International Journal of Theoretical & Applied Sciences, 13(2): 01-06(2021)

ISSN No. (Print): 0975-1718
ISSN No. (Online): 2249-3247

A Detailed Analysis of Applying the K Nearest Neighbour Algorithm for Detection of Breast Cancer

Somya Singh, Aditi Sneh and Vandana Bhattacharjee*

Birla Institute of Technology, Mesra Ranchi, (Jharkhand), India.

(Corresponding author: Vandana Bhattacharjee)*
(Received 26 May, 2021, accepted 20 July, 2021)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: Breast cancer is a serious disease plaguing women all over the world. However, if detected early, it can be treated and one can live a healthy life. Mammography images are used by radiologists to diagnose the presence or absence of breast cancer. The use of machine learning techniques in health care is gaining popularity, and this paper attempts to present a detailed analysis of the use of K Nearest Neighbour (KNN) classifier to diagnose breast cancer in particular. The application of any machine learning algorithm presents some challenges such as handling missing data, removing erroneous data and sometimes even data generation. The main contribution of this work is in predicting the presence of Cancer on the Breast Cancer Data Set (BCD) taken from the UCI Machine Learning Repository. The proposed work has achieved a best accuracy of 96.51% on the test dataset.

Keywords: K Nearest Neighbour classifier, health care, accuracy, feature selection

I. INTRODUCTION

Breast Cancer occurs when the breast cells begin to grow in an amount that is not normal. The cells start growing more than as compared to a healthy cell and start accumulating forming a lump. This further leads to blood discharge. Breast Cancer is also the second leading cause of cancer death in women after lung cancer. One of the major steps in the cure is to detect it early and machine learning algorithms have proved to be helpful in the same. Breast Cancer diagnosis is one of the most serious problems in the medical field. Various researches have been carried out by researchers in order to improve the performance and obtain satisfactory outcomes. Machine learning is particularly used in Breast cancer diagnosis. This paper presents a detailed analysis of the application of KNN classifier for cancer classification. We have not only conducted extensive experiments for finding the best set of features, but also for finding the best value of K. The figures in experiment section showing the Elbow curves present a clear idea to the reader about our experiments, and would provide further guidance to the prospective researchers. This depth of analysis on the KNN algorithm has not been found to the best of our knowledge. A comparison of the previous work done has been presented in the results section. The rest of the paper is organized as follows. Section II presents the related work. Section III presents the methods and Section IV presents the experiments and results. Finally Section V concludes the paper.

II. LITERATURE REVIEW

Bakthavachalam and Raj (2020) used the University of Wisconsin's WBCD (Wisconsin breast cancer database). They used the k-Nearest Neighbor method to investigate various distance measures, multiple K values, and different classification rules, resulting in a 98.70 percent accuracy [1]. Seyyid *et al*., (2013) used KNN with different types of distances and classification rules in function of the parameter K [2]. They experimented on Wisconsin Breast Cancer Data (WBCD) obtained by UCI machine repository. They reached 98.70% accuracy for Euclidean distance and 98.48% accuracy for Manhattan distance. Joshi and Mehta (2018) used the KNN technique on Wisconsin (Diagnostic) Breast Cancer dataset showing that KNN technique with Linear Discriminant Analysis technique (Dimensionality Reduction Technique gives an accuracy of 97.06% [3].

Priya *et al.,* (2018) reached an average of 96% using KNN algorithm [4]. Victoria *et al.,* (2018) leveraged the KNN algorithm using six distance measurements against a mammographic mass dataset to determine the predictive capability of malignancy [5]. The best overall results were found by using the Manhattan distance measurement and a K value of 7, yielding accuracy of 81.67%. Shagun *et al.,* (2018) achieved the highest accuracy of 98.24% with KNN implementation using the "Manhattan" distance metric [6]. Rana *et al.,* (2015) used classification algorithms, including Support Vector Machines, Logistic Regression, KNN, and Naive Bayes.

They examined the accuracy of each strategy and discovered that SVM was better for predictive analysis and KNN was best for our overall methodology [7]. Assegie (2021) analysed the performance of the model on breast cancer detection using the testing set, revealed that the accuracy of the proposed optimized model is 94.35% and the performance of the KNN with the default hyper-parameter is 90.10% [8]. Eyupoglu

(2018) implemented KNN for different k values for 2-fold, 5-fold and 10-fold cross validation and different k values and the obtained classification accuracies. The classification accuracy of nearly 97% was achieved for k = 5 [9].

Ben (2014) experimented with Wisconsin Diagnostic Breast Cancer data reached 96% of right prediction using K-nearest neighbors (KNN) technique [10]. Aidarus *et al*., (2016) experimented on the Digital Database for Screening Mammography (DDSM). They obtained the classification results using KNN and SVM (Support Vector Machine) for k means-max pooling and Bag-of-features. KNN technique performed better than SVM, with a high accuracy of 98.19% [11]. Palaniammal and Chandrasekaran (2014) experimented on Wisconsin breast cancer data (200 rows and 11 columns) and applied a fourfold Cross Validation method for testing in which each fold contains 50 instances. They reached 97% accuracy using KNN technique [12]. Jha *et al* (2021) have predicted the presence of thyroid disease among patients using machine learning techniques and feature selection algorithms [13]. Sivalenka and Bai (2021) performed comparison of Decision Tree, Support Vector Machine, K Nearest Neighbour and Radius Nearest Neighbour Classifier yielding accuracies of 91.23%, 94.7%, 95.61% and 98.9% respectively [15]. In recent research works, Husain *et al.*, (2021) have applied soft computing approach to solve a problem related to mentally stressed students [16], whereas Sodhar *et al.*, (2021) has built an E health application for kids [17]. Other researchers have applied machine learning techniques for detecting cervical cancer, autism detection and also Covid 19 disease prediction [18-20].

III. METHODS

A. K-Nearest Neighbour Algorithm

The k-nearest neighbours (KNN) algorithm uses the idea of calculating distance between two points on a graph. Selecting the right K is our task and for that we run our algorithm multiple times to reduce the number of errors such that a more accurate result is obtained.

The K-NN working can be explained on the basis of the below algorithm:

Algorithm: KNN
Input: *k*, Dataset D, test data *t*,
Output: class label *c*
Start
1. For each $d \in D$,
(a) compute $dist(t, d)$ and store in sorted list *l* // sorted in ascending order
2. find top *k* closest neighbors of *t*
3. find the label *c* by majority voting of the *k* neighbors
4. return *c*// the label assigned to *t*
End

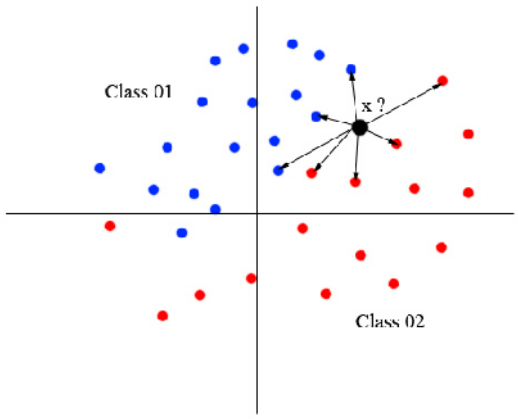

Fig. 1. Graphic visualization of KNN algorithm.

B. Evaluation Parameters

Confusion matrix

Predicted actual	True	False
True	True positive	False negative
false	False positive	True negative

The evaluation parameters we have used are as follows: Precision measures the number of positive class predictions that belong to the positive class.

$$\text{Precision (P)} = \frac{\text{True Positive}}{\text{TruePositive+False Positive}}$$

Recall measures the number of positive class predictions made out of all positive examples in the dataset.

$$\text{Recall (R)} = \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

F-Measure gives a value that balances both the concerns of precision and recall in one number.

$$\text{F-Measure (FM)} = 2 \times \frac{\text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

IV. EXPERIMENTS AND RESULTS

The University of Wisconsin's WBCD (Wisconsin breast cancer database) [14] was used in this study. Breast cancer data from human breast tissue is included in the database. There are a total of 699 instances, 458 of which are benign tumours (class 0) and 241 of which are malignant tumours (labeled as class 1). The K-Nearest Neighbor algorithm has been applied with the different K values. The dataset is partitioned into Training Set with 468 data values and Testing Set with 231 data points.

Table 1: Description of Datasets.

Class	Wisconsin breast cancer database
0 (benign tumours)	458
1 (malignant tumours)	241

Table 2: The Confusion Matrix values for various feature settings.

Features	Name of Features	Best K	Class	Confusion	Matrix
1-4	clump_thickness, Uniformity_Cell_size, Uniformity_cell_shape, Marginal_adhesion.	15	0	145	6
			1	6	76
2-5	Uniformity_Cell_size, Uniformity_cell_shape, Marginal_adhesion, Single_e_cell_size	5	0	143	6
			1	4	78
2-6	Uniformity_Cell_size, Uniformity_cell_shape, Marginal_adhesion, Single_e_cell_size, Bare_nuclei	5	0	143	6
			1	3	79
3-6	Uniformity_cell_shape, Marginal_adhesion, Single_e_cell_size, Bare_nuclei	15	0	145	3
			1	5	79
8-10	Normal_nucleloi, Mitoses, Class	5	0	139	10
			1	11	71

Table 3: Performance analysis for different feature sets for Class-1 (cancerous).

Features	Name of Features	Best K	Accuracy	Class	F1	Precision	Recall	Support
1-4	clump_thickness, Uniformity_Cell_size, Uniformity_cell_shape, Marginal_adhesion.	15	0.9567	1	0.93	0.93	0.93	82
2-5	Uniformity_Cell_size, Uniformity_cell_shape, Marginal_adhesion, Single_e_cell_size	5	0.9567	1	0.94	0.93	0.95	82
2-6	Uniformity_Cell_size, Uniformity_cell_shape, Marginal_adhesion, Single_e_cell_size, Bare_nuclei	5	0.9610	1	0.95	0.93	0.96	82
3-6	Uniformity_cell_shape, Marginal_adhesion, Single_e_cell_size, Bare_nuclei	15	0.9653	1	0.95	0.94	0.96	82
8-10	Normal_nucleloi, Mitoses, Class	5	0.9090	1	0.87	0.88	0.87	82

Table 4: Performance analysis for different feature sets for Class-0 (Non cancerous).

Features	Name of Features	Best K	Accuracy	Class	F1	Precision	Recall	Support
1-4	clump_thickness, Uniformity_Cell_size, Uniformity_cell_shape, Marginal_adhesion.	15	**0.9567**	0	0.96	0.96	0.96	149
2-5	Uniformity_Cell_size, Uniformity_cell_shape, Marginal_adhesion, Single_e_cell_size	5	**0.9567**	0	0.97	0.97	0.96	149
2-6	Uniformity_Cell_size, Uniformity_cell_shape, Marginal_adhesion, Single_e_cell_size, Bare_nuclei	5	**0.9610**	0	0.97	0.98	0.96	149
3-6	Uniformity_cell_shape, Marginal_adhesion, Single_e_cell_size, Bare_nuclei	15	**0.9653**	0	0.97	0.98	0.97	149
8-10	Normal_nucleloi, Mitoses, Class	5	**0.9090**	0	0.93	0.93	0.93	149

Bar-Graph Representation for the various results are presented in Fig. 2 (a) and (b).

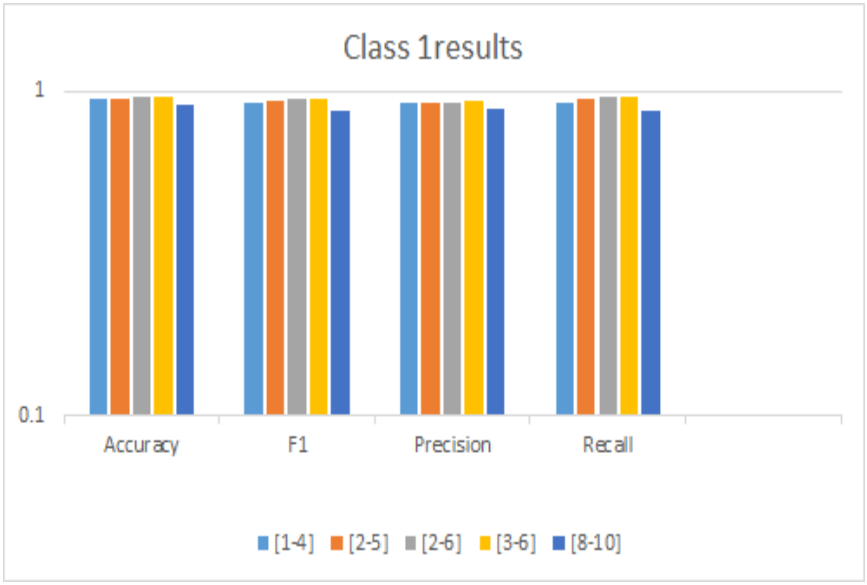

Fig. 2 (a) Bar graph for Class 1.

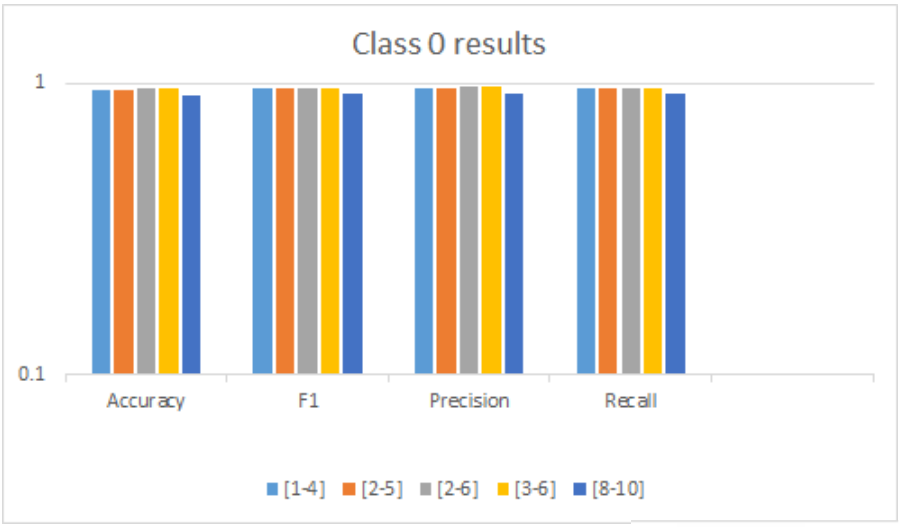

Fig. 2 (b) Bar graph for Class 0.

Whereas, accuracy is the total number of correct predictions divided by the total number of data points used for prediction, *N*.

$$\text{Accuracy (A)} = \frac{\text{True Positive} + \text{True Negative}}{\text{N}}$$

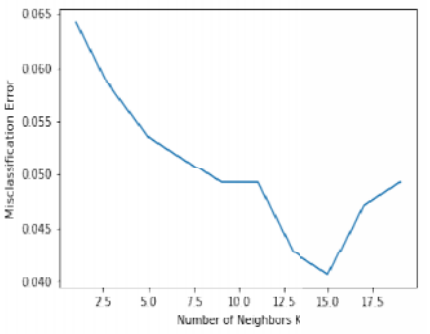

Fig. 3 (a) Elbow curve for feature columns 1-4.

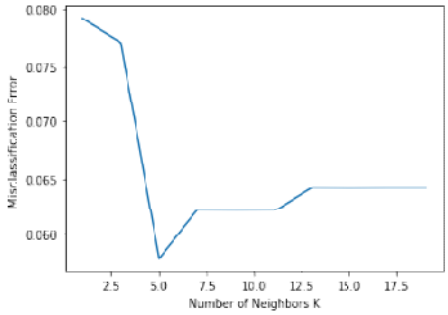

Fig. 3 (b) Elbow curve for feature columns 2-5.

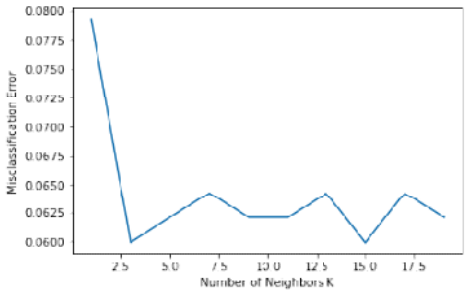

Fig. 3 (c) Elbow curve for feature columns 3 – 6.

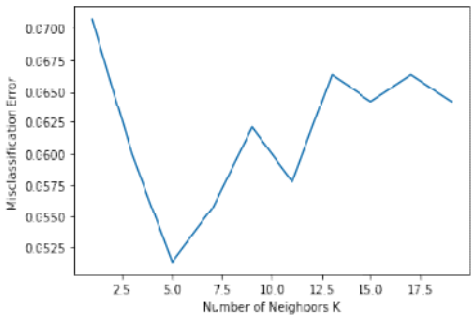

Fig. 3 (d) Elbow curve for feature columns 2 – 6.

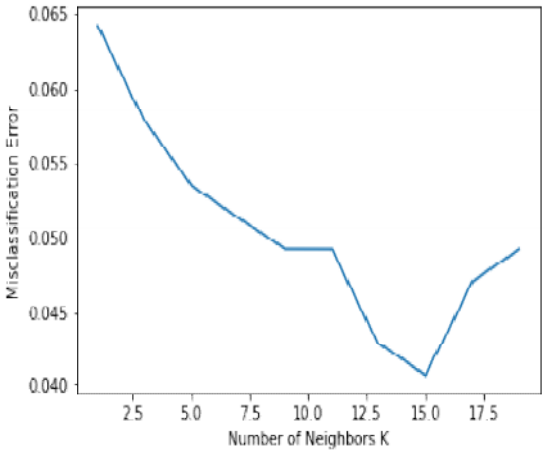

Fig. 3 (e) Elbow curve for feature columns 8 – 10.

The elbow curve graphs for various settings are given in Figs. 3.a – 3.e. Based on the lowest misclassification error, the best values of K are obtained. The corresponding accuracy values have been reported in Table 3 and 4. The comparative analysis with other researchers has been presented in Table 5, and it shows that our results are comparable or better than others.

Table 5: Comparative analysis with other research work.

Classifier	Accuracy(%)
K-Nearest Neighbours [4]	96.0
K-Nearest Neighbours [5]	81.67
K-Nearest Neighbours [8]	90.10
K-Nearest Neighbours [9]	96.99
K-Nearest neighbors [10]	96.0
K-Nearest Neighbour [15]	95.51
K-Nearest Neighbours [proposed]	96.53

V. CONCLUSION AND FUTURE SCOPE

This paper proposes a classification approach to predict Breast Cancer. The study was conducted on the WBC datasets. We have conducted extensive experiments for finding the best set of features, and also for finding the best value of K. The figures in experiment section showing the Elbow curves present a clear idea to the reader about our experiments, and would provide further guidance to the prospective researchers. This depth of analysis on the KNN algorithm has not been found to the best of our knowledge. A comparison of the previous work done has been presented in the results section. From our experimental results presented in Table 3, it is observed that KNN yields the highest accuracy of 96.53%. The comparative analysis with other researchers has been presented in Table 5, and it shows that our results are comparable or better than others. As part of our ongoing work, we aim to apply dimension reduction techniques in our classification model. Data generation and augmentation techniques shall also be used for improving the classification accuracy further. Further, deep learning techniques can be applied for disease prediction using image datasets.

Acknowledgement. The authors thank the reviewers for their comments which greatly helped in preparing the paper to its present form.

Conflict of Interest. The authors declare that they have no conflict of interest.

REFERENCES

[1]. Bakthavachalam, M. D., and Raj, D. S. (2020). 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*(3), 4842-4851.

[2]. Seyyid, A., Tamazouzt, S., and Abdelkader, B. (2013). Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules. *International Journal of Computer Applications*, *62*(1).

[3]. Joshi, A., and Mehta, A. (2018). Analysis of K-nearest neighbour technique for breast cancer disease Classification. *International Journal of Recent Scientific Research*, *9*, 4(I), 26126-26130.

[4]. Priya, S., Agilandeeswari, L., and Prabukumar, M. (2017). "A Case Study on effective approach to predict class

of breast cancer on numerical data. *International Journal of Pure and Applied Mathematics, 117*(17), 161-167.
[5]. Victoria, R., Karan, S., and Dana, W. (2018). Breast Cancer Prediction with K-Nearest Neighbor Algorithm using different Distance Measurements. Software Engineering Project (SWEN 670), University of Maryland University College.
[6]. Shagun, C., Rajat, K., Ekansh, A., Sarthak, S. (2018). Breast Cancer Detection Using K-Nearest Neighbour Algorithm. *International Journal of Computational Intelligence & IoT, 2*(4).
[7]. Rana, M., Chandorkar, P., Dsouza, A., & Kazi, N. (2015). Breast cancer diagnosis and recurrence prediction using machine learning techniques. *IJRET: International Journal of Research in Engineering and* Technology, 372-376.
[8]. Assegie, T. A. (2021). An optimized K-Nearest Neighbor based breast cancer detection. *Journal of Robotics and Control (JRC), 2*(3), 115-118.
[9]. Eyupoglu, C. (2018). Breast cancer classification using k-nearest neighbors algorithm. *The Online Journal of Science and Technology*, *8*(3), 29-34.
[10]. Ben, F. (2014). k-Nearest Neighbors to Diagnose Cancer", benfradet.github.io
[11]. Aidarus, I. M., Baharum, B., Abas, S.M., and Hashimah N.P. (2016). Classification of Breast Tumor in Mammogram Image Using Unsupervised Feature Learning. *American Journal of Applied Sciences, 13*(5), 552-561
[12]. Palaniammal, V. and Chandrasekaran, R.M. (2014). Analysis for breast cancer diagnosis using KNN classification. *International Journal of Applied Engineering Research, 9*(22): 14233-14241
[13].Jha, R., Bhattacharjee, V., & Mustafi, A. (2021). Increasing the Prediction Accuracy for Thyroid Disease: A Step Towards Better Health for Society. *Wireless Personal Communications*, 1-18 https://doi.org/10.1007/s11277-021-08974-3.
[14]. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]
[15]. Sivalenka, V., & Bai, A. (2021). An analysis on prediction of breast cancer using radius Nearest Neighbor algorithm over other classification algorithms. *Materials Today: Proceedings*.
[16]. Husain, I. and Ali, A. (2021). Fuzzy Matrix Approach to Study the Maximum Age Group of Stressed Students Studying in Higher Education. *International Journal on Emerging Technologies, 12*(1), 31-35.
[17]. Sodhar, I. N., Buller, A. H., Memon D. N. I., Sodhar, A. N., and Mirani, A. A. (2021). E-Health Application for Kids. *International Journal on Emerging Technologies, 2*(1), 36-38.
[18]. Geeitha, S., and Thangamani, M., A. (2020). Cognizant Study of Machine Learning in Predicting Cervical Cancer at various Levels-A Data Mining concept. *International Journal on Emerging Technologies, 11*(1), 23-28.
[19]. Sudha, V. P., and Vijaya, M. S. (2020). Gated Recurrent Neural Network for Autism Spectrum Disorder Gene Prediction. *International Journal on Emerging Technologies, 11*(1), 136-141.
[20]. Asif, M., Ibrar M., Ahmad S., Farooq M. A., Ullah, H., Abbasi M. K. and Afzal, Z. (2021). Detection of COVID-19 from C-X-Ray Scans Empowered by Machine Learning. *International Journal on Emerging Technologies, 12*(2), 104-109.