HIGH SCHOOL EDITION
Journal of Student Research

An Analysis of the k-Nearest Neighbor Classifier to Predict Benign and Malignant Breast Cancer Tumors

Sahasra Chatakondu[1] and Kevin Zhai[#]

[1]Metea Valley High School
[#]Advisor

ABSTRACT

Because of Breast Cancer's high mortality rate and being a leading cause of death among women worldwide, there has been importance given to machine learning (ML) algorithms to detect early signs of benign and malignant tumors effectively. Assistance from ML classifiers allows for a more efficient evaluation of mammographic results, surpassing the capabilities of radiologists who manually classify extensive patient data. This study aims to evaluate the effectiveness of the k-Nearest Neighbor (kNN) classifier in characterizing cancer tumor stages based on concavity, texture, area, perimeter, and smoothness. We employ scatterplots to differentiate between benign and malignant classes using the Breast Cancer Wisconsin Dataset (WBCD) from the University of California at Irvine Machine Learning Repository. Employing the k-Fold Cross Validation (k-FCV) technique, we determine the optimal value for k to assign anonymous data to their respective categories. The analysis conducted in this study finds that the most favorable value for the hyperparameter k is 12, resulting in a highly effective diagnostic outcome from administering four distinct tests. Given the absence of a predefined value for the k parameter, guesswork could lead to accuracy errors and misdiagnosis; therefore, employing k-FCV provides a more precise approach to determining the optimal class for unknown tumor attributes. Additionally, meticulous preprocessing of this dataset and measuring how different data splits impact accuracy are used to organize the data effectively and achieve reliable results. Recognizing that early detection is essential in preventing Breast Cancer-related deaths, ML techniques like kNN can greatly reduce mortality rates associated with the disease.

Introduction

Statistics obtained from the International Agency for Research on Cancer (IARC) reveal an increase in the incidence of Breast Cancer, surging from 10 million cases in 2000 to 19.3 million cases in 2020. Treatment efficacy is significantly enhanced when the cancer is detected at early stages, yielding a survival probability of 90% or higher[1]. However, a considerable challenge arises due to the frequency of breast lumps, where non-cancerous abnormalities are deemed benign, while malignant tumors pose a menacing and rapidly progressing threat. Accurate diagnosis of Breast Cancer relies on the classification of tumors. Yet, radiologists experience a 15% rate of misjudgment in interpreting mammogram scans, resulting in both false positives and negatives, both unfavorable outcomes[2]. Given these challenges, machine learning (ML) algorithms, such as the k-Nearest Neighbors (kNN) approach, have been adapted to precisely ascertain tumor characteristics and classify them as either benign or malignant[3].

In this study, we investigate the properties of the k-Nearest Neighbors (kNN) algorithm, along with the underlying processes that contribute to its optimal outcomes. This approach to classification, initially proposed by Fix and Hodges in 1951 and subsequently refined by Cover and Hart in 1967[4], has found applications in diverse domains such as pattern recognition, object recognition, text categorization, and medicine. In addition to its prominent usage within the medical field, technological corporations leverage this technique to customize user experience and adapt technology to their liking[5].

kNN is a supervised lazy-learning and non-parametric classifier. The term "non-parametric" signifies that prior knowledge of the underlying data distribution is unnecessary, and that the model does not rely on assumptions about the underlying distribution. It does not require any predefined structure and adapts flexibly to the provided dataset. On the other hand, "lazy-learning" implies that the generalization of test data is postponed until the system encounters the actual testing data. Such classification methods typically necessitate a shorter training time but tend to have longer prediction times[4], as they memorize the presented data instead of conducting extensive calculations[6]. kNN's response time can be quite slow when presented with large, high-dimensional data sets.

The proficiency of kNN lies in its ability to generalize prominent attributes within a dataset, allowing kNN to classify unknown points for classification tasks. Through the learning process, the model acquires the ability to recognize the crucial attributes within the data, enabling accurate categorization and prediction[4].

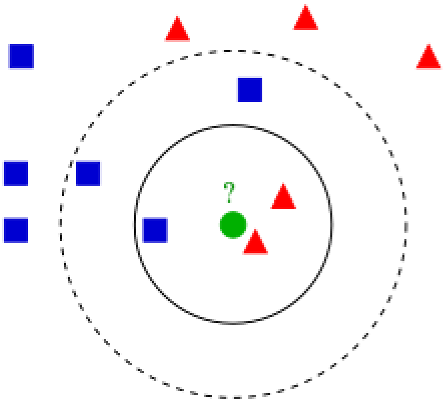


Figure 1[7]**.** Example of k-nearest neighbor classification. The green circle is an unclassified point and relies on a *k* number of neighbors to determine if it is a red triangle or a blue square.

The *k* value in the classifier's title denotes how many neighboring points are evaluated to categorize an unclassified data point. This can be seen in Figure 1.1, in which the green circle depicts a sample to be classified, and the blue squares and orange triangles represent labeled data corresponding to two distinct classes. In addition, the inner solid circle depicts when the *k* value is set to 3 and the outer dotted one depicts when k is set to 5. A smaller value of *k* results in a more flexible model, but can lead to overfitting since noise and outliers may dominate. In contrast, a larger value of *k* can lead to over-smoothing and potentially underfitting while missing local patterns. The optimal value of *k* may depend on the dataset's characteristics and should be selected using techniques like k-Fold Cross Validation[8] (k-FCV), proving the "no-free-lunch theorem"[20]. The classified point is compared to its *k* nearest classified points.

Previous Works

Here, we briefly review other papers that use various ML techniques for classifying Breast Cancer tissues.

Hiba and Hajar et al.[9] studied the accuracy of kNN and various other classifiers such as Support Vector Machine, Decision Tree, and Naive Bayes. kNN took 0.01 seconds to build a model, and the accuracy was between 95.12% and 95.28%. Their experiment of the WBCD showed that the Mean Absolute Error was 0.04 or 4%, equivalent to 33 incorrect instances by kNN.

Research conducted by Mandeep and Pooja et al.[10] examined the performance of kNNs- along with Support Vector Machine, Logistic Regression, and Naive Bayes—using specific distance metrics. The Euclidean Distance scored the same as the Manhattan Distance (see Methods & Materials) for training accuracy, 100%, but proved slightly higher at 95.68%, in contrast to 94.96% on the testing data.

Anjali and Chintan[11] compared kNN to a Decision Tree and Bayesian Network. Using the dataset of the University of Wisconsin Hospital, their findings assured kNN took 0.02 seconds to come up with a model, 94.9928% Correctly Classified Instances, and 0.0487 Mean Absolute Error. Their work concluded that Naïve Bayes was a superior algorithm due to its high accuracy and low error percentage.

Amrane et al.[12] also implemented Naïve Bayes and kNN classifiers. Using the WBCD, they divided the data to test the Euclidean distance between the sample points and implemented k-FCV to evaluate each classifiers' accuracy. After measuring the mean and standard deviations of the predictions on the test set, the comparison showed that kNN was the most accurate, with a 97.51% success and a minimum error rate compared to the other model. The study intended to find the finest classifier to distinguish different tumor types.

In another study conducted by Zoelkarnain and Herman et al.[13], a comprehensive analysis of cervical cancer data was performed, including characteristics similar to those present in the WBCD. The researchers observed that employing the kNN methodology with the k-FCV algorithm improved the classification process. Specifically, an optimal k-value proved important, with their findings indicating a preference for a 3-Fold approach, consistently achieving classification accuracies exceeding 90% for their specific dataset.

Methods & Materials

Dataset

The WBCD[14] is a typical dataset used for binary classification tasks where the goal is to train a model to predict the diagnosis based on the provided features. The characteristics were derived from digitized images of fine needle aspirate (FNA) of breast mass, then used to classify breast cancer as benign or malignant. The dataset includes 569 patient samples, 357 benign and 212 malignant, with 32 attributes.

Attribute Information (WBCD)

1. Patient ID number: unique identification number for each sample in the dataset, not valuable for diagnosis
2. Diagnosis (M = malignant, B = benign): represents the target variable and contains the classification labels for each sample
3. (3-32) ten real-valued features are computed for each cell nucleus:
    a) Radius: mean of distances from the center to points on the perimeter
    b) Texture: standard deviation of gray-scale values
    c) Perimeter: the perimeter of the cell nuclei
    d) Area: the area of the cell nuclei
    e) Smoothness: local variation in radius lengths
    f) Compactness: $(perimeter^{2} \div area) - 1$
    g) Concavity: the severity of concave portions of the contour
    h) Concave points (number of concave portions of the contour)
    i) Symmetry
    j) Fractal dimension: ("coastline approximation" -1)

Distance Metrics

Here displayed are a few distance measures kNN typically uses to compute space between neighbors. These formulas assume two vectors are given x and y, with x = $(x_{1,} x_{2,}\ldots,x_{n})$ and y = $(y_{1,} y_{2,}\ldots,y_{n})$ having numerical values[15]. For this study, we focus on a typical distance metric for continuous or numerical data: Euclidean Distance (2), based on the

Pythagorean Theorem and performed in Euclidean Space, which was derived from the generalized Minkowski Distance (2) by setting p=2 p=2. Distance can be manipulated, but in this paper, we control this variable to focus on another aspect of achieving accuracy.

Minkowski Distance

$$D(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^{p}\right)^{1/p}$$

The absolute difference between each feature value of the two points is raised to a positive power of p, summed across all features, and then the result is raised to the power of 1/p. This computation yields the Minkowski distance between the two points.

Euclidean Distance

$$D(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)}^{2}$$

Subtract x_i and y_i- the respective coordinates of the two points in each dimension (i = 1 to n)- square each difference, sum up the squared differences in each dimension and take the square root of the sum to obtain the Euclidean distance.

Confusion Matrix

Confusion matrices, commonly employed for binary classification tasks, provide valuable insights into a model's accuracy. In the context of Breast Cancer prediction, the occurrence of misdiagnosis poses a concern. Specifically, false positives and false negatives are possible, resulting in adverse consequences for patients as these outcomes yield incorrect and misleading results[16].

A true positive (TP) refers to the number of instances from the positive set that the classification model correctly identifies as belonging to the positive set.

A true negative (TN) refers to the number of instances outside the positive set that the model correctly recognizes as not belonging to the positive set.

A false positive (FP) represents the number of instances from the negative set that the model incorrectly classifies as belonging to the positive set.

A false negative (FN) represents the number of instances from the positive set that the model incorrectly classifies as not belonging to the positive set[4].

Predicted Class

Actual Class +TP (True Positive) FN (False Negative)
FP (False Positive) TN (True Negative)

Accuracy & Margin of Error

The margin of error plays a crucial role in assessing the proximity between a model's prediction and the actual proportion value. It serves as a statistical measure that quantifies the sampling variability inherent in an estimate. As the margin of error increases, the precision of the prediction diminishes, whereas a smaller margin of error signifies a higher degree of precision in the model's performance.

Mathematically, the margin of error is calculated using the formula: Margin of error = $z * \sigma/\sqrt{n}$

In this formula, the margin of error is determined by dividing the standard deviation (σ) by the square root of the sample size (n), and then multiplying it by a z-score (z) corresponding to the desired confidence level. This computation provides an estimation of the error present in the model[17].

Although statistical estimates are not inherently perfect, they provide valuable insights into survey results. When evaluating the accuracy of a kNN model, the sample size varies with each iteration. Consequently, during each run of the model, the accuracy might deviate slightly from the previous run due to the distribution of testing points across the graph. Using this formula, we attempt to interpret the reliability and precision of the model accurately.

Normalization

The variation in attribute ranges within a given dataset requires the application of normalization as a preprocessing technique for data cleaning. This procedure ensures that the data values are transformed to a standardized scale. Numerous methods can be employed to accomplish this process; In a scientific investigation conducted by Henderi et al.[18], a comparative analysis of z-score and min-max normalization techniques were conducted on the WBCD. Notably, min-max normalization showed an accuracy rate of 98%, while z-score normalization demonstrated a slightly lower accuracy of 97%. Despite the marginal disparities, our study uses the min-max normalization technique for all continuous attributes to achieve optimal outcomes.

Cross-Validation

As previously indicated within this research paper, k-FCV is paramount in accurately diagnosing Breast Cancer, preferred over assigning an uninformed value to the parameter *k*. The use of k-FCV allows the model to systematically evaluate multiple data splits, enabling the identification of an optimal *k* value for predicting outcomes on unseen data. Conventionally, adopting a fixed value for *k*, such as 5 or 10, is commonly practiced, particularly when dealing with large datasets[19]. Nonetheless, in our study, we opt to explore various *k* values to determine the most favorable choice at the cost of considerable time and effort. By employing a stepwise approach within our code, we establish a training method with values ranging from 1 to 70, which are systematically tested, aiming to obtain the most influential variable and produce the highest achievable accuracy.

Software implementation

The R programming language (Version 4.3.0) implemented under R Studio was employed for this study. To enhance the performance and facilitate the implementation of various algorithms, several libraries were downloaded. These libraries include readr, gmodels, dplyr, ggplot2, and caret.

Testing

We conducted an empirical study to assess the impact of varying values of *k* on the accuracy of the provided model. Given that kNN is categorized as a lazy learning algorithm, we seek to evaluate the performance of the training dataset using both smaller and larger values of *k*. After normalizing all the values within a feature, and replicating the WBCD dataset, it becomes necessary to partition the variable into separate training and test sets. To ensure consistency across all attributes being divided, the creation of a new variable for the first column, "diagnosis", is needed for predictive purposes. Subsequently, the kNN algorithm is employed to forecast the labels for the test set. An essential component of this procedure involves the generation of a cross-tabulation table, which allows for the examination of true positives, false positives, true negatives, and false negatives in relation to the predicted and actual labels. During this phase, it is essential to use the frequencies (instead of proportions) for the purpose of calculating the chi-square statistic.

The time required for making predictions using kNN or other lazy learning algorithms can be greater than that of other machine learning techniques. In our case, the dataset under consideration possesses multiple dimensions, as diverse columns that require processing. However, when subjected to testing, the model exhibits fast results. The empirical measurements reveal consistent performance, with each prediction cycle yielding a consistent time interval of approximately 1.05 seconds, leading to the generation of the corresponding confusion matrix. The conclusion is based on how fast each attempt of kNN took to generate a cross-tabulation table with respective preprocessing done prior to testing. We measure this time by pressing run on our code and timing it with a stopwatch to calculate the efficiency. There is a possibility for human error in this attempt, but the report is an accumulated average of 10 iterations run. Each is in close proximity to one another avoiding any outliers to be considered.

Results

While trying to use the kNN classifier, we implement other forms of noise reduction techniques to find the most effective method to predict benign and malignant Breast Cancer tumors. Based on the available data, the absence of cross-validation during model development typically results in utilization of a *k* value greater than 5 in order to obtain the most precise outcomes. In comparison to the reported percentages observed in previous studies, where the training dataset consisted of a larger number of predictive instances, our results display a similar trend. This similarity is caused because of the diminished significance of each individual prediction within a larger dataset. Although our model does not incorporate consideration of all potential factors—such as additional preprocessing techniques beyond normalization, alternative distance metrics, or noise elimination—our implemented code successfully generated a highly accurate model.

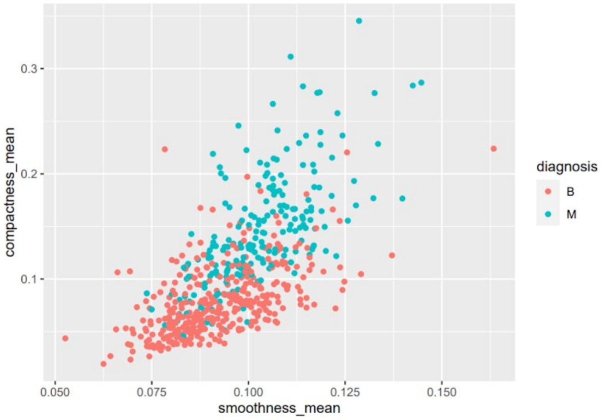


Figure 2. Indication of Benign (B) and Malignant (M) Breast Cancer Tumors based on Compactness and Smoothness Attributes

In Figure 5.1, the representation demonstrates the overlap of numerous data points associated with two randomly selected column features, smoothness_mean and compactness_mean. This figure shows how kNN can be used to visualize the different classes; the specific attributes in a data set do not make a difference in seeing the area of distinguishable benign and malignant classes. As seen in Figure 5.2, the use of two different attributes: texture_mean and radius_mean, still shows a distinguishable similar relationship like that from figure 5.1.

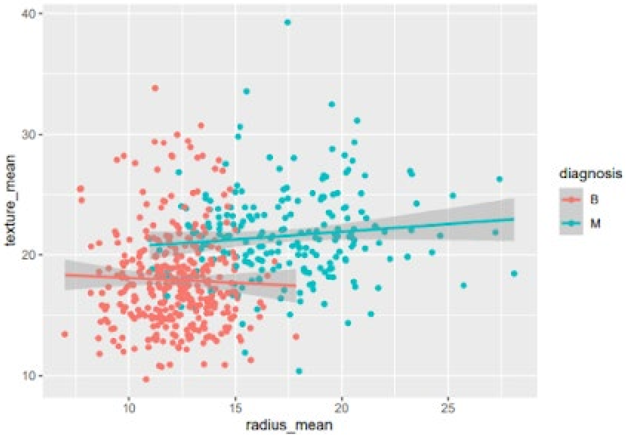
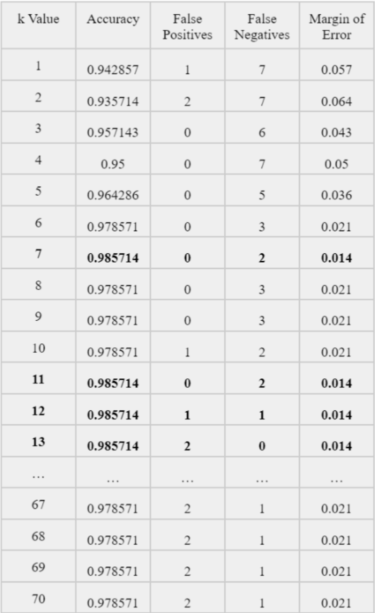


Figure 3. Indication of Benign (B) and Malignant (M) Breast Cancer Tumors based on Texture and Radius Attributes

For certain datasets, considering these factors becomes necessary, motivating our choice of employing k-FCV to determine the optimal *k* value of 12 specifically for this dataset. While the majority of the calculated percentages exhibited inaccuracies by incorrectly estimating only one or two instances, it is essential to recognize that such inaccuracies could have significant adverse consequences in real-world scenarios, particularly for individuals requiring a diagnosis.

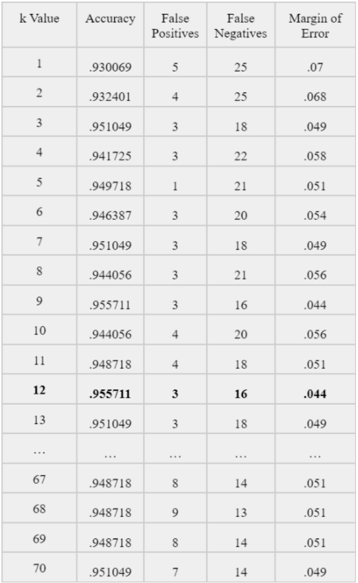
In the initial testing attempt, we employ a training-to-testing ratio of 75:25, resulting in the use of 140 patient data instances for classification by the model. We modified the approach by adopting a training-to-testing ratio of 25:75, therefore providing the model with 429 patient data instances for classification. Since kNN does not require a dedicated training phase, our objective is to discover how effective the *k* values produced by k-FCV would be when running kNN. Notably, the testing data is consistently randomized within the extensive dataset after establishing the percentage split. Each test has one to several numbers that satisfy the highest accuracy, but when compared with all four tests, there is a distinct answer common throughout all tests.

Table 1. Testing different kNN values to see accuracy and error on the data set with a 75:25 split of training to testing data. The highest accuracy was achieved by setting k = 7, 11, 12, 13



k Value	Accuracy	False Positives	False Negatives	Margin of Error
1	0.942857	1	7	0.057
2	0.935714	2	7	0.064
3	0.957143	0	6	0.043
4	0.95	0	7	0.05
5	0.964286	0	5	0.036
6	0.978571	0	3	0.021
7	**0.985714**	**0**	**2**	**0.014**
8	0.978571	0	3	0.021
9	0.978571	0	3	0.021
10	0.978571	1	2	0.021
11	**0.985714**	**0**	**2**	**0.014**
12	**0.985714**	**1**	**1**	**0.014**
13	**0.985714**	**2**	**0**	**0.014**
…	…	…	…	…
67	0.978571	2	1	0.021
68	0.978571	2	1	0.021
69	0.978571	2	1	0.021
70	0.978571	2	1	0.021

Table 2. Testing different kNN values to see accuracy and error on the data set with a 25:75 split of training to testing data. The highest accuracy was achieved by setting k = 12

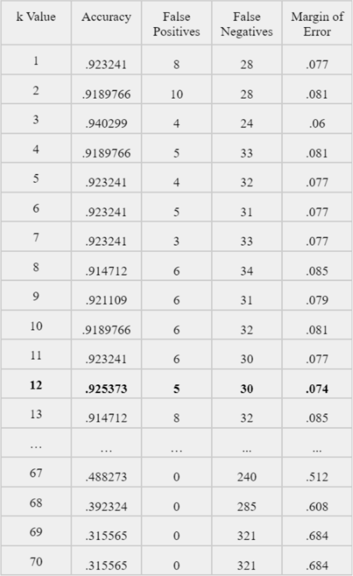


k Value	Accuracy	False Positives	False Negatives	Margin of Error
1	.930069	5	25	.07
2	.932401	4	25	.068
3	.951049	3	18	.049
4	.941725	3	22	.058
5	.949718	1	21	.051
6	.946387	3	20	.054
7	.951049	3	18	.049
8	.944056	3	21	.056
9	.955711	3	16	.044
10	.944056	4	20	.056
11	.948718	4	18	.051
12	**.955711**	**3**	**16**	**.044**
13	.951049	3	18	.049
…	…	…	…	…
67	.948718	8	14	.051
68	.948718	9	13	.051
69	.948718	8	14	.051
70	.951049	7	14	.049

Each iteration of our cross-validation model yielded varying optimal *k* values, prompting us to examine and compare their respective accuracy percentages in a systematic manner. In our initial investigation, as depicted in Table 4.3, four distinct values, namely 7, 11, 12, and 13, were tested, and their corresponding accuracy percentages and margin of error were analyzed. Notably, the value with the highest accuracy percentage and the lowest margin of error emerged as our primary focus. Subsequently, in an effort to identify an optimal value with greater precision, we modified the number of nearest neighbors. Our findings, presented in Table 4.4, narrowed down to two numbers: 9 and 12. The value 12 exhibited consistently high accuracy levels in both large and small datasets, achieving accuracy rates of 95.97% and 98.57%, respectively.

Previous research in the field has extensively examined the efficacy of utilizing kNN for detecting tumor outcomes. However, we hypothesized that further improvements could be achieved by employing k-FCV to optimize specific results within distinct datasets rather than relying on a predefined *k* value, such as 5 or 10. Our experimental findings indicated that, in the case of the WBCD dataset, an adaptable *k* value is necessary to fit various ML techniques applied to this particular dataset while accommodating different training and testing dataset sizes: for example, a 20-80 or 30-70 split. Utilizing a training dataset size below 100—an 18:82 percent training to testing split and would cause the testing data set to have high numbers—would not constitute an adequate measure, as it necessitates evaluating up to 70 nearest neighbors, therefore generating an excessively dense classification space that blocks the model's capacity to distinguish between benign and malignant since there is so much noise from other data points. When this scenario occurs, the error rate becomes extremely high, especially with higher values of *k*: Table 5.5 is a sample test run to indicate this problem. Notice how the k-FCV score of 12 was the single highest accuracy in the testing space of 469 patient data, further proving the importance of k-FCV.

Table 3. Testing different kNN values to see accuracy and error on the data set with an 18:82 split training to testing data. The highest accuracy was achieved by setting k = 12

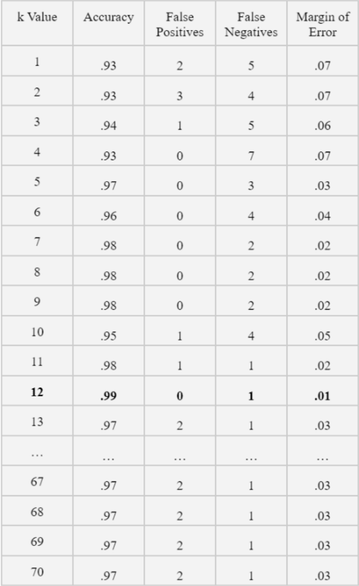


k Value	Accuracy	False Positives	False Negatives	Margin of Error
1	.923241	8	28	.077
2	.9189766	10	28	.081
3	.940299	4	24	.06
4	.9189766	5	33	.081
5	.923241	4	32	.077
6	.923241	5	31	.077
7	.923241	3	33	.077
8	.914712	6	34	.085
9	.921109	6	31	.079
10	.9189766	6	32	.081
11	.923241	6	30	.077
12	**.925373**	**5**	**30**	**.074**
13	.914712	8	32	.085
…	…	…	…	…
67	.488273	0	240	.512
68	.392324	0	285	.608
69	.315565	0	321	.684
70	.315565	0	321	.684

Along with the higher values being: 67, 68, 69, and 70, the margin of error went significantly above the threshold of 4%-8% to be considered decently accurate models with values reaching 68.4%[17].

One final examination conducted aims to investigate the contrasting subset of data from the previous data table. This particular test sought to establish an 82:18 percent division of the data, as shown through Table 5.6, thereby reevaluating the hypothesis that kNN is independent of the training data set and can effectively utilize 100 data points to accurately determine the nature of tumors using unclassified patient data. The outcomes of this analysis demonstrate significantly enhanced accuracy, given the reduced number of data points, as kNN is capable of assessing the proximity of benign and malignant instances based on a higher average. Remarkably, the k-FCV output demonstrates the highest level of accuracy among the examined set of numbers, yielding a value of 12.

Table 4. Testing different kNN values to see accuracy and error on the data set with an 82:18 split of training to testing data. The highest accuracy was achieved by setting k = 12



k Value	Accuracy	False Positives	False Negatives	Margin of Error
1	.93	2	5	.07
2	.93	3	4	.07
3	.94	1	5	.06
4	.93	0	7	.07
5	.97	0	3	.03
6	.96	0	4	.04
7	.98	0	2	.02
8	.98	0	2	.02
9	.98	0	2	.02
10	.95	1	4	.05
11	.98	1	1	.02
12	**.99**	**0**	**1**	**.01**
13	.97	2	1	.03
…	…	…	…	…
67	.97	2	1	.03
68	.97	2	1	.03
69	.97	2	1	.03
70	.97	2	1	.03

Conclusions

The k-Nearest Neighbors (kNN), an algorithm belonging to the category of supervised and lazy-learning methods, exhibits its high classification accuracy in identifying a given data point by considering the characteristics of a given data set. Preprocessing techniques, such as normalization and resampling procedures, such as k-Folds Cross Validation, were employed on datasets of varying sizes containing 140 and 429 values, respectively. It was observed that blindly assuming a fixed value for *k* to achieve optimal classification outcomes may not always be as accurate as the performance obtained through cross-validation procedures. In the case of the Wisconsin Breast Cancer Dataset, rigorous testing was conducted by applying the optimal *k* values from cross-validation to the kNN model itself, resulting in the optimal *k* value of 12, achieving a 98.57% accuracy on 25% of data and 95.57% accuracy on 75% of data. This study focused on exploring misdiagnosis reduction techniques and maximizing accuracy, utilizing only a single machine learning algorithm but thoroughly examining how certain techniques can enhance a model's accuracy. Regarding benign and malignant Breast Cancer tumors, implementing the kNN algorithm provides radiologists with a robust avenue to enhance the accuracy and efficiency of classifying a significant volume of patient data.

Although kNN does not require explicit training, alternative ML algorithms such as Support Vector Machines, Random Forest, Gradient Boosting algorithms, Naive Bayes, and Neural Networks necessitate both training and testing phases. While we did not execute all of these algorithms on the WBCD dataset, we successfully identified an optimal *k* value, performed data cleansing procedures to enhance accuracy, and developed a comprehensive understanding of the underlying objective of constructing a precise model.

This study conducted an extensive analysis on the kNN algorithm and several techniques in statistical learning, resulting in remarkable outcomes. However, future research aims to investigate alternative distance measures beyond the conventional Euclidean Distance in order to evaluate their impact on the accuracy of the classification model. By maintaining a control experiment with the Euclidean Distance as the baseline factor, valuable insights can be obtained regarding the effectiveness of common distance metrics in one, two, and multi-dimensional spaces. By imposing a measure that encompasses all coordinates of a point, distance measures exhibit versatility, enabling their application in diverse planes and spaces. Consequently, data analysis and comparison across different dimensions become feasible[15]. This potential future study possesses the potential for an intriguing comparison, exploring whether the achieved accuracy of 95.57% can be further improved.

References

[1] Preventing cancer. (n.d.). World Health Organization (WHO). Retrieved July 10, 2023, from 'WHO | Breast cancer', WHO. http://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/ (accessed Feb. 18, 2020).

[2] Rafid, A. K. M. R. H., Azam, S., Montaha, S., Karim, A., Fahim, K. U., & Hasan, M. Z. (2022, November 11). An Effective Ensemble Machine Learning Approach to Classify Breast Cancer Based on Feature Selection and Lesion Segmentation Using Preprocessed Mammograms. NCBI. Retrieved July 11, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9687739/

[3] Abdulla, S. H., Sagheer, A. M., & Veisi, H. (2021, August 14). 1979Breast Cancer Classification Using Machine Learning Techniques: A Review. View of Breast Cancer Classification Using Machine Learning Techniques: A Review. Retrieved July 10, 2023, from Abdulla, S. H., Sagheer, A. M., & Veisi, H. (2021, August 19). Breast Cancer Classification Using Machine Learning Techniques: A Review. urkish Journal of Computer and Mathematics Education. Retrieved June 29, 2023, from https://turcomat.org/index.php/turkbilmat/article/view/10604/8162

[4] Ehsani1, R., & Drabløs, F. (2020, September 19). Robust Distance Measures for kNN Classification of Cancer Data. Cancer Informatics. Retrieved July 10, 2023, from Ehsani, R., & Drabløs, F. (2020, September 19). Robust Distance Measures for kNN Classification of Cancer Data. Cancer Informatics. Retrieved June 30, 2023, from https://journals.sagepub.com/doi/pdf/10.1177/1176935120965542

[5] Bolandraftar, M., & Imandoust, S. B. (2017, December 7). Application of K-nearest neighbor (KNN) approach for predicting economic events theoretical background. ResearchGate. Retrieved July 10, 2023, from Imandoust, S. B., & Bolandraftar, M. (2013). Application of K-nearest neighbor (KNN) approach for predicting ... International Journal of Engineering Research and Applications. https://www.researchgate.net/profile/Mohammad-Bolandraftar/publication/304826093_Application_of_K-nearest_neighbor_KNN_approach_for_predicting_economic_events_theoretical_background/links/5a296efba6fdccfbbf816edf/Application-of-K-nearest-neighbor-KNN-approach-for-predicting-economic-events-theoretical-background.pdf

[6] Wettschereck, D., Aha, D. W., & Mohri, T. (n.d). A Review and Empirical Evaluation of Feature Weighting Methods for a Class of Lazy Learning Algorithms. Citeseerx. Retrieved July 10, 2023, from https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=5675f05a2e10e436218a0432678cb0416e606306

[7] Ajanki, A. (2007, May 28). File:KnnClassification.svg. Wikimedia Commons. Retrieved July 11, 2023, from https://commons.wikimedia.org/wiki/File:KnnClassification.svg

[8] Li, Y., & Zhang, X. (2011). Improving k Nearest Neighbor with Exemplar Generalization for Imbalanced Classification. Springer Link. Retrieved July 10, 2023, from https://link.springer.com/chapter/10.1007/978-3-642-20847-8_27

[9] 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. Retrieved July 10, 2023, from 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. https://doi.org/10.1016/j.procs.2016.04.224

[10] Kharya, S. (2015). BREAST CANCER DIAGNOSIS AND RECURRENCE PREDICTION USING MACHINE LEARNING TECHNIQUES. IJRET. Retrieved July 10, 2023, from https://ijret.org/volumes/2015v04/i04/IJRET20150404066.pdf

[11] Shah, C., & Jivani's, A. G. (2015, July 22). (PDF) Comparison of data mining classification algorithms for breast cancer prediction. ResearchGate. Retrieved July 10, 2023, from https://www.researchgate.net/publication/269270867_Comparison_of_data_mining_classification_algorithms_for_breast_cancer_prediction

[12] Amrane, M., Oukid, S., Gagaoua, I., & Ensarİ, T. (2018). Breast cancer classification using machine learning. IEEE Xplore. Retrieved July 10, 2023, from M. Amrane, S. Oukid, I. Gagaoua and T. Ensarİ, "Breast cancer classification using machine learning," 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), Istanbul, Turkey, 2018, pp. 1-4, doi: 10.1109/EBBT.2018.8391453.

[13] Tembusai, Z. R., Mawengkang, H., & Zarlis, M. (2021, January 11). K-Nearest Neighbor with K-Fold Cross Validation and Analytic Hierarchy Process on Data Classification | International Journal of Advances in Data and Information Systems. ijadis. Retrieved July 10, 2023, from http://www.ijadis.org/index.php/IJADIS/article/view/k-nearest-neighbor-with-k-fold-cross-validation-and-analytic-hie

[14] Machine Learning, U. (2016, September 25). Breast Cancer Wisconsin (Diagnostic) Data Set. Kaggle. Retrieved July 10, 2023, from Learning, U. M. (2016, September 25). Breast cancer wisconsin (diagnostic) data set. Kaggle. https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

[15] Alfeilat, H. A., Hassanat, A. B. A., Lasassmeh, O., Tarawneh, A. S., Alhasanat, M. B., Salman, H. S. E., & Prasath, V. B. S. (2019, December 7). Effects of Distance Measure Choice on K-Nearest Neighbor Classifier Performance: A Review. PubMed. Retrieved July 10, 2023, from Lewis, H. G., & Brown, M. (2010, November 25). A generalized confusion matrix for assessing area estimates from remotely sensed data. Taylor & Francis Online. Retrieved July 10, 2023, from https://www.tandfonline.com/doi/epdf/10.1080/01431160152558332?needAccess=true

[16] Lewis, H. G., & Brown, M. (2010, November 25). A generalized confusion matrix for assessing area estimates from remotely sensed data. Taylor & Francis Online. Retrieved July 10, 2023, from https://www.tandfonline.com/doi/epdf/10.1080/01431160152558332?needAccess=true

[17] n.d. (n.d.). Margin of Error - Definition, Usage, and Calculator. Zoho. Retrieved July 11, 2023, from https://www.zoho.com/survey/margin-of-error.html

[18] Henderi, H., Wahyuningsih, T., & Rahwanto, E. (2021, March 1). 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 | Henderi. International Journal of Informatics and Information Systems. Retrieved July 10, 2023, from http://ijiis.org/index.php/IJIIS/article/view/73

[19] Wong, T. T., & Yeh, P. Y. (2020, August 1). Reliable Accuracy Estimates from k-Fold Cross Validation. Research NCKU. Retrieved July 11, 2023, from https://researchoutput.ncku.edu.tw/en/publications/reliable-accuracy-estimates-from-k-fold-cross-validation

[20] James, G., Witten, D., Hastie, T., & Tibshirani, R. (n.d, n.d n.d). Corrected 7th Printing. Squarespace. Retrieved July 28, 2023, from https://static1.squarespace.com/static/5ff2adbe3fe4fe33db902812/t/6062a083acbfe82c7195b27d/1617076404560/ISLR%2BSeventh%2BPrinting.pdf