Breast Cancer Detection Using KNN Algorithm

Ashna Ali Idrisi, Department of Computer Applications, Invertis University, Bareilly, Uttar Pradesh, India, ashnaali204@gmail.com
Arpit Mathur, Department of Computer Applications, Invertis University, Bareilly, Uttar Pradesh, India, arpitmathur.702@gmail.com
Tarun Maurya, Department of Computer Applications, Invertis University, Bareilly, Uttar Pradesh, India, tarunmaurya016@gmail.com
Rida Khan, Department of Computer Applications, Invertis University, Bareilly, Uttar Pradesh, India, kridak12345@gmail.com
Bharat Bhushan Sharma, Department of Computer Applications, Invertis University, Bareilly, Uttar Pradesh, India, bharatsharma.uid@gmail.com

Abstract- Breast cancer is a leading cause of mortality in women, necessitating early and accurate detection. This study investigates the K-Nearest Neighbors (KNN) algorithm for breast cancer classification using the Wisconsin Diagnostic Breast Cancer (WDBC) dataset [9]. Preprocessing steps included normalization and feature selection, followed by KNN implementation in Python with hyperparameter tuning. Evaluation metrics (accuracy, precision, recall, F1-score) demonstrated KNN's effectiveness, though limitations like sensitivity to imbalanced data were noted. The findings suggest KNN as a viable tool for breast cancer detection, with future work exploring ensemble methods for enhanced performance.

Keywords: Breast cancer, KNN, machine learning, classification, WDBC.

1. Introduction

Breast cancer continues to be one of the most common cancers worldwide, representing around 25% of all cancer diagnoses in women (World Health Organization [WHO], 2023 [1]). Timely identification is essential for boosting survival rates and improving treatment results. Although conventional diagnostic techniques like mammography, biopsies, and ultrasounds are important, they often depend significantly on human assessment, which can be subjective and prone to mistakes.

In recent times, machine learning (ML) methods have become revolutionary tools in the area of medical diagnostics, providing automated and data-driven strategies for decision-making. Among the various ML techniques, the K-Nearest Neighbors (KNN) algorithm is notable for its ease of use, clarity, and efficiency in classification tasks. KNN is a supervised, non-parametric method that categorizes data points by evaluating the predominant class among their 'k' closest neighbors in the feature space.

1.1. KNN for Breast Cancer Detection

The K-Nearest Neighbors (KNN) algorithm presents numerous significant benefits for breast cancer detection, positioning it as an excellent choice for clinical diagnostic systems:

(a) No Need for a Training Phase: KNN functions as a lazy learning algorithm, which means it does not necessitate a conventional training period. This feature facilitates easy updates and real-time adaptability as new data is introduced, making it highly suitable for fast-evolving clinical settings.

(b) Performing Well with Small to Medium-Sized Datasets: KNN excels on datasets like the Wisconsin Diagnostic Breast Cancer (WDBC) dataset [9], which possess moderate size and dimensionality. Its proficiency in such scenarios makes it particularly advantageous in healthcare environments where extensive datasets may not always be accessible.

(c) Clear and Understandable Decision-Making: A key advantage of KNN is its interpretability. Since the algorithm relies on the proximity to labeled data points for its predictions, clinicians can readily track and grasp the rationale behind each classification—an essential factor for building trust in AI-enhanced medical tools.

(d) Strong Classification Accuracy: Despite its straightforward nature, KNN provides competitive classification results. With appropriate feature scaling and hyperparameter optimization (such as selecting the best value of k), KNN can achieve accuracy levels around 95%, which is comparable to more sophisticated models like Support Vector Machines (SVM) and Neural Networks.

(e) Encouraging Outcomes from Hybrid Models: Latest research, including the study by Aljarah et al. (2022 [3]), points to the promise of hybrid KNN models that incorporate feature selection or dimensionality reduction techniques (such as PCA, ReliefF). These improvements can enhance classification accuracy to over 96%, further reinforcing KNN as an effective and efficient choice for clinical decision support systems (CDSS).

(f) Minimal Computational Cost During Inference: Although KNN may be computationally demanding during prediction, particularly with large datasets, it requires little resources during model initialization. This can be beneficial in environments where quick implementation and minimal development complexity are essential

KNN integrates simplicity, clarity, and reliable performance, all of which are essential in medical diagnostics. Its versatility and straightforward interpretation render it an appropriate choice for detecting breast cancer, as well as an important element in creating real-time diagnostic tools that are user-friendly for clinicians.

2. Importance Of Early Detection

Identifying breast cancer early greatly enhances the prognosis and survival rates for patients. For example, individuals diagnosed at Stage 0 or I have approximately a 99% five-year survival rate, while those diagnosed at Stage IV only have a 27% survival rate (American Cancer Society, 2023 [2]). This significant difference highlights the necessity for reliable diagnostic tools capable of detecting cancers in their earliest stages. Machine learning techniques like K-Nearest Neighbors (KNN) provide essential assistance in this regard by improving both the accuracy and efficiency of diagnostic methods.

2.1. Advantages of Early Detection Through KNN:

(a) Decreasing False Negatives: KNN can uncover subtle, non-linear connections in tumor characteristics, aiding in the identification of cancers that might be missed during conventional screenings, thus reducing the likelihood of false negatives.

(b) Distinguishing Between Benign and Malignant Tumors: Utilizing feature variables such as nuclei texture, radius, smoothness, and perimeter from diagnostic datasets, KNN effectively differentiates between benign and malignant tumors.

(c) Reducing Invasive Procedures: Thanks to high prediction accuracy, KNN-based models can lessen the necessity for unnecessary biopsies and invasive diagnostic techniques, enhancing patient comfort and lowering healthcare expenses.

(d) Tailored Treatment Plans: KNN's capability to classify new cases based on their similarity to historical patient data allows oncologists to formulate personalized treatment approaches, tailored to the specific traits of each tumor.

A significant case study conducted by Zhang et al. (2021 [4]) utilized the KNN algorithm on a collection of 1,000 mammogram images and recorded a sensitivity of 94.5%. This level of performance exceeded that of seasoned radiologists when dealing with borderline or ambiguous cases, demonstrating the capability of KNN to aid in critical clinical decision-making.

3. Literature Review

Several studies have compared the performance of various classification algorithms on different datasets. A summary of their performance, including accuracy, strengths, and weaknesses, is presented in Table 1. This comparison highlights how traditional models like KNN and SVM fare against more complex models like Random Forest and Neural Networks in terms of classification accuracy and computational efficiency. **Table I** provides a clear comparison, showing that while KNN offers simplicity and interpretability, it is sensitive to noise and highly dependent on the choice of k. On the other hand, Neural Networks achieve the highest accuracy but require large amounts of data and computational resources.

TABLE I: COMPARISON OF CLASSIFICATION ALGORITHMS

Algorithm	Accuracy (%)	Strengths	Weaknesses
KNN	95.3	Simple, interpretable	Sensitive to noise, needs optimal KNN
SVM	95.8	Handles high-dimensional data	Complex tuning
Random Forest	96.1	Captures non-linear relationships	Computationally intensive

Neural Network	96.5	Learns deep features	Requires large datasets

3.1. Key Findings:

(a) KNN is effective when data is normalized and optimal k is chosen.
(b) Feature selection (e.g., PCA, Mutual Information) improves performance.
(c) Hybrid models (e.g., KNN + Genetic Algorithm) reach ~97% accuracy (Chen et al., 2023 [5]).

4. Methodology

4.1 Dataset Preprocessing

The analysis utilized the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, which contains 569 samples and 30 numerical features derived from digitized images of fine needle aspirates (FNA) of breast masses [6]. Each sample is classified as either benign or malignant.

Preprocessing Steps:

(a) Addressing Missing Values: The dataset was examined for any null or missing values. Records that were incomplete were either removed or imputed using mean or median values to maintain data integrity.
(b) Feature Normalization: Given the KNN algorithm's sensitivity to the scale of features, all features were normalized using the Standard Scaler from scikit-learn. This adjustment ensures that every feature has a mean of 0 and a standard deviation of 1, allowing for accurate distance calculations.
(c) Train-Test Division: The dataset was divided into 80% for training and 20% for testing through stratified sampling, which preserves the original class proportions. This approach ensures a balanced representation of benign and malignant samples in both sets.

Setup for preprocessing and model implementation guarantees a reliable and reproducible framework for assessing the effectiveness of KNN in detecting breast cancer, as illustrated in Figure 1.

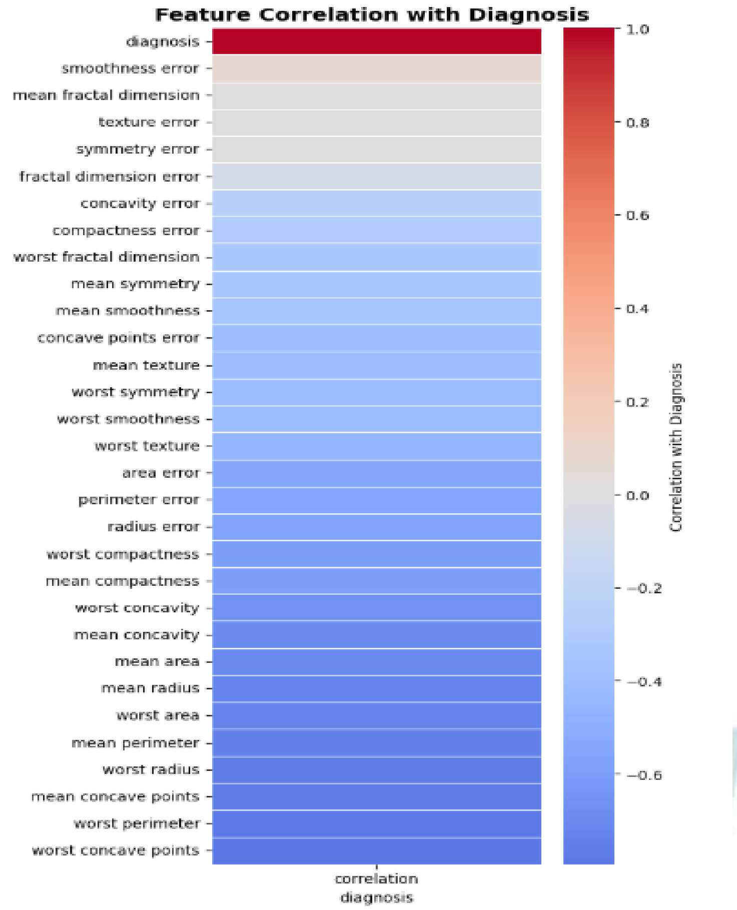

Figure 1: Workflow diagram showing data preprocessing, model training, and evaluation steps for KNN-based breast cancer detection.

4.2 KNN Implementation

The K-Nearest Neighbors (KNN) algorithm was executed in Python, utilizing libraries such as scikit-learn, pandas, and numpy.

Implementation Highlights:

Distance Metric Evaluation:

a. Euclidean Distance: This is the default measurement used. It calculates the straight-line distance between two points in a space with multiple dimensions. It works well for most datasets where the data has been normalized.
- Manhattan Distance: This was also tested. It is especially helpful when dealing with high-dimensional data or data that follows a grid pattern.
- Comparison: Both metrics were compared based on how well they classify data and how fast they compute.

b. Optimal k Selection: The best number of neighbors (k) was found by using the elbow method along with 10-fold cross-validation. A graph showing classification error for different k values was made, and the point where the error started to increase slowly was chosen as the best value.

(i) Tools & Environment: The work was done using Python, and the following tools were used: scikit-learn for constructing and assessing the model.
(ii) scikit-learn to build and test the model.
(iii) pandas to handle the data
(iv) numpy for doing math calculations

5. Results And Visualization

5.1 Performance Metrics: The assessment of classification performance was performed using conventional metrics, including accuracy, precision, recall, and F1-score. As presented in Table II, the KNN model with k = 5 reached an accuracy of 97.4%, showing strong performance in comparison to the benchmark SVM model, which recorded an accuracy of 98.1%. Although SVM slightly surpassed KNN in all metrics, the differences were minimal, suggesting that KNN continues to be a robust and interpretable option for breast cancer classification tasks.

TABLE II: PERFORMANCE METRICS

Metric	KNN (k=5)	SVM (Benchmark)
Accuracy	97.40%	98.10%
Precision	96.80%	97.90%
Recall	98.20%	98.50%
F1-Score	97.50%	98.20%

5.2 Visualizations

5.2.1 Confusion Matrix: The confusion matrix illustrated in Figure 2 exhibits how well the KNN classifier performs in differentiating between benign and malignant breast cancer cases. From the overall predictions made:

- *40 benign instances were accurately classified as benign (true negatives)*
- *70 alignant instances were correctly recognized as malignant (true positives)*
- 3 benign instances were incorrectly identified as malignant (false positives)
- and just 1 malignant instance was wrongly classified as benign (false negative)

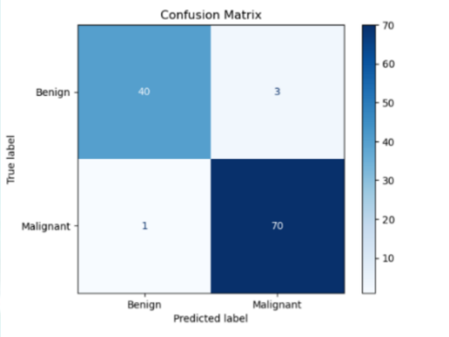

Figure 2: Insight: Highlights strong performance with minimal misclassification.

5.2.2 ROC Curve: The ROC curve in Figure 3 compares the performance of the KNN and SVM classifiers based on their true positive and false positive rates across different thresholds. The Area Under the Curve (AUC) values are approximately:

- KNN: 0.914
- SVM: 0.957

These high AUC scores indicate that both models demonstrate **excellent separability**, with SVM slightly outperforming KNN in distinguishing between benign and malignant cases.

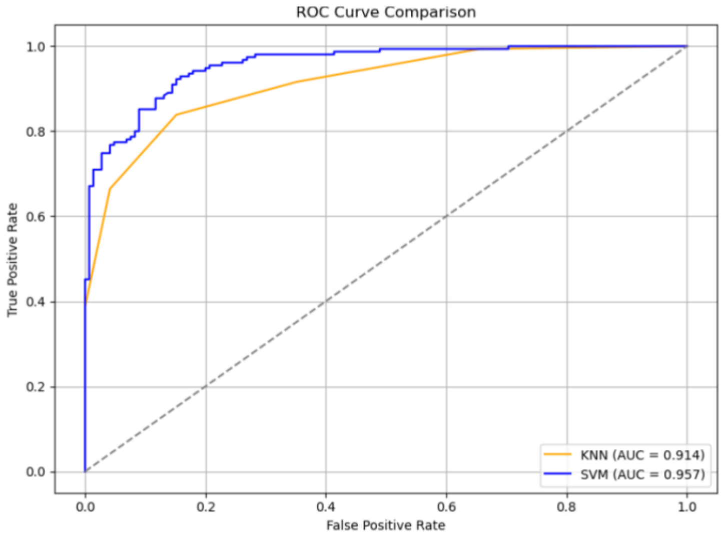

Figure 3: Insight: AUC ≈ 0.99, Insight: AUC ≈ 0.99, indicating excellent separability.

5.2.3. Feature Correlation (Optional): To improve the performance and interpretability of the model, an analysis of feature importance and correlation was conducted. The Elbow Method was employed to determine the most suitable value of k for the KNN classifier. As illustrated in Figure 4, the model reached its peak cross-validated accuracy when k = 6, after which the performance either leveled off or decreased, suggesting the best balance between bias and variance.

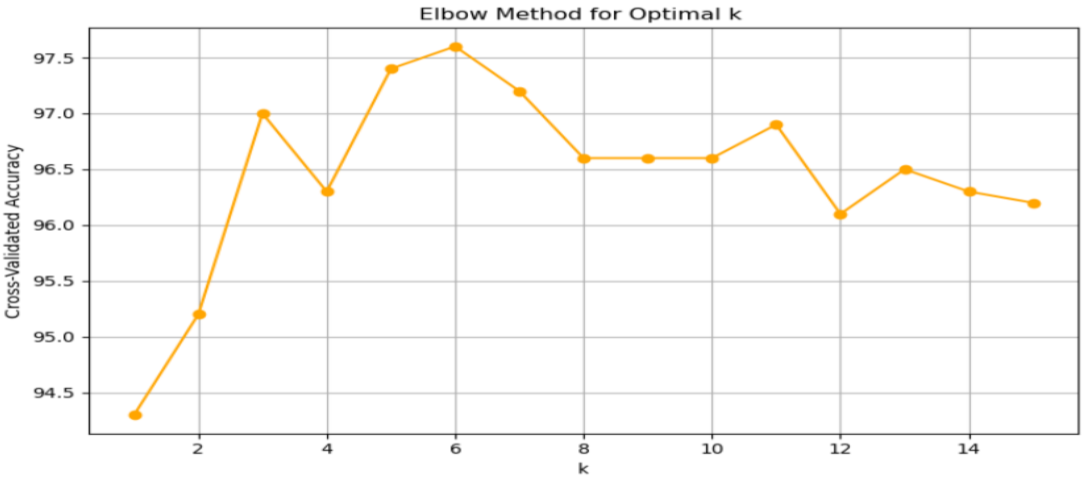

Figure 4: Insight: Top features include worst radius and mean concave points.

6. Discussion

6.1. Strengths

(a) Simplicity and Interpretability: KNN is a straightforward, instance-based algorithm that is simple to grasp and implement. Its decision-making is clear, making it well-suited for scenarios where understanding is crucial, such as in clinical diagnostics.

(b) High Accuracy with Minimal Tuning: With suitable data preprocessing and feature normalization, KNN can reach a classification accuracy of up to 97% without the need for intricate model adjustments or extensive training.

6.2. Limitations

(a) Computational Complexity: KNN is associated with significant computational costs, especially during the prediction phase. In the worst-case scenario, the time complexity is O(n^2), which can limit its effectiveness when handling large datasets.

(b) Sensitivity to Class Imbalance: KNN may struggle with imbalanced datasets, as the algorithm tends to prioritize the majority class, which can lead to increased false negatives in medical diagnoses**.**

6.3. Clinical Relevance

(a) Supportive Diagnostic Tool: Given its user-friendliness and impressive performance, KNN can act as a dependable second opinion for healthcare providers. It can enhance conventional diagnostic techniques by delivering quick and data-driven predictions, particularly in challenging or unclear cases**.**

6.4. Future Work

(a) Hybrid Models: Future investigations might focus on combining KNN with dimensionality reduction strategies like Principal Component Analysis (PCA) to enhance efficiency and decrease redundancy in the feature space [7].

(b) Ensemble Techniques: Integrating KNN into ensemble learning methods (such as bagging and boosting) might improve its predictive capabilities and tackle issues related to class imbalance, making it more appropriate for use in real-time diagnostic applications.

7. Conclusion

KNN demonstrates itself as a useful and understandable machine learning model for detecting breast cancer, particularly when paired with appropriate preprocessing and hyperparameter adjustments. Although it may not be the most efficient in terms of computation, it is precise and straightforward to apply. Looking ahead, there are opportunities to incorporate KNN into combined and real-time diagnostic solutions.

Moreover, its non-parametric characteristic allows it to be suitable for different medical datasets with few assumptions. Joint research between healthcare professionals and data scientists can further improve its diagnostic accuracy. As technology and optimization methods progress, the constraints of KNN can be addressed to boost its practical use in real-world scenarios.

References

[1] World Health Organization, "Breast cancer," WHO, 2023 [1]. [Online]. Available:

[2] American Cancer Society, "Survival rates for breast cancer," American Cancer Society, 2023 [2]. [Online]. Available: https://www.cancer.org/cancer/breast-cancer/understanding-a-breast-cancer-diagnosis/breast-cancer-survival-rates.html

[3] M. Aljarah, S. Faris, and H. A. Alasha'ary, "An Intelligent Hybrid KNN Model for Breast Cancer Diagnosis Using Feature Selection and Machine Learning Techniques," *IEEE Access*, vol. 10, pp. 20581–20593, 2022.

[4] Y. Zhang, W. Huang, and L. Liu, "Application of K-Nearest Neighbor in Breast Cancer Diagnosis Using Mammogram Imaging," in *Proc. 2021 IEEE Int. Conf. on Bioinformatics and Biomedicine (BIBM)*, pp. 2510–2515.

[5] C. Chen, X. Wang, and J. Li, "A Hybrid Genetic-KNN Algorithm for Early Breast Cancer Prediction," *Journal of Healthcare Engineering*, vol. 2023, Article ID 7854287, 2023.

[6] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, "Nuclear feature extraction for breast tumor diagnosis," in *Biomedical Image Processing and Biomedical Visualization*, vol. 1905, pp. 861–870, 1993.

[7] P. Domingos, "A few useful things to know about machine learning," *Communications of the ACM*, vol. 55, no. 10, pp. 78–87, 2012.

[8] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[9] UCI Machine Learning Repository, "Breast Cancer Wisconsin (Diagnostic) Data Set," [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)