ANALYSIS OF BREAST CANCER DATA USING *k*NN ALGORITHM

Abstract

Comparing the performance of machine learning algorithms over the same dataset is the current research trend. The fact that each algorithm has its underlying assumptions indicate that not all algorithms should be compared with each other on the same dataset. Hence, every algorithm should be compared only when the dataset satisfies the underlying assumptions. Algorithms that are suitable for a certain dataset should only be compared with each other at the optimal level. To pave a way for this, this study investigates the performance of several variations of the k-Nearest Neighbors (kNN) algorithm on a dataset comprising 569 breast cancer cases from the United States. The research evaluates the impact of three distance metrics, namely; Chebyshev, Manhattan, and Euclidean, across various values of k. The analysis reveals that the optimal metric is Euclidean distance metric and the optimal k values are k = 21 and 23. The optimal results obtained is 97.37%accuracy, 97.26% (Benign) and 97.56% (Malignant)precision, 98.61% (Benign) and 95.24% (Malignant)recall, and 97.93% (Benign) and 96.39% (Malignant)F1-scores.Additionally, the two optimal models (for k = 21 and k = 23) exhibit strong agreement on the features importance except compactness feature. Further analysis is recommended to better understand the role of compactness in breast cancer diagnosis.

Keywords

Machine Learning;*k* Nearest neighbor;Breast Cancer; Hyperparameter tuning; Distance Metrics

1 INTRODUCTION

As individuals age, a natural process occurs wherein cells die and necessitate replacement. Under normal circumstances, healthy cells undergo division to replace those that are dead or dying, thereby maintaining tissue homeostasis. However, cancer emerges when a normal cell experiences unregulated growth and division instead of undergoing apoptosis (Weinberg, 1996). The nomenclature of a specific cancer typically reflects the tissue or organ from which it originates; for example, aberrant cells in the breast are responsible for breast cancer. In addition to environmental factors, hereditary genetic mutations and DNA damage that can be transmitted to offspring further contribute to the etiology of cancer. Breast cancer, in particular, is characterized by the onset of abnormal cellular proliferation within the breast tissue, frequently culminating in the formation of tumors. According to Giaquinto et al. (2022), approximately 13% of women are expected to receive a breast cancer diagnosis during their lifetime. While palpable lumps are often the first sign of breast cancer, other manifestations may include the development of denser lumps, alterations in breast or

nipple size and shape, changes in the color of the breast surface, and unusual fluid discharge not related to lactation. Breast tumors can be classified into two primary categories: malignant (cancerous) or benign (non-cancerous). The process of diagnosis and subsequent treatment typically commences upon the confirmation of breast cancer through imaging modalities such as breast ultrasound, mammography, and Fine Needle Aspiration (FNA) biopsies (Lim et al., 2022). These diagnostic procedures may entail the administration of ionizing radiation exposure or pharmacological interventions. Recently, attention has been drawn to machine learning as an alternative to multiple tests and scans.

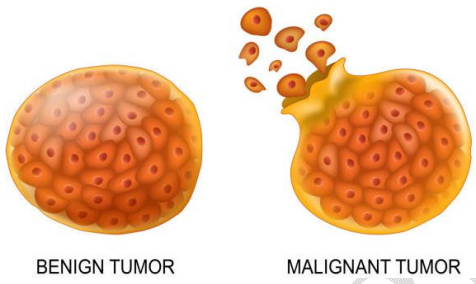

Figure 1: Benign and Malignant cells

The development of algorithms for machine learning (ML) is essential for enabling computers to learn from data and draw logical conclusions without the need for explicit programming instructions. In recent years, machine learning has risen to prominence within the field of artificial intelligence (AI), attracting considerable interest and investment from cancer researchers seeking to enhance diagnostic and treatment approaches (Chan et al., 2022; Singh et al., 2023). Fundamentally, machine learning involves utilizing advanced computer algorithms to identify hidden patterns within large datasets. These algorithms are trained to analyze input data, recognize pertinent patterns, and subsequently construct predictive models that can be applied to make informed decisions based on new, unseen data (Beam and Kohane, 2018). This iterative learning process allows ML systems to continuously refine their accuracy and effectiveness over time. A notable application of machine learning in oncology, particularly in breast cancer treatment, is the early diagnosis of the disease. By leveraging advanced imaging datasuch as mammograms, ultrasound images, and MRI scans, ML models can detect subtle abnormalities that may indicate the presence of cancer at its earliest stages. This innovative approach not only improves diagnostic accuracy but also has the potential to significantly enhance patient outcomes by allowing for timely interventions, as highlighted by research from Balkenende et al. (2022). Through these advancements, machine learning is poised to transform the landscape of cancer detection and treatment, offering renewed hope in the ongoing battle against this widespread disease.

Studies on breast cancer have leveraged the publicly available Wisconsin Breast Cancer Dataset (WBCD). For example, Asri et al. (2016) evaluated the effectiveness of four machine learning algorithms using this

dataset and discovered that the Support Vector Machine (SVM) approach outperformed the others. Similarly, Kumar et al. (2022) employed a stacking ensemble method for classification, whereas Abdulkareem and Abdulkareem (2021) assessed two ensemble models utilizing the WBCD. However, the conclusions regarding the superiority of one algorithm over another may be flawed. Each algorithm operates under specific assumptions about the dataset, which means that comparisons should only occur between algorithms that share similar assumptions. To facilitate a fair comparison, two essential questions must be addressed: (1) Do the algorithms have an equal opportunity to perform well with the dataset? and (2) Are the algorithms compared using their optimal parameters? Unfortunately, many studies neglect these critical inquiries. Each algorithm possesses its own strengths and weaknesses, and they should only be utilized when the underlying conditions are satisfied. Moreover, certain parameter values are necessary for an algorithm to achieve optimal performance. Therefore, all algorithms should be compared only when they are functioning at their best. This paper is dedicated to identifying the requirements for k-nearest neighbors (kNN) to operate at an optimal level when applied to the WBCD.

The kNNis a supervised machine learning algorithm used for classification and regression tasks (Hu et al., 2020). It is based on the concept of similarity between data points and operates by finding the closest data points (neighbors) to make predictions. The kNN algorithm classifies an arbitrary data into the most common class among the k closest datapoints (Nguyen ET AL., 2023). For regression, it averages all the k closest datapoints and assign it to the arbitrary data. The fact that kNN is simple, non-parametric, and flexible, makes its application in image recognition (face and handwriting recognition), finance (stock market analysis and fraud detection) and healthcare (disease prediction and medical image analysis) more assessable.Although kNN is non-parametric and does not make any assumption about the underlying dataset, the algorithm has some inherent assumptions on which its performance depends. The assumptions include, (1) Similar datapoints are close to each other. (2) The dataset is representative of the entire population. (3) All features contribute equally. (4) The dataset is not extremely large. Based on the assumptions, kNN algorithm is useful when the dataset is relatively small. The algorithm requires space to store the entire dataset since it calculates the distance of each instance from the query point. Hence, large datasets are computationally expensive. Secondly, kNN is applicable to situations where feature scaling is applied. Scaling makes each feature compare on the same scale and allows distance metric more reasonable.

Numerous studies are currently being conducted that focus on comparing various algorithms using the WBCD (Wisconsin Breast Cancer Diagnostic) dataset. Researchers have sought to demonstrate the effectiveness of Support Vector Machines (Muralidharr et al., 2023), ensemble models (Jabbar, 2021), and several other algorithms in achieving superior performance (Koca and Aktepe, 2024; Abunasser et al., 2023).

However, despite these efforts, none of the studies have recognized kNN as a leading algorithm in this context. This particular study aims to explore the optimal conditions under which kNN can be utilized with the WBCD dataset. By doing so, it seeks to enhance the understanding of the capabilities and limitations in this domain. The insights gained from this research will serve as a valuable resource for future investigations into the WBCD dataset, providing guidance for subsequent studies aimed at more effective algorithmic approaches.

2 k-NEAREST NEIGHBORS

The algorithm for kNN is discussed in this section, along with the metrics and parameter tuning for optimality.The algorithm, based on the code of Halder et al. (2024), for executing kNN is given below;

Inputs: (1) A dataset $D = \{(\mathbf{x}_i, y_i), i = 1 \cdots N\}$, $\mathbf{x}_i \in \mathbb{R}^d$ is the feature vector of the ith instance and y_i is class label. (2) A query instance $\boldsymbol{q} = \mathbb{R}^d$. (3) The number of neighbors k. (4) A distance metric $d(\mathbf{x}_i, \boldsymbol{q})$

The output will be a predicted class $\hat{y}$ for the query instance q. The algorithm is therefore given below;

Step #1: Compute distance $d_i = d(\mathbf{x}_i, \boldsymbol{q})$ for each x_i.

Step #2: Sort the dataset D based on the computed distances so that

$$d(\mathbf{x}_1, \boldsymbol{q}) \leq d(\mathbf{x}_2, \boldsymbol{q}) \leq \cdots \leq d(\mathbf{x}_N, \boldsymbol{q}).$$

Step #3: Select a subset D_k of D so that it contains the first k entries from D, that is,

$$D_k = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_k, y_k)\}.$$

Step #4: Count the class label y in D_k so that the number of instances with label y is s and the number of instances with different label is t.

Step #5: If $s > t$, then classify $\boldsymbol{q}$ as y. If $s < t$, then classify $\boldsymbol{q}$ as the other label. If $s = t$, then randomly select.

2.1 Distance Metrics

Distance is key concept in kNN. Although there are some other metrics, the most common generalized metric used to measure the distance between a point $\mathbf{q} = (q_k)_{k=1\cdots n}$ and the instances $\mathbf{x}_i = (x_{i,k})_{k=1\cdots n}$ is the Minkowski metric defined as

$$d(\mathbf{q}, \mathbf{x}_j) = \left(\sum_{k=1}^{n} |q_k - x_{i,k}|^p \right)^{\frac{1}{p}}$$

where p can be adjusted depending on the nature of the data. Specifically, the Manhattan distance is obtained for $p = 1$, and the Euclidean distance is obtained for $p = 2$ (see Liang et al., 2023; Halder et al., 2024). The algorithm is structured to rank the k closest points and make decision based on the most common class (for classification) among the k classifications or calculate the weighted average of the k classifications (for regression).

2.2 Hyperparameter

Hyperparameter tuning are required to optimize the algorithms (Bischl et al., 2023). The hyperparameters of kNN are the number of neighbors (k), the distance metric, weighting scheme, the search algorithm and the power parameter p. For the choice of k, the rule of thumb is to start with $k = \sqrt{n}$ and check if the optimal k can be found within the neighborhood of $k = \sqrt{n}$ (such as $k = \sqrt{n} \pm 5$). This does not necessarily find the optimal k (Kusonkhum et al., 2023). However, cross-validation is a common practice in choosing k. This approach starts by splitting the dataset into k-folds. The model is trained on $k - 1$ folds and tested on the remaining fold, and the process is repeated for each fold as the candidate k. The average performance metric is calculated to get identify the optimal k.

2.3 Performance Metrics

The performance metrics are used to measure the effectiveness of the model in classifying the query point correctly. The confusion matrix serves as the first metric that can be easily seen and is shown in Figure (2). The True positive is the number of correct classifications of Class 1 instances into Class 1, the false positive is the number of wrong classification of instances into Class 1, the false negative is the number of wrong classifications into Class 2, and the true negative is the number of correct classifications of Class 2 instances into Class 2.

	Class 1	Class 2
Class 1	True Positive	False Positive
Class 2	False Negative	True Negative

Figure 2: Confusion Matrix

The confusion matrix gives a quick overview of the counts, but more in-depth metrics include precision, recall, and F1-score and accuracy and they are defined by Foody (2023) as

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive + False Positive}} \times 100\%,$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive + False Negative}} \times 100\%$$

$$\text{F1score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision + Recall}} \times 100\%$$

$$\text{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{Total Instances}} \times 100\%$$

Higher values of these metrics indicate better performance of the model.

3 Data Collection, Description and Preprocessing

The data used in this study was downloaded from Kaggle website. The dataset was curated using fine needle aspirations to remove some tissues from the solid breast mass of 569 breast cancer patients in Wisconsin. Shafique et al. (2023), Abdulkareem and Abdulkareem (2021), and Asri et al. (2016) have all used this data. Of the 569 observations, 212 are cancerous and 357 are benign. The diagnosis types are denoted as M for malignant and B for benign. Ten (10) characteristics of the nucleus of the digitised image of the breast mass were recorded. The measured characteristics are radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. The mean, the standard error and the worst measurement of each of the ten characteristics were recorded. This implies that there are a total of 30 features in the dataset. For example, the feature named `texture` is recorded three times as **`texture_worst`** (calculated by computing the mean of the three largest values), **`texture_mean`** (the mean of the measured textures), and **`texture_se`** (the standard error of the textures). There are therefore a total of 30 features and one (1) label. The means of each feature are extracted from the dataset and we have a total of ten features and one label for the analysis. The distribution of each class of diabetes is shown in Table 1.

Table 1: Data Summary

diagnosis	frequency	percentage
B	357	62.74165
M	212	37.25835

The summary statistics is shown in Table (2). There are a total of 569 instances with no missing value in any of the instance.

N Rows	N	Mean	Stdev	Min	Max	Missing	CV
radius	569	14.1273	3.524	6.981	28.11	0	24.945
texture	569	19.2896	4.301	9.71	39.28	0	22.2971

perimeter	569	91.969	24.299	43.79	188.5	0	26.4208
area	569	654.8891	351.9141	143.5	2501	0	53.7364
smoothness	569	0.0964	0.0141	0.0526	0.1634	0	14.5954
compactness	569	0.1043	0.0528	0.0194	0.3454	0	50.6155
concavity	569	0.0888	0.0797	0	0.4268	0	89.7753
concave_points	569	0.0489	0.0388	0	0.2012	0	79.3204
symmetry	569	0.1812	0.0274	0.106	0.304	0	15.1325
fractal_dimension	569	0.0628	0.0071	0.05	0.0974	0	11.243

List 1- Results of Data statistics

Comparing the mean of concave points with the mean of area, it becomes clear that the dataset is not on the same scale and require scaling. The mean of the dataset is normalized and the variance is scaled using the formula

$$\mathbf{x}_{i,scaled} = \frac{\mathbf{x}_i - \mu_i}{\sigma_i},$$

where $\mathbf{x}_i$ is the feature i, $\mathbf{x}_{i,scaled}$ is the scaled version of $\mathbf{x}_i$, μ_i is the mean of the feature i and σ_i is the variance of the feature i. This scaling brings the all the means to 0 and standard deviation 1. Now, the new dataset has ten features (each with mean 0 and standard deviation 1) and one label.

4 Search for Optimal Parameters

The dataset is split into two datasets. The first dataset consists of 80% of the entire dataset and is used to train the model. The second dataset consists of the remaining 20% of the entire dataset for testing. In the search for the optimal value of k, the values of k are varied for three most common distance metrics and the results are compared. The distance metrics are the Chebyshev metric, Euclidean metric and Manhattan metric and they are defined as follows

$$d_{cheby}(\mathbf{q}, \mathbf{x}_j) = \max_k |q_k - x_{i,k}|,$$

$$d_{Euclid}(\mathbf{q}, \mathbf{x}_j) = \left(\sum_{k=1}^{n} |q_k - x_{i,k}|^2 \right)^{\frac{1}{2}},$$

$$d_{Manh}(\mathbf{q}, \mathbf{x}_j) = \sum_{k=1}^{n} |q_k - x_{i,k}|.$$

By starting with the rule of thumb, we set $k = \sqrt{569} \approx 24$. Values of k in the neighborhood of $k = 24$ are chosen from $k = 18,19,\cdots,29$. Since it common to consider smaller k's, $k = 3$ and $k = 8$ are also considered. The results are presented in Table 2. The precision, recall, and F1-score are recorded in percentage and it is clear that although Chebyshev and Manhattan metrics perform well at some points, the optimal behaviour occurs when $k = 21$ and $k = 23$ for Euclidean metric. At these values of k, the model has 97.26% precision for Benign, 97.56% precision for Malignant, 98.61% recall for Benign, 95.24% for Malignant, F1-score of 97.93% for Benign, and 96.39% F1-score for Malignant.

Table 2: Classification Table (P=Precision, R=Recall, F1 = F1-score

k		Chebyshev			Euclidean			Manhattan		
		P	R	F1	P	R	F1	P	R	F1
3	B	97.14	94.44	95.77	95.65	91.67	93.62	94.52	95.83	95.17
	M	90.91	**95.24**	93.02	86.67	92.86	89.66	92.68	90.48	91.57
8	B	97.14	94.44	95.77	95.65	91.67	93.62	94.52	95.83	95.17
	M	90.91	**95.24**	93.02	86.67	92.86	89.66	92.68	90.48	91.57
18	B	97.14	94.44	95.77	95.95	**98.61**	97.26	94.52	95.83	95.17
	M	90.91	**95.24**	93.02	97.50	92.86	95.12	92.68	90.48	91.57
19	B	97.14	94.44	95.77	95.65	91.67	93.62	94.52	95.83	95.17
	M	90.91	**95.24**	93.02	86.67	92.86	89.66	92.68	90.48	91.57
20	B	97.14	94.44	95.77	95.65	91.67	93.62	94.52	95.83	95.17
	M	90.91	**95.24**	93.02	86.67	92.86	89.66	92.68	90.48	91.57
21	B	93.33	97.22	95.24	**97.26**	**98.61**	**97.93**	92.21	**98.61**	95.30
	M	94.87	88.10	91.36	**97.56**	**95.24**	**96.39**	97.30	85.71	91.14
22	B	92.11	97.22	94.59	94.67	**98.61**	96.60	92.21	**98.61**	95.30
	M	94.74	85.71	90.00	97.44	90.48	93.83	97.30	85.71	91.14
23	B	92.11	97.22	94.59	**97.26**	**98.61**	**97.93**	92.21	**98.61**	95.30
	M	94.74	85.71	90.00	**97.56**	**95.24**	**96.39**	97.30	85.71	91.14
24	B	92.11	97.22	94.59	93.42	**98.61**	95.95	92.21	**98.61**	95.30
	M	94.74	85.71	90.00	97.37	88.10	92.50	97.30	85.71	91.14
25	B	92.11	97.22	94.59	93.42	**98.61**	95.95	93.42	**98.61**	95.95
	M	94.74	85.71	90.00	97.37	88.10	92.50	97.37	88.10	92.50
26	B	92.11	97.22	94.59	93.42	**98.61**	95.95	93.42	**98.61**	95.95
	M	94.74	85.71	90.00	97.37	88.10	92.50	97.37	88.10	92.50
27	B	97.14	94.44	95.77	95.65	91.67	93.62	94.52	95.83	95.17

	M	90.91	**95.24**	93.02	86.67	92.86	89.66	92.68	90.48	91.57
28	B	97.14	94.44	95.77	95.65	91.67	93.62	94.52	95.83	95.17
	M	90.91	**95.24**	93.02	86.67	92.86	89.66	92.68	90.48	91.57
29	B	97.14	94.44	95.77	95.65	91.67	93.62	94.52	95.83	95.17
	M	90.91	**95.24**	93.02	86.67	92.86	89.66	92.68	90.48	91.57

The results obtained by calculating the accuracy from using the three metrics for each k are recorded in Table (3). The table further confirms that the optimal value of k is 21 or 23. Figure (3) shows that the Chebyshev metric has similar behaviour for most of the choice of k, except at few points where the performance drops. Manhattan hits its optimal performance at $k = 24$ and 25, but has similar behaviour other values of k. Finally, the Euclidean metric performs the best with the performance optimal at $k = 21$ and 23.

Table 3: Accuracy for the threemetricsforvariouschoiceofk

k	3	8	18	19	20	21	22	23	24	25	26	27	28	29
Chebyshev	94.74	94.74	94.74	94.74	94.74	93.86	92.98	92.98	92.98	92.98	92.98	94.74	94.74	94.74
Euclidean	92.11	92.11	96.49	92.11	92.11	**97.37**	95.61	**97.37**	94.74	94.74	94.74	92.11	92.11	92.11
Manhattan	93.86	93.86	93.86	93.86	93.86	93.86	93.86	93.86	93.86	94.74	94.74	93.86	93.86	93.86

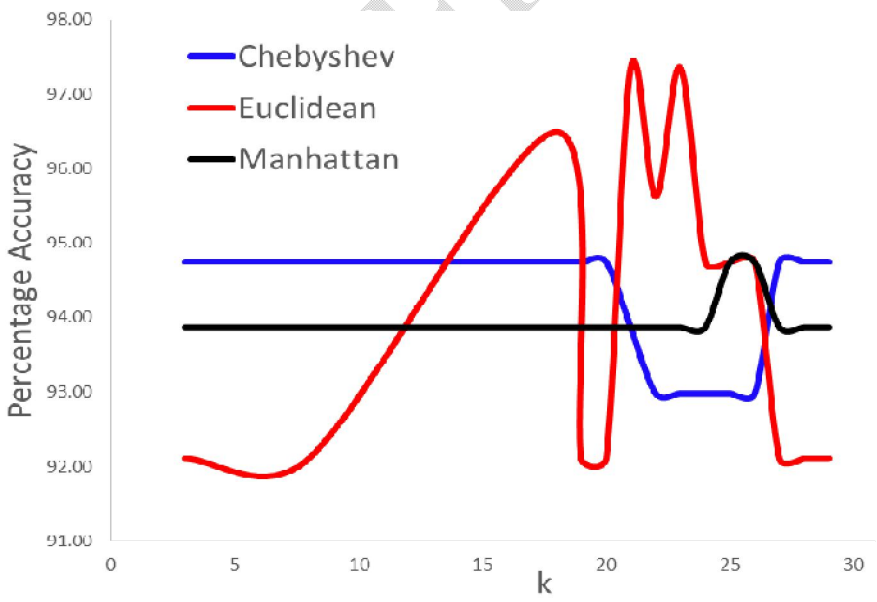

Figure 3: Accuracy curve for the three metrics for various k varies

5 Feature Importance

The kNN algorithm does not inherently have the feature importance property because it is a non-parametric instance-based algorithm. Hence, proposed approaches for obtaining the importance of the features include

permutation importance, weighted distance metrics, and wrapper method. The permutation importance approach is the most common method and it is adopted in this study. The approach follows the following algorithm;

Step #1: Train the kNN model.

Step #2: Calculate the performance metrics of the model.

Step #3: Randomly shuffle the values of one feature in the dataset.

Step #4: Recalculate the performance with the shuffled dataset.

Step #5: The drop in performance indicates the feature's importance.

Having established the optimal performance of kNN for the dataset occur for Euclidean metric at $k = 21$ and 23, the feature importance of each feature is demonstrated in Table (4) and Figure (4). The figure shows that both models (models for $k = 21$ and $k = 23$) agree on the impact of all features except the compactness. The Spearman rank correlation for the feature importance produced from the two models is 0.9341. This shows that the two results agree strongly positively.

Table 4: Feature Importance

Feature	Importance for $k = 21$	Importance for $k = 23$
Area (A)	0.005232	0.005232
Compactness (C)	-0.00175	0.00177
Concavity (Cv)	0.001739	0.001754
Perimeter (P)	0.00177	0.005294
Concave Points (Pt)	-0.00177	0.003524
Radius (R)	0.01759	0.015852
Smoothness (S)	-0.00177	-0.00175
Symmetry (Sy)	0.003524	0.005294
Texture (T)	0.003509	0.003524

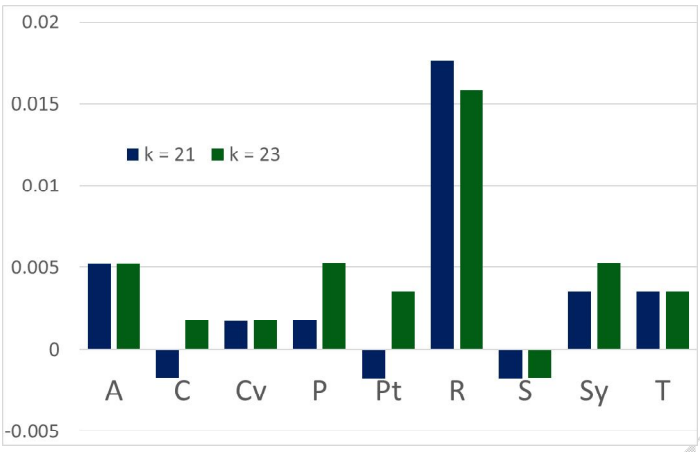

Figure 4: Feature Importance Bar Chart

6 Conclusion

This study has analyzed the dataset from 569 breast cancer patients in the United States using the kNN machine learning algorithm. The study explored the impact of distance metrics on the dataset by evaluating three different distance metrics, namely; Chebyshev, Manhattan, and Euclidean metrics. Each metric is used on the dataset for various values of k. By evaluating the performance of the resulting models, it is found that the optimal k for this dataset is obtained at $k = 21$ and $k = 23$. The optimalprecision for Benign and Malignant are 97.26% and 97.56% respectively, the optimal recall for Benign and Malignant are 98.61% and 95.24% respectively, and the optimal F1-score for Benign and Malignant are 97.93% and 96.39% respectively. The optimal accuracy is 97.37%. Moreso, there is a strong positive agreement between the two optimal models. They both agree on the importance of all features except the compactness. More analysis is therefore required to estimate the importance of compactness.

Availability of data and materials: The dataset(s) supporting the conclusions of this article is(are) available in the Kaggle Website, https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

REFERENCES

1. Abdulkareem, S. A. & Abdulkareem, Z. O. (2021), 'An Evaluation of the Wisconsin Breast Cancer Dataset using Ensemble Classifiers and RFE Feature Selection Technique', *International Journal of Sciences: Basic and Applied Research (IJSBAR)* **55**(2), 67–80.
2. Abunasser, B. S., AL-Hiealy, M. R. J., Zaqout, I. S., & Abu-Naser, S. S. (2023, May). Literature review of breast cancer detection using machine learning algorithms. In *AIP Conference Proceedings* (Vol. 2808, No. 1). AIP Publishing.

3. Asri, H., Mousannif, H., Moatassime, H. A. & Noel, T. (2016), 'Using machine learning algorithms for breast cancer risk prediction and diagnosis', *Procedia Computer Science* **83**, 1064–1069. Balkenende
4. Balkenende, L., Teuwen, J. & Mann, R. M. (2022), 'Application of deep learning in breast cancer imaging', *Seminars in Nuclear Medicine* **52**(5), 584–596.
5. Beam, A. L. &Kohane, I. S. (2018), 'Big data and machine learning in health care', *JAMA* **319**(13), 1317.
6. Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., ... & Lindauer, M. (2023). Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, ***13***(2), e1484.
7. Chan, R. C. K., To, C. K. C., Cheng, K. C. T., Yoshikazu, T., Yan, L. L. A. & Tse, G. M. (2022), 'Artificial intelligence in breast cancer histopathology', *Histopathology* **82**(1), 198–210.
8. Foody, G. M. (2023). Challenges in the real-world use of classification accuracy metrics: From recall and precision to the Matthews correlation coefficient. *Plos one*, *18*(10), e0291908.
9. Giaquinto, A. N., Sung, H., Miller, K. D., Kramer, J. L., Newman, L. A., Minihan, A., ... & Siegel, R. L. (2022). Breast cancer statistics, 2022. *CA: a cancer journal for clinicians*, ***72***(6), 524-541.
10. Halder, R. K., Uddin, M. N., Uddin, M. A., Aryal, S., &Khraisat, A. (2024). Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. *Journal of Big Data*, ***11***(1), 113.
11. Hu, J., Peng, H., Wang, J., & Yu, W. (2020). kNN-P: A kNN classifier optimized by P systems. *Theoretical Computer Science*, *817*, 55-65.
12. Jabbar, M. A. (2021). Breast cancer data classification using ensemble machine learning. *Engineering & Applied Science Research*, *48*(1).
13. Koca, Y. B., & Aktepe, E. (2024). Evaluation of Missing Data Imputation Methods and PCA Techniques for Machine Learning Models in Breast Cancer Diagnosis Using WBCD. *Türk Doğave Fen Dergisi*, *13*(3), 109-116.
14. Kumar, M., Singhal, S., Shekhar, S., Sharma, B. & Srivastava, G. (2022), 'Optimized stacking ensemble learning model for breast cancer detection and classification using machine learning', *Sustainability* **14**(21), 13998.
15. Kusonkhum, W., Srinavin, K., &Chaitongrat, T. (2023). The Adoption of a Big Data Approach Using Machine Learning to Predict Bidding Behavior in Procurement Management for a Construction Project. *Sustainability*, *15*(17), 12836.
16. Liang, Y., Pan, Y., Yuan, X., Jia, W., & Huang, Z. (2023). Surrogate modeling for long-term and high-resolution prediction of building thermal load with a metric-optimized KNN algorithm. *Energy and Built Environment*, *4*(6), 709-724.
17. Lim, Z. L., Ho, P. J., Khng, A. J., Yeoh, Y. S., Ong, A. T. W., Tan, B. K. T., ... & Hartman, M. (2022). Mammography screening is associated with more favourable breast cancer tumour characteristics and better overall survival: case-only analysis of 3739 Asian breast cancer patients. *BMC medicine*, *20*(1), 1-19.
18. Muralidharr, T., Madhav, P. S., Kumar, P. P., & Tiwari, H. (2023). Analyzing the Performances of Different ML Algorithms on the WBCD Dataset. In *Deep Learning: Theory, Architectures and Applications in Speech, Image and Language Processing* (pp. 73-89). Bentham Science Publishers.
19. Nguyen, L. V., Vo, Q. T., & Nguyen, T. H. (2023). Adaptive knn-based extended collaborative filtering recommendation services. *Big Data and Cognitive Computing*, *7*(2), 106.

20. Singh, L. K., Khanna, M. & Singh, R. (2023), 'Artificial intelligence based medical decision support system for early and accurate breast cancer prediction', *Advances in Engineering Software* **175**, 103338.
21. Weinberg, R.A., (1996). How cancer arises. *Scientific American*, *275*(3), pp.62-70.