Predicting Breast Cancer Diagnosis Based on Tumor Characteristics and Features Using a Multi-Faceted Algorithmic Approach: K-Means Clustering, Decision Trees, and Random Forests.

Elena Ajayi

St. John's University

Abstract

This study examines the power of machine learning algorithms such as K-means clustering, random forests, and decision trees and their ability to enhance breast cancer diagnosis and treatment. We utilized the Wisconsin Diagnostic Breast Cancer (WDBC) dataset (Al Aswad, 2023), containing 569 breast tissue samples with 32 features related to cell characteristics and biopsy images that reflect tumor diagnostics: Benign or Malignant. Data preprocessing, including handling missing and undetermined values and feature engineering, such as min-max normalization and the construction of rations, were performed to optimize data exploration. K-means clustering revealed three breast cancer clusters with unique clinical and texture features. Subsequently, a random forest model trained on these engineered devices achieved a 93.4 % accuracy in predicting benign or malignant tumors. The interpretability of decision trees within the random forest model introduced important diagnostic channels highlighting which features were associated with each tumor type. The complex nature of breast cancer necessitates a multifaceted approach wherein the strength of diverse machine learning tools synergize to produce a comprehensive diagnostic and prognostic tool. As a result, integrating clustering, classification, and interpretable models enables a nuanced understanding of this heterogeneous disease. Our findings underscore the potential of these machine learning algorithms to identify different subtypes, improve diagnostic accuracy, and guide personalized treatment plans and decisions, ultimately advancing clinical care and the field of breast cancer research.

Keywords: Breast cancer, machine learning, K-means clustering, random forests, decision trees, WDBC dataset, personalized treatment, feature engineering, tumor diagnosis

Predicting Breast Cancer Diagnosis Based on Tumor Characteristics and Features Using a Multi-Faceted Algorithmic Approach: K-Means Clustering, Decision Trees, and Random Forests.

Literature Review

In data mining, machine learning methods such as K-means and SVM have been widely used to analyze healthcare information, with breast cancer being a prime example. For instance, the metabolic stratification of human breast tumors study investigated tumor subtypes' clinical and therapeutic relevance by identifying robust metabolic signatures based on samples from seven previously published studies (Iqbal, 2023). This involved differentiating between 134 primary cancers in three independent validation sets and mapping out the metabolic environment of breast cancers through experimental inhibition of metabolically related pathways (Iqbal, 2023). The study found that cell lines representing different metabolic subtypes showed distinct sensitivities to specific treatments, suggesting the potential for tailored and combination therapies.

In another study, titled *Applications of Support Vector Machine (SVM) Learning in Cancer Genomics (Huang et al., 2021)*, SVM and other machine learning algorithms were integrated to classify gene expression microarray data from two breast cancer subtypes (Huang et al., 2021). Although the s training set was limited to 38 patients, the study demonstrated SVM's superior ability to categorize gene features, illustrating the potential of machine learning in genomic analysis.

Overall, machine learning algorithms are powerful tools to analyze complex biological data, such as gene expression metabolic profiles or patterns, and uncover hidden patterns or relations that can inform diagnosis, prognosis, and treatment decisions in breast cancer.

Some key points of difference that make this study worthwhile are its multifaceted approach, feature engineering, and consideration of the aspect of texture features. Introducing biopsy image data and clinical features shifts the paradigm towards emerging viewpoints and methods. Other precursors likely lead to tumor outbreaks being revealed by this approach. Overall, the examination of subtypes, together with a prediction of diagnosis, provides a comprehensive assessment for classifying breast cancer tumors.

Data

This dataset was taken from Wisconsin Diagnostic Breast Cancer (WDBC). The dataset includes 569 breast tissue samples with all its nuclei-derived features in digitized images. The dataset provided ten mean values, standard errors, and extreme values for the cell characteristics. In particular, the diagnosis column classified a tumor as benign or malignant based on its characteristics.

We also enriched the data with newly generated features through ratio calculations. We calculated ratios between pairs of existing features, such as radius_mean / radius_se, to describe the possible correlation or relationship not easily captured by the raw data. In total, we developed 11 new features for feature engineering, expanding the possibilities of exploring subtle patterns in the data.

- Radius_ratio: captures the relationship between a mean and standard error of radius, which can expose cell size differences, which may indicate cancer.
- Perimeter_area_ratio: The perimeter_area_ratio is a measure of the irregular shape of the tumor. Higher ratios mean more complex and possibly cancerous growth patterns.
- Concavity_ratio: Aggressive Tumors have uneven margins.
- Texture_worst_mean_ratio and texture_mean_se_ratio: These ratios show the

heterogeneity of tumor texture; higher values indicate more significant asymmetry and possible malignancy.

- Symmetry_mean_se_ratio, symmetry_worst_mean_ratio, and symmetry_asymmetry_ratio: Symmetry of cell distribution inside a tumor is examined employing these indicators. Malignant tumors are usually asymmetrically developed.
- Fractal_dimension_mean_se_ratio and fractal_dimension_worst_mean_ratio: Using the fractal dimension mean se ratio and fractal dimension worst mean ratio to measure the complexity of shape at different scales for the tumor. High scores may reflect malignant evolutionary trends.
- Fractal_dimension_complexity_difference: Highlights the difference between the tumor's most and least complex regions, representing areas of rapid growth.

This feature engineering process, followed by applying K-means clustering and random forests, allowed for a comprehensive dataset exploration and revealed hidden patterns and relationships that could be more apparent when using a single algorithm alone.

From the practical and biological relevance angle, these features can give the individual more information than absolute values because they reflect differences and trends among tumor characteristics, ultimately crucial for comprehending the heterogeneous nature of breast cancer and tailoring personalized treatment plans. Furthermore, with the help of different tumor medical dimension standards, the selected feature gauges a comprehensive range from symmetry to texture and shape. Together, these features provide a detailed analysis of the overall characteristics of the tumor. The morphological and textural features identified in this study align with well-established indicators of malignancy observed in breast cancer, such as asymmetric growth, heterogeneous texture, and irregular shape, as highlighted in Beck et al.'s (2011)

systematic analysis of breast cancer morphology.

Methodology

Clustering Analysis

K-means clustering was employed to understand the structural properties more comprehensively. Three different combinations of features are chosen for analysis: (a) radius_ratio and perimeter_area_ratio, (b) concavity_ratio and radius_ratio, and (c) texture_worst_mean_ratio and texture_mean_se_ratio. The optimal number of clusters was determined using a combined approach of the elbow method and silhouette analysis to ensure a robust statistical strategy.

Figure 1

Silhouette Analysis for Determining Optimal Number of Clusters

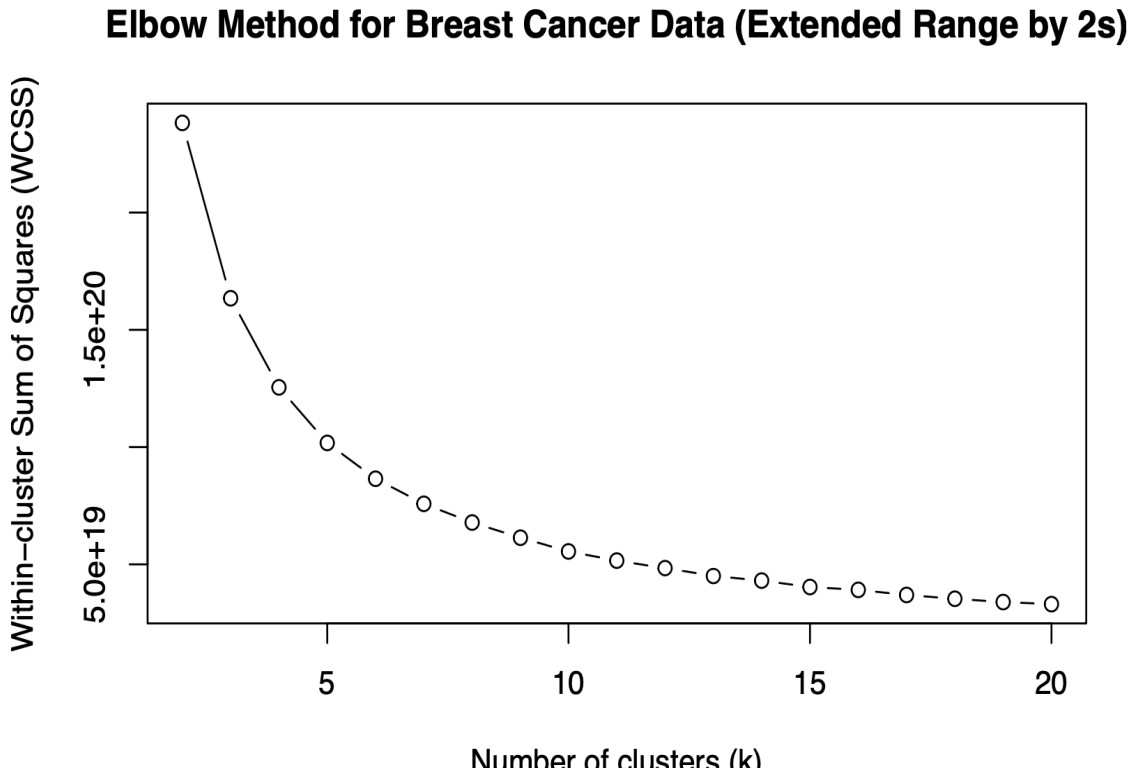

Figure 1 depicts the silhouette scores against the k-value used for clustering in k-means clustering. As shown in the figure, the maximum value of the silhouette score is approximately 0.55 with three clusters, indicating better separation and coherence among the cluster partitions, hence likely representing the different stages of breast cancer or subtypes. With higher k-values, the clustering yielded decreasing silhouette scores, supporting this choice of three clusters as optimal.

Figure 2

Scatter Plot of Radius Ratio vs Perimeter Area Ratio, Colored by Cluster and Shaped by Diagnosis

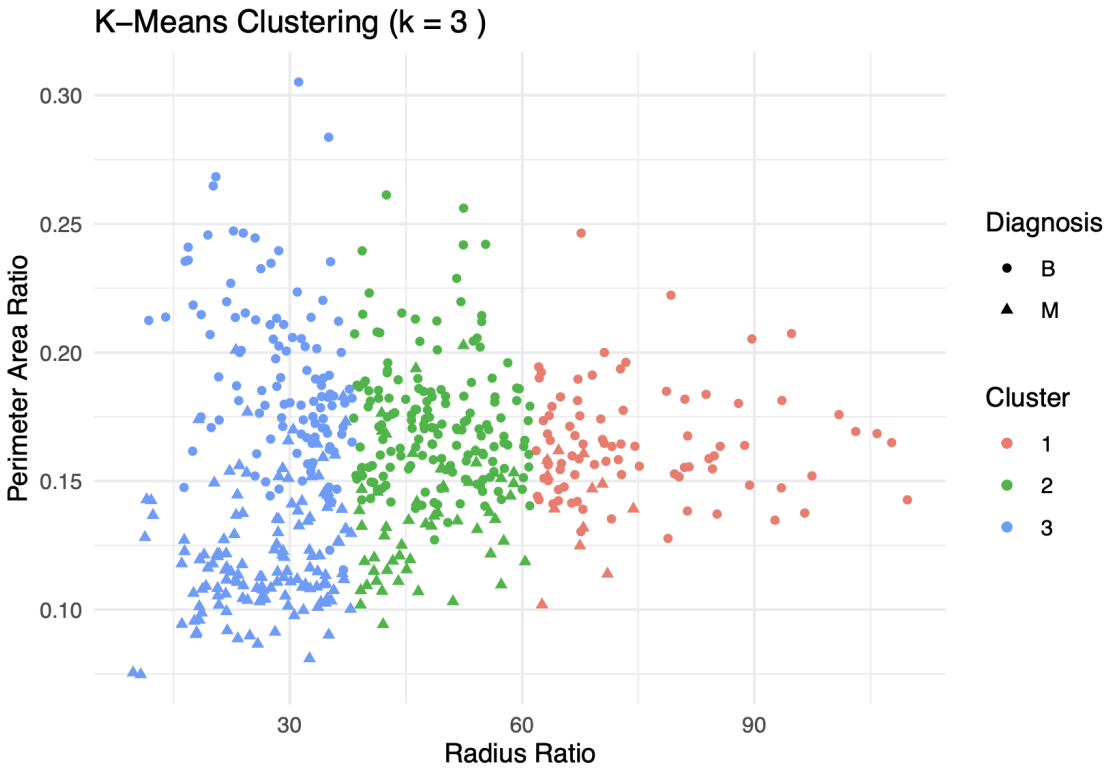

Figure 2 shows the distribution of samples of breast tissue using radius ratio and perimeter area ratio, after clustering with the k-means algorithm. Each point represents a sample,

colored by its assigned cluster, of which there are three, and shaped by its actual diagnosis: circle for benign, triangle for malignant. The separation among the clusters is reasonable, as most of the malignant cases are mainly in the lower cluster, while the benign cases are more scattered. Observed dispersion in the distribution of benign cases suggests that there might be some limitations of the k-means algorithm in this context, and future studies could try to use more robust clustering methods for the separation of groups. This dispersion could be improved by the application of more robust clustering algorithms, such as K-medoids or K-means++, which are more resilient against outliers and which usually have better centroid initialization.

Diagnosis Prediction

Two well-known supervised learning algorithms were integrated into predict diagnoses (Benign vs. Malignant):

a. Decision tree: Using the R CTree package, a decision tree model was trained on the designed characteristics (Zhang, 2019). This understandable approach illuminates vital features that influence tumor diagnosis.
b. Random forest: The randomForest package was also used to create a random forest model, an ensemble of decision trees renowned for its accuracy and robustness (Liaw & Wiener, 2019). With a 93.4 % predictive accuracy on the test set, our model demonstrated its potential for accurate diagnosis prediction.

Subsequently, these models were paired for t-testing to assess statistical significance.

Results

Regarding evaluation and statistical interpretation, the random forest model showed remarkable performance with a 93.41% accuracy rate in breast cancer diagnosis prediction. Feature engineering methods helped the model to extract more informative features from the

data. The robustness and better performance of random forest compared to simpler models may be attributed to an ensemble learning method. This technique combines the predictions of different decision trees to decrease the risk of overfitting and improve generalization.

A paired t-test was used to analytically evaluate the performance difference between the random forest and a baseline decision tree model. Based on this test, the random forest was found to have a statistically significant advantage (p-value = 0.01867). This suggests that the random forest model consistently beats the Decision Tree across various evaluation metrics, including accuracy, precision, recall, and F1-score, with a mean difference of -0.1152633, signifying an approximately 11.5 percentage point improvement.

Limitations

Despite the high accuracy of the random forest model for breast cancer diagnosis, the study has several limitations. The results might have low generalizability due to the specific dataset and model parameters applied in the study. It would be necessary to investigate the influence of multiple datasets and model configurations to validate and generalize the findings. Moreover, the median imputation strategy applied to address missing data is effective; however, it may introduce bias if missingness is not entirely random.

Discussion

Notwithstanding these limitations, the high accuracy of the random forest model and its statistically significant improvement over the baseline model emphasizes its potential as a valuable tool for diagnosing breast cancer. Its potential for further optimization through advanced feature engineering and hyperparameter tuning suggests it could become even more powerful and reliable.

In this regard, future research should validate the model's performance across diverse populations, explore additional feature engineering techniques, and optimize hyperparameters to maximize its predictive capabilities. These efforts could lead to earlier detection, more personalized treatment planning, and improved patient outcomes.

Conclusion

The accuracy of the random forest model in predicting breast cancer diagnosis is a testament to the impact of machine learning in the healthcare sector. Using algorithms to analyze large amounts of medical data can help create more effective, faster, and individual approaches to diagnostics. This could transform the way breast cancer is diagnosed and managed, leading to better detection, personalized treatment, and better prognosis.

However, as machine learning algorithms become more complex, it is crucial to solve the problem of the 'black box' – the inability to understand how the algorithm makes its decision. This lack of clarity can undermine confidence among patients and healthcare workers, thus slowing the use of these possibly lifesaving technologies that prioritize the development of interpretable machine learning models that provide clear and understandable explanations for their predictions. This would not only foster trust but also empower clinicians to gain a deeper understanding of the underlying factors contributing to a breast cancer diagnosis.

The use of artificial intelligence in healthcare is expected to deliver quality healthcare to the doorsteps of patients and clinicians through the use of data. By promoting transparency, fairness, and accountability in the development and implementation of machine learning models, we can create a roadmap for the integration of AI tools into clinical practice that will increase the accuracy of diagnoses, the effectiveness of treatments, and the overall quality of care, thus creating a better future for healthcare.

References

Al Aswad, M. (2023). Wisconsin Diagnostic Breast Cancer Dataset. Kaggle. https://www.kaggle.com/uciml/breast-cancer-wisconsin-data/kernels

Beck, A. H., Sangoi, A. R., Leung, S., Marinelli, R. J., Nielsen, T. O., van de Vijver, M. J., ... & Koller, D. (2011). Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Sci Transl Med*, 3(108), 108ra113. https://doi.org/10.1126/scitranslmed.3002564

Cao, X. H., Stojkovic, I., & Obradovic, Z. (2016). A robust data scaling algorithm to improve classification accuracies in biomedical data. *BMC Bioinformatics*, *17*(1). https://doi.org/10.1186/s12859-016-1236-x

Guo, L., Kong, D., Liu, J., Zhan, L., Luo, L., Zheng, W., Zheng, Q., Chen, C., & Sun, S. (2023). Breast cancer heterogeneity and its implication in personalized precision therapy. *Experimental Hematology & Oncology*, p. *12*. https://doi.org/10.1186/s40164-022-00363-1

Huang, S., Cai, N., Pacheco, P. P., Narandes, S., Wang, Y., & Xu, W. (2018). Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. *Cancer Genomics & Proteomics*, *15*(1), 41-51. https://doi.org/10.21873/cgp.20063

Iqbal, M. A., Siddiqui, S., Smith, K., Singh, P., Kumar, B., Chouaib, S., & Chandrasekaran, S. (2023). Metabolic stratification of human breast tumors reveals subtypes of clinical and therapeutic relevance. *IScience*, *26*(10). https://doi.org/10.1016/j.isci.2023.108059

Jäger, S., Allhorn, A., & Bießmann, F. (2021). A Benchmark for Data Imputation Methods. *Frontiers in Big Data*, *4*, 693674. https://doi.org/10.3389/fdata.2021.693674

Liaw & Wiener (2018). Liaw A, Wiener M. Classification and regression by random Forest. R package Version 4.6-14 https://cran.r-project.org/package=randomForest

Zhang, Z. (2016). Decision tree modeling using R. *Annals of Translational Medicine*, *4*(15). https://doi.org/10.21037/atm.2016.05.14