$

3.3.2.2 Manhattan distance

Manhattan distance (Black, 2006) is the distance between two points measured along axes at right angles. In other words, it is the total of the lengths of the projections of the line sections between the points onto the axes of the coordinate system (Eugene, 1987). This distance, originally proposed by Minkowsky is defined as follows.

$$D = \sum |q_i - p_i| \quad (5)$$

3.3.2.3 Chebyshev distance

It is a metric defined on a vector space where the distance between two vectors is the greatest of their differences along any coordinate dimension (Cantrel, 2000) (Abello, Pardalos & Resende, 2002).

$$D = max_i(p_i - q_i) \quad (6)$$

3.3.2.4 Cosine distance

Cosine similarity (Singhal, 2001) is the measure of the cosine of the angle between the two non-zero vectors. Cosine distance is used to represent the complement in positive space, that is, D (d1,d2)=1-S(d1,d2), where D is the cosine distance and S is the cosine similarity(Singhal, 2001). It can be formulated as follows.

$$D(d1,d2) = \frac{(d1.d2)}{||d1|| \times ||d2||} \quad (7)$$

Where d1 and d2 are vectors,

(d1.d2) = Dot product =$d1[0] \times d2[0] \times d1[1] \times d2[1],\ldots d1[n] \times d2[n]$,

$||d1|| = \sqrt{(d1[0]^2 + d1[1]^2 + \ldots d1[n]^2)}$ and

$||d2|| = \sqrt{(d2[0]^2 + d2[1]^2 + \ldots d2[n]^2)}$

3.4 Block Diagram

The data has been acquired from WB CD. After acquisition, pre-processing of the data has been performed through normalization. Three types of

normalization techniques have been used which include Min-Max, Z-Score and Decimal Scaling as shown Figure 2. After normalization of the dataset, distance between the test sample and training instances are calculated through different distance metrics which are Euclidean, Manhattan, Chebyshev and Cosine. The results obtained through various combinations of normalization techniques and distances are then compared to find out the combination which gives the most accurate result.

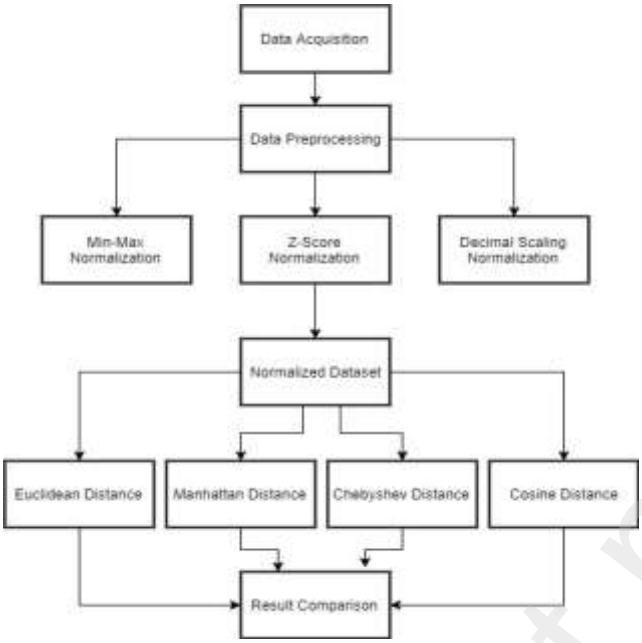

Fig. 2 - Block diagram of workflow

4. Experimental results and discussions

In this segment, the outcomes obtained after performing the experiments are reported.

We used 3 different types of normalization techniques, which are, Min-Max, Z-Score and Decimal Scaling along with 4 different types of distances, which are, Euclidean, Manhattan, Chebyshev and Cosine. KNN is implemented using each of these 12 combinations. The accuracy percentage is noted for different values of K. For a particular method, maximum accuracy obtained at particular K is considered for the final output.

The following graphs depict the output KNN at various K values using Min-Max Normalization (Dodge, 2003)), Z-Score Normalization and Decimal Scaling Normalization (Dodge, 2003), respectively.

4.1 KNN using Min-Max Normalization technique

Figure 3 shows various accuracy percentages obtained when KNN algorithm was implemented with different types of distances along with Min-Max normalization technique. When Euclidean distance was chosen as the distance metric, the highest accuracy percentage was obtained at K=8 and the accuracy equalled to 95.61%. Similarly, when Manhattan distance was chosen as the distance metric, the highest accuracy percentage was obtained at K=6 and the accuracy equalled to 97.36%. When Chebyshev distance was chosen as the distance metric, the highest accuracy percentage was obtained at K=6 and the accuracy equalled to 92.10%. Finally, when Cosine distance was chosen as the distance metric, the highest accuracy percentage was obtained at K=14 and the accuracy equalled to 86.84%. It is clear from the above graph that the maximum accuracy of 97.36% was achieved at K=6 when Manhattan distance was chosen as the distance metric.

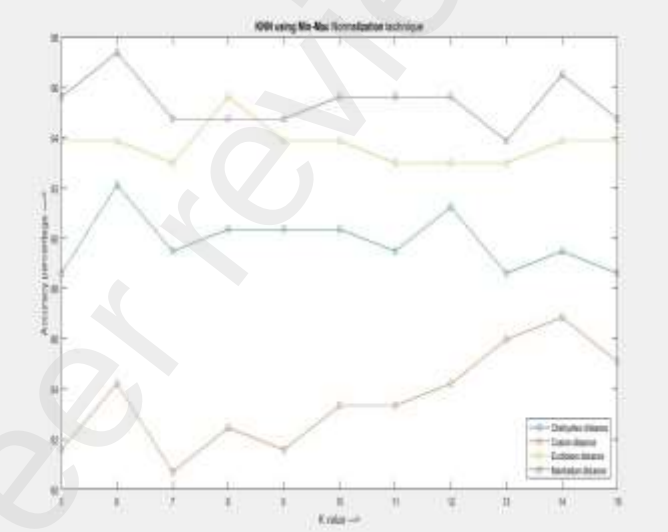

Fig. 3 - KNN using Min-Max normalization

4.2 KNN using Z-Score Normalization technique

Figure 4 shows various accuracy percentages obtained when KNN algorithm was implemented with different types of distances along with Z-Score normalization technique. When Euclidean distance was chosen as the distance metric, the highest accuracy percentage was obtained at K=4 and the accuracy equalled to 95.61%. Similarly, when Manhattan distance was chosen as the distance metric, the highest accuracy percentage was obtained at K=6 and the accuracy equalled to 95.61%. When Chebyshev distance was chosen as the distance metric, the highest accuracy percentage was obtained at K=4 and the accuracy equalled to 94.73%. Finally, when Cosine distance was chosen as the distance metric, the highest accuracy percentage was obtained at K=4 and the accuracy equalled to 94.73%. It is clear from the above graph that the maximum accuracy of 95.61% was achieved for both Euclidean and Manhattan distance at K=4 and K=6 respectively.

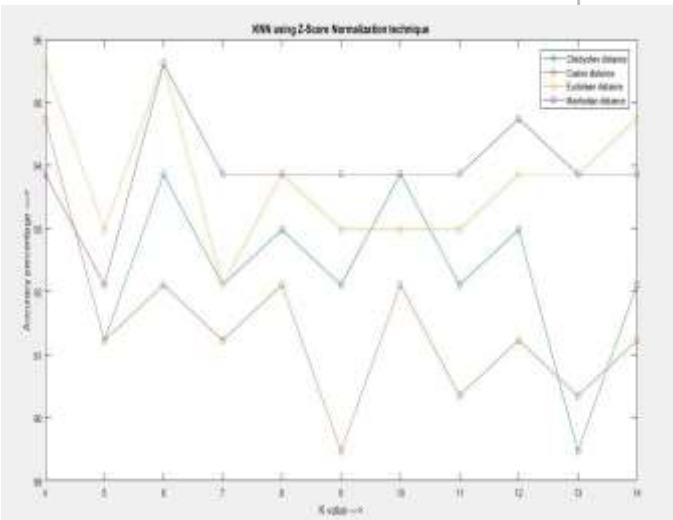

Fig. 4 - KNN using Z-Score normalization

4.3 KNN using Decimal Scaling technique

Figure 5 shows various accuracy percentages obtained when KNN algorithm was implemented with different types of distances along with Decimal Scaling normalization technique. When Euclidean distance was chosen as the distance metric, the highest accuracy percentage was obtained at K=4 and the accuracy equalled to 94.73%. Similarly, when Manhattan distance was chosen as the distance metric, the highest accuracy percentage was obtained at K=14 and the accuracy equalled to 98.24%. When Chebyshev distance was chosen as the distance metric, the highest accuracy percentage was obtained at K=4 and the accuracy equalled to 48.24%. Finally, when Cosine distance was chosen as the distance metric, the highest accuracy percentage was obtained at K=10 and the accuracy equalled to 68.42%. It is clear from the above graph that the maximum accuracy of 98.24% was achieved at K=14 when Manhattan distance was chosen as the distance metric.

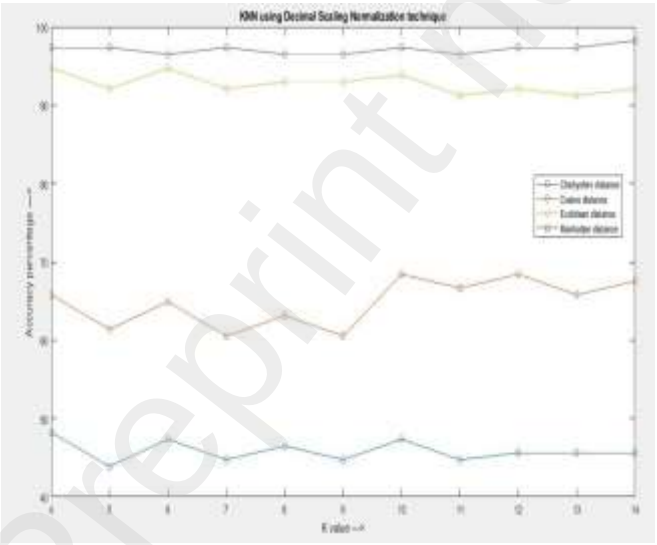

Fig. 5 - KNN using Decimal Scaling normalization

4.4 Collective summarization of obtained results

Based on the above results, a table can be constructed as follows.

Table 2 - KNN accuracy using various normalization techniques and distance functions

Distance Function Normalization Technique	Euclidean distance	Manhattan distance	Chebyshev distance	Cosine distance
Min-Max	95.6140, K=8	97.3684, K=6	92.1053, K=6	86.8421, K=14
Z-Score	95.6140, K=4	95.6140, K=6	94.7368, K=4	94.7368, K=4
Decimal Scaling	94.7368, K=4	98.2456, K=14	48.2456, K=4	68.4211, K=10

Table 2 shows the maximum accuracy percentage using KNN for each normalization technique using the three distance functions.

The maximum accuracy obtained at a particular K for each normalization is depicted graphically as

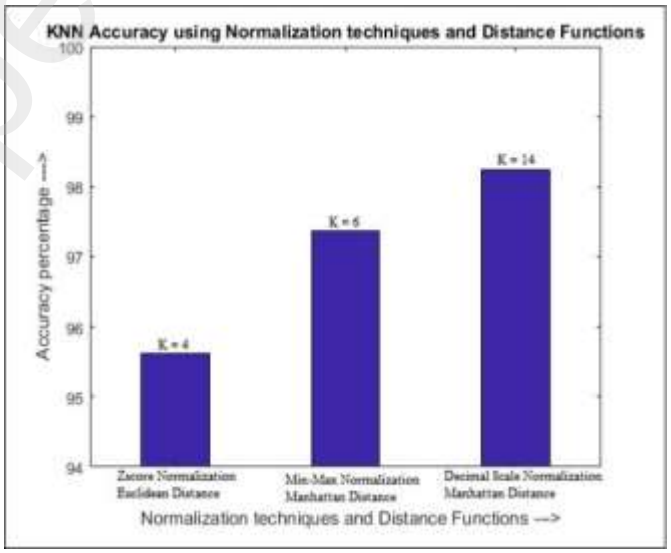

Fig. 6 - Maximum KNN accuracy using various normalization techniques

As shown in Figure 6 the highest accuracy of K-Nearest Neighbour algorithm, when applied on the dataset, was obtained at K = 14. The normalization technique used was Decimal Scaling and the distance metric used was Manhattan Distance.

5. Conclusion

K-Nearest Neighbour (KNN) algorithm has been implemented along with various normalization techniques and distance metrics. A significant improvement in the accuracy was noticed when different combinations of normalization techniques and distance were tested out against the dataset.

The best accuracy achieved was 98.24 % when Decimal Scaling Normalization technique was used along with Manhattan distance at K = 14.

REFERENCES

Berry, J. (2017, April 26). "Worldwide statistics on breast cancer: Diagnosis and risk factors." *Medical News Today*. Retrieved from https://www.medicalnewstoday.com/articles/317135.php.

Yun Liu *et al., Detecting Cancer Metastases on Gigapixel Pathology Images",* 2017. Available:arXiv:1703.02442

Seyyid Ahmed Medjahed *et al.,"Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules",* in International Journal of Computer Applications (0975 - 8887), Vol. 62-No.1,January 2013.

Hazra Animesh, Mandal K. Subrata and Gupta Amit, "Study and Analysis of Breast Cancer Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms." *International Journal of Computer Applications* 145(2):39-45, July 2016.

R.R.Janghel *et al*.,"Breast cancer diagnosis using Artificial Neural Network models"**,** IEEE **,** August 2010.

Hiba Asri, Hajar Mousannif, Hassan Al Moatassime, Thomas Noel, "*Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis*", Elsevier, 2016, DOI 10.1016/j.procs.2016.04.224,

Bin Othman M.F., Yau T.M.S. (2007),"*Comparison of Different Classification Techniques Using WEKA for Breast Cancer*." In: Ibrahim F., Osman N.A.A., Usman J., Kadri N.A. (eds) 3rd Kuala Lumpur International Conference on Biomedical Engineering 2006. IFMBE Proceedings, vol 15. Springer, Berlin, Heidelberg.

Rana, Mandeep, "Breast Cancer Diagnosis And Recurrence Prediction Using Machine Learning Techniques." *International Journal of Research in Engineering and Technology*, 2015.

Hamou, Reda Mohamed. "Handbook of Research on Biomimicry in Information Retrieval and Knowledge Management." IGI Global, 2018. 1-429. Web. 11 Mar. 2018.

Mandelbrot, B.B. (1977), The Fractal Geometry of Nature, W.H. Freeman and Company, New York.

Dodge Y, "*The Oxford Dictionary of Statistical Terms"*,(2003)

N. Katayama and S. Satoh, "Distinctiveness-sensitive nearest-neighbor search for efficient similarity retrieval of multimedia information,"*Proceedings 17th International Conference on Data Engineering*, Heidelberg, 2001, pp. 493-502.

Anton, Howard (1994), "Elementary Linear Algebra" (7th ed.), John Wiley & Sons, pp. 170–171, ISBN 978-0-471-58742-2

Deza, Elena; Deza, Michel Marie, "*Encyclopedia of Distances",*(2009) Springer. p. 94.

Paul E. Black, "Manhattan distance", in Dictionary of Algorithms and Data Structures [online], Vreda Pieterse and Paul E. Black, eds. 31 May 2006.

Eugene F. Krause, "*Taxicab Geometry",*(1987) Dover.

Cyrus. D. Cantrel, "*Modern Mathematical Methods for Physicists and Engineers"* (2000), Cambridge University Press.

James M. Abello, Panos M. Pardalos, and Mauricio G. C. Resende (editors) (2002). *Handbook of Massive Data Sets*. Springer.

Singhal, Amit, "*Modern Information Retrieval: A Brief Overview*", (2001) Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24 (4): p. 35–43.

Shalabi, L.A., Z. Shaaban and B. Kasasbeh, Data Mining: A Preprocessing Engine, J. Comput. Sci., 2: 735-739, 2006

Street, W. N., W. H. Wolberg, O. L. Mangasarian, and Dmitry B. Goldgof , "Biomedical Image Processing and Biomedical Visualization", 1993.