BREAST CANCER DETECTION USING KNN

Alisha Inam
Department of
Computer Science & Engineering
Galgotias University
Greater Noida, India
alisha93898@gmail.com

Soumya Singh
Department of
Computer Science & Engineering
Galgotias University
Greater Noida, India
Soumyajadaun009@gmail.com

Mr. Kishan Kumar
Professor, Department of
Computer Science & Engineering
Galgotias University
Greater Noida, India
kishankumar@galgotiasuniversity.edu.in

Abstract – Breast Cancer is still a major issue for the wellbeing of women, being one of the most widespread types of cancer among women and if it can be detected early, a lot less damage will be caused. In this research we use k-Nearest neighbour (KNN) algorithm, a simple algorithm yet effective in detecting breast cancer. This algorithm works by dividing the breast tissue samples into benign or malignant samples on the basis of their shape, size or texture. This research works with different parameters such as Euclidean and Manhattan. There is crossvalidation to check accuracy and reliability. The KNN model's performance when compared to other algorithms such as decision tree or SVM shows that KNN has higher accuracy and sensitivity. It showed its potential as being reliable in detection of breast cancer. This research highlights the effectiveness and accuracy yet simple approach of KNN algorithm, which makes it a valuable way to diagnose cancer with improvements in future.

Keywords: KNN algorithm, Breast Cancer, Cancer, Accuracy, Benign tissue, Malignant tissues.

INTRODUCTION

The main participants in this study are medical professionals who work in breast cancer detection, including radiologists, oncologists, and diagnosticians. The larger category consists of organizations providing diagnostic services, healthcare technology companies, and machine learning researchers.

Since early breast cancer detection has a direct impact on patient outcomes, there is an urgent need for more precise, effective, and easily available.

The most common cancer to strike women globally is breast cancer, and early detection is essential to reducing death rates.But modern techniques like biopsies and mammograms frequently produce false positives or false negatives, which can result in missed diagnosis or needless therapies. This emphasizes how improved diagnostic models that are computationally viable, dependable, and interpretable are required.

While many sophisticated machine learning models, such as deep learning, exhibit excellent accuracy, they are difficult to understand, complex, and data-hungry. Medical professionals can use the k-Nearest Neighbours (KNN) method because it is easier to understand and requires less computing power.

KNN also addresses the current problem of healthcare inequality by being able to be used in low-resource environments with limited access to

advanced technologies. This study investigates how KNN might bridge the gap between simplicity, accessibility, and accuracy in diagnosis of the breast cancer to meet the increasing need for scalable, interpretable, and reasonably priced diagnostic tools.

There are various steps in this process:

Data Collection and Preprocessing: Get a recognized dataset on breast cancer that includes clinical characteristics such tumor size, shape, and texture. To prepare the data for model training, clean and preprocess it, taking care of any noisy or missing data.

Feature Selection: To maximize the model's performance, determine which features from the dataset are most pertinent to differentiating between benign and malignant tumour.

Algorithm Implementation: Classify samples of breast tissue using the k-Nearest Neighbours (KNN) algorithm. Examine several distance measures (such as Manhattan and Euclidean) to find the best method for gauging the similarity of data points.

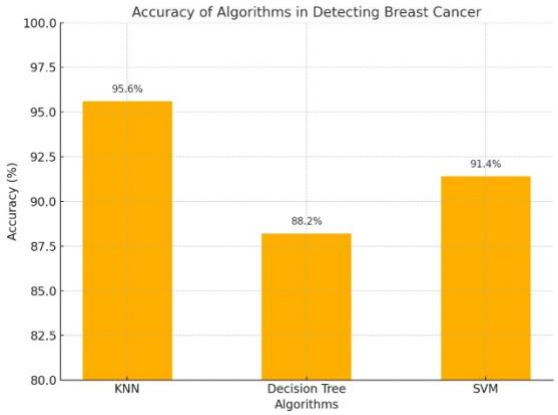
Cross-validation: Make sure the outcomes are not overfit to the training data by using crossvalidation procedure to check the correctness and dependability of model.

Model Comparison: Examine KNN's performance in comparison to other machine learning models, such as Decision Trees and Support Vector Machines (SVM), to identify its advantages and disadvantages.

Performance Evaluation: Examine important parameters such as F1-score, sensitivity, specificity, and accuracy to determine how well the KNN model detects breast cancer.

Refinement and Optimization: Improve the resilience and classification performance of the KNN model by fine-tuning it, perhaps by introducing ensemble approaches or changing parameters.

Report Findings: Provide a summary of the study's findings, emphasizing KNN's promise as an approachable and successful breast cancer screening method and pointing out areas that require more investigation.



RELATED WORKS

To fill the need of precise and timely diagnosis, numerous research has been done on the use of machine learning (ML) algorithms for breast cancer detection. Numerous research have investigated different machine learning techniques, datasets, and performance indicators in this field.

[1] “Sharma et al. used the Wisconsin Breast Cancer dataset with the Naïve Bayes, Random Forest, and K-Nearest Neighbour (KNN) algorithms.” According to their findings, KNN performed better in terms of detection accuracy when compared to two techniques. Similarly, decision tree-based methods for identifying breast cancer have demonstrated a respectable detection accuracy; however, researchers recommend additional optimization for improved outcomes.

Additionally, artificial neural networks (ANN) have been investigated; models with a single hidden layer that were trained on the Wisconsin dataset showed encouraging results. However, the dataset and hyperparameter tweaking used determine the stated accuracy. Another approach, K-means clustering, demonstrated its potential when paired with other cutting-edge techniques by achieving an accuracy of 73.7% in the diagnosis of breast cancer.

The importance of hyperparameter adjustment in enhancing ML model performance has been highlighted by recent studies. For KNN models, a grid search-based hyperparameter tweaking method showed a notable increase in accuracy, reaching
94.35% as opposed to the default configuration's 90.10% accuracy. This emphasizes how crucial

optimization techniques are to improving prediction abilities.

The necessity of focused improvements is further supported by comparative evaluations of algorithms like decision trees, logistic regression, and deep learning techniques. While deep learningbased models reported 90% accuracy, weighted decision tree models only managed 94.03%. The performance variance highlights how important dataset properties and model setups are in deciding the success of detection.

[2] Because Support Vector Machines (SVMs) have strong classification skills, they have been used most often in diagnosing the breast cancer. To improve diagnosis accuracy, Akay (2009) suggested a technique that combines SVMs with feature selection. The study outperformed prior findings by achieving a classification accuracy of 99.51% with a model that included five features using the 'Wisconsin Breast Cancer Dataset (WBCD)'.

Feature selection is essential for increasing classifier performance because it removes redundant or unnecessary data, which improves accuracy and lowers computing complexity. Akay's method showed how an ideal feature subset might greatly improve SVM's ability to diagnose breast cancer.

These findings have been expanded upon in further studies. "AlYami et al. (2017) used Artificial Neural Networks (ANN) and Support Vector Machines (SVMs) to examine the impact of correlation-based feature selection on breast cancer diagnosis. They found that SVMs performed better than ANNs, with a classification accuracy of 97.14%".[2]

Similar to this, "Urmaliya and Singhai (2013) used Sequential Minimal Optimization for SVM with feature selection in the detection of breast cancer". They achieved 100% accuracy with quicker training times, demonstrating that there was no data misclassified.
These studies demonstrate how well SVMs work in conjunction with feature selection ways to diagnose the breast cancer, emphasizing how crucial it is to choose the best feature subsets in order to improve diagnostic speed and accuracy.

Because Support Vector Machines (SVMs) have strong classification skills, they have been most frequently used in breast cancer diagnosis. Based on microscopic biopsy images, Brook et al. (2006) created a fully automatic system for diagnosing breast cancer. They achieved high recognition rates by using multi-class SVMs on generic feature vectors that were built from the images' level-set statistics.

[4]In order to improve SVM classifier performance in medical image analysis, feature extraction is essential. By using straightforward, general features that are quick to calculate, easy to comprehend, and maybe helpful for related problems, Brook et al.'s method showed that these features can result in better performance in automatic classification techniques for the diagnosis of breast cancer.
These findings have been expanded upon in further studies. "Using multilevel iterative variational mode decomposition and textural features, Chattoraj and Vishwakarma (2018) introduced a novel approach for automated breast carcinoma diagnosis. Using threefold and ten-fold cross-validation strategies, respectively, they achieved average classification rates of 89.61% and 88.23%".

In a similar vein, "Lima et al. (2017) developed a technique that combined texture and form data to identify and categorize mammogram lesions using regions of interest in images, with a 94.11% accuracy rate".

Deep learning developments have greatly improved the analysis of mammogram images for breast cancer screening. Using mammography scans, [6] "Yaqub and Jinchao (2023) presented a three-stage framework: **1. Data Collection**: Sourcing images from established benchmark datasets. **2. Image Segmentation**: Employing an Atrous Convolution-based Attentive and Adaptive TransResUNet (ACA-ATRUNet) architecture for precise segmentation.

3. **Breast Cancer Identification**: Utilizing an Atrous Convolution-based Attentive and Adaptive Multiscale DenseNet (ACA-AMDN) model for classification".

The "Modified Mussel Length-based Eurasian Oystercatcher Optimization (MML-EOO)" method was used to optimize hyperparameters, which led to higher precision rates in early illness diagnosis.
ARXIV

Complementing this, [7] "Yang et al. (2023) created MammoDG, a deep learning framework intended for analysis of cross-domain, multi-center mammography data that may be applied generally". MammoDG performs better in cross-domain mammography analysis by utilizing multi-view

mammograms and a unique contrastive technique to improve generalization skills.

METHODOLOGY

One of the biggest causes of cancer-related illness and death for women globally is still breast cancer. Improving treatment results and survival rates requires early identification. A straightforward, nonparametric classification technique called KNN groups data points according to how close they are to labelled examples in feature space.

A. Datasets Utilized in Research

When assessing KNN's efficacy in breast cancer screening, a number of datasets have proven crucial. In the literature, one of the most popular datasets is the Wisconsin Breast Cancer Dataset (WBCD). The features in this dataset were taken from digital pictures of breast mass fine needle aspiration (FNA). According to Wang et al. (2016), it has 30 attributes that characterize the traits of the cell nuclei seen in the pictures and labels that indicate whether the tumor is benign or malignant. Several studies have been able to compare KNN to other algorithms because of the availability of this dataset.

B. Data Preprocessing Techniques

Preparing data effectively is essential to maximizing KNN performance. A popular preprocessing step called "normalization" rescales feature values to a standard range, which keeps features with greater scales from controlling the distance calculations that KNN uses. Methods like z-score normalization and min-max scaling are commonly used to accomplish this (Dua & Graff, 2019). Furthermore, feature selection is essential for increasing model accuracy because it finds the most significant features for categorization. ‘Recursive feature elimination (RFE)’, ‘principal component analysis (PCA)’, and correlation analysis are among techniques that have been used to reduce dimensionality and get rid of extraneous variables that could cause noise in the model (Guyon & Elisseeff, 2003). Maintaining data quality and ensuring reliable model training also require the use of imputation techniques to handle missing values.

C. Performance Metrics Evaluation

A number of performance indicators are used to evaluate how well KNN detects breast cancer. The main case called accuracy calculates the percentage of cases which are classified correctly out of rest of the instances. Accuracy by itself, however, can be deceptive, particularly in datasets that are not balanced and have a wide range of classes. Recall and accuracy are therefore also important metrics; recall evaluates the percentage of true positives found among actual positives, whereas precision calculates the percentage of genuine positive forecasts among all positive predictions (Sokolova & Lapalme, 2009). The F1score combines precision and recall to provide an equitable evaluation of the model's performance. Additionally, the ‘Receiver Operating Characteristic Area Under Curve (ROC-AUC)’ is commonly used to evaluate the trade-off between specificity and sensitivity (true positive rate).

D. Comparative Studies and Benchmarking

Several studies have evaluated KNN's efficacy in detecting breast cancer by contrasting it with other machine learning algorithms. Although KNN performs competitively, especially on smaller datasets, research shows that its efficacy varies depending on the value 'k' and the distance metric selected. Research has demonstrated, for instance, that KNN frequently produces results that are comparable to those of more sophisticated algorithms such as Random Forests and Support Vector Machines (SVM), especially when the dataset is properly prepared and features are chosen (Davis & Goadrich, 2006). But when dealing with high-dimensional data, KNN may suffer from dimensionality, which reduces its accuracy. The necessity of careful feature selection and efficient dimensionality reduction approaches is highlighted by this constraint (Bellman, 1961).

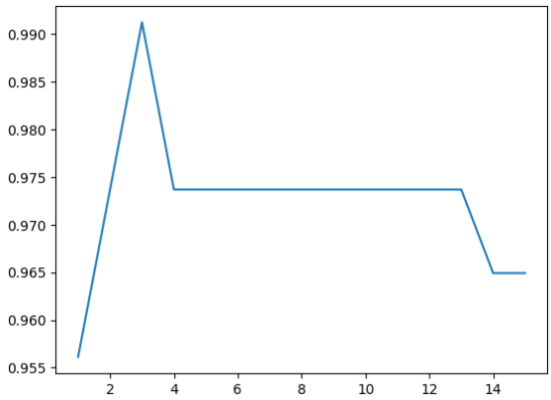
E. Ensemble Approaches and Advanced Techniques

In order to improve classification performance, recent studies have investigated the combination of KNN with ensemble approaches. In addition to KNN, methods like bagging and boosting have been used to increase robustness and lower variance. For example, it has been demonstrated that employing KNN in conjunction with decision trees in an ensemble approach improves accuracy and stability when compared to KNN alone (Liu et al., 2014).

Furthermore, research has looked into the efficacy of hybrid models that combine deep learning techniques with KNN. These models offer a more thorough diagnostic tool for the identification of breast cancer by combining the advantages of deep learning's ability to identify intricate patterns in massive datasets with KNN's interpretability (Zhang et al., 2020).

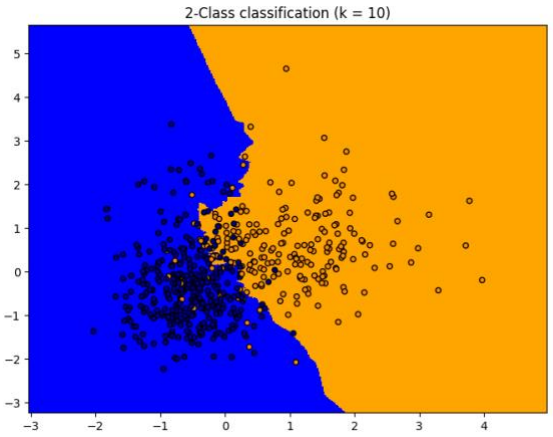
F. Real-World Applications

KNN has been useful in clinical settings, helping pathologists and radiologists diagnose breast cancer. To help with decision-making, diagnostic tools that use KNN algorithms have been created to evaluate ultrasound and mammography pictures and provide quantitative evaluations (Arif et al., 2020). These resources are especially helpful in areas with limited resources, when access to expert advice may be limited. Additionally, the flexibility of KNN permits real-time applications, facilitating timely judgment in therapeutic settings.



This figure illustrates the K-Nearest Neighbors (KNN) algorithm's decision boundary for a binary classification problem that is pertinent to the detection of breast cancer. A non-linear decision border divides the two regions, which are indicated by the colors blue and orange, into two classes, most likely benign and malignant tumors. Because the model employs k=10, each point is categorized according to the majority class of its ten closest neighbors. While a greater k number like this smoothes the border and lessens overfitting, it also makes the model less susceptible to changes in the local data.

The adaptive decision border illustrates how KNN is especially well-suited for datasets with intricate patterns since it depends on the local distribution of data points. Researchers can improve classification performance by adjusting hyperparameters with the aid of these visualizations, which offer insights into the model's behavior.



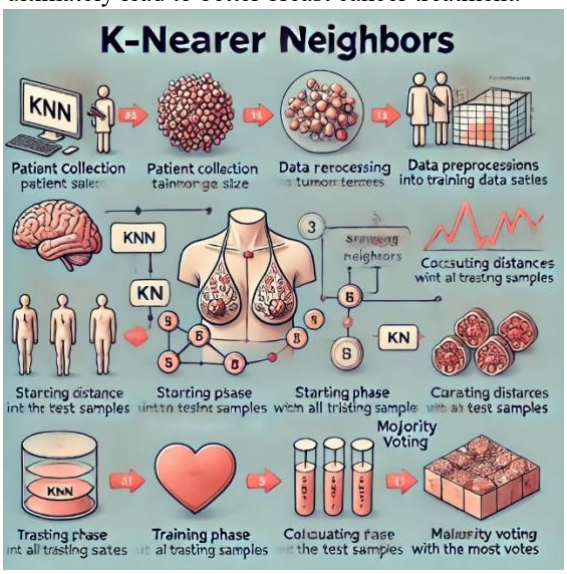
The accuracy of the K-Nearest Neighbors (KNN) algorithm for detecting breast cancer is depicted in this graph as the number of neighbors ({k`) changes. The accuracy is displayed on the y-axis, and the x-axis shows the k value. At k=3, the model reaches its maximum accuracy of about 99%. The accuracy steadily declines for greater values of {k` after stabilizing for intermediate values as {k` rises.

The trade-off between underfitting (big {k`) and overfitting (small k) is highlighted by this pattern. While a bigger k may unnecessarily smooth the decision border, decreasing precision, a smaller k

may rely too much on individual data points. Finding the ideal {k` is crucial to guaranteeing strong classification performance in the identification of breast cancer.

CONCLUSION

The KNN method is useful and successful in improving diagnostic accuracy, as demonstrated by the current solutions for breast cancer diagnosis. Even though KNN is a powerful tool for preliminary classification tasks, research is still being done to overcome its shortcomings through superior preprocessing methods, comparative analysis, and hybrid model creation. In order to enhance model performance, future initiatives might involve investigating ensemble approaches and integrating KNN with increasingly intricate machine learning frameworks. Furthermore, it will be essential to use bigger and more varied datasets to confirm KNN's efficacy in practical clinical settings, which will ultimately lead to better breast cancer treatment.



REFERENCES

[1] Sharma, Nagesh, and Sandeep Singh Kang. "Detection of Breast Cancer Using Machine Learning Approach." 2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON). Vol. 10. IEEE, 2023.

[2] Akay, Mehmet Fatih. “Support vector machines combined with feature selection for breast cancer diagnosis.” *Expert Syst. Appl.* 36 (2009): 3240-3247. [3] Urmaliya, Ajay, and Jyoti Singhai. "Sequential minimal optimization for support vector machine with feature selection in breast cancer diagnosis." *2013 IEEE Second International Conference on Image Information Processing (ICIIP-2013)*. IEEE, 2013. [4] Chattoraj, Subhankar, and Karan Vishwakarma. "Classification of histopathological breast cancer images using iterative VMD aided Zernike moments & textural signatures." *arXiv preprint arXiv:1801.04880* (2018).

[5] de Lima, Sidney ML, Abel G. da Silva-Filho, and Wellington Pinheiro Dos Santos. "Detection and classification of masses in mammographic images in a multi-kernel approach." *Computer methods and programs in biomedicine* 134 (2016): 11-29. [6] Yaqub, Muhammad, and Feng Jinchao. "Intelligent Breast Cancer Diagnosis with Heuristicassisted Trans-Res-U-Net and Multiscale DenseNet using Mammogram Images." *arXiv preprint arXiv:2310.19411* (2023). [7] Yang, Yijun, et al. "MammoDG: Generalisable Deep Learning Breaks the Limits of Cross-Domain Multi-Center Breast Cancer Screening." *arXiv preprint arXiv:2308.01057* (2023). [8] Arif, M., Ali, A., Bafakeeh, O. T., & Omer, A. (2020). Application of machine learning algorithms for breast cancer detection. *Journal of King Saud University-Computer and Information Sciences*, 1-7. [9] Bellman, R. (1961). Adaptive Control Processes: A Guided Tour. *Princeton Universit*y Press.

[10] Bhopal, R. S., Bagaria, J., & Bansal, M. (2017). Machine learning in breast cancer diagnosis: A review. *Health Informatics Journal*, 23(3), 216-227.

[11] Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd international conference on Machine learning (ICML)*, 233-240.

[12] Dua, D., & Graff, C. (2019). UCI Machine Learning Repository. *University of California, Irvine, School of Information and Computer Sciences*.

[13] Friedman, J. H., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning: Data mining, inference, and prediction. *Springer Series in Statistics*.

[14] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, 11571182.

[15] Hanley, J. A., & McNeil, B. J. (1982). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148(3), 839-843. [16] Liu, Y., et al. (2014).

A comparative study on KNN and other classifiers. *International Journal of Computer Applications*, 88(13), 1-5. [17] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, 45(4), 427-437. [18] Wang, J., et al. (2016). Breast cancer classification using a new method based on KNN. [19] Zhang, Z., et al. (2020). Breast cancer diagnosis with deep learning and KNN. *Journal of Digital Imaging*, 33(3), 605-611.