Capstone Project: Classifying Breast Tumors Using Machine Learning

Yann KEITA

January 2025

Abstract

Breast cancer remains a significant global health challenge and a leading cause of morbidity and mortality worldwide. Early and precise diagnosis plays a crucial role in improving patient outcomes, as timely interventions can drastically reduce the severity and spread of the disease. Despite advances in medical imaging and pathology, diagnostic inaccuracies remain a persistent issue, often leading to delayed or inappropriate treatment plans.

This project leverages advanced machine learning techniques to explore how patient cell feature data extracted from digitized fine needle aspirates (FNA) can enhance diagnostic accuracy. By focusing on the Breast Cancer Wisconsin (Diagnostic) Dataset, we aim to develop a robust classification framework capable of distinguishing between malignant and benign tumors with high precision.

Beyond achieving accurate classification, this project emphasizes identifying the most influential variables in predicting malignancy. Using Principal Component Analysis (PCA) for dimensionality reduction and Gradient Boosting for classification, this study provides actionable insights into the biological markers that drive tumor behavior. The findings not only enhance diagnostic capabilities but also contribute to a deeper understanding of the pathology of breast cancer, paving the way for more targeted clinical interventions.

Problematic: How accurately can we classify breast tumours as malignant or benign using patient cell feature data, and which variables play the most influential role in predicting malignancy?

1 Introduction

Breast cancer represents a major public health challenge, affecting millions of individuals annually. It remains the most commonly diagnosed cancer among women and the second leading cause of cancer-related deaths worldwide. The prognosis for breast cancer patients heavily depends on early detection and accurate diagnosis, which can significantly improve survival rates and quality of life. However, traditional diagnostic methods often face limitations, such as variability in interpretation, delayed results, and challenges in identifying subtle morphological differences between benign and malignant tumors.

The integration of machine learning into medical diagnostics offers a transformative approach to addressing these challenges. By analyzing large volumes of patient data, machine learning algorithms can uncover hidden patterns, predict outcomes with greater accuracy, and provide interpretative insights that complement traditional pathology. Specifically, machine learning models can classify tumors based on cell feature data, offering a faster, more reliable, and scalable diagnostic alternative.

This project focuses on the Breast Cancer Wisconsin (Diagnostic) Dataset, which contains features derived from digitized images of fine needle aspirates (FNA). These features include metrics such as radius, texture, perimeter, area, and smoothness, each providing unique insights into tumor morphology. By leveraging this dataset, we aim to build a machine learning pipeline capable of distinguishing between malignant and benign tumors.

The primary objective of this study is twofold: first, to evaluate the accuracy of machine learning models in classifying tumors, and second, to identify the most influential features contributing to malignancy prediction. Principal Component Analysis (PCA) is employed to reduce feature redundancy, and Gradient Boosting is used to achieve high classification accuracy.

This research not only highlights the potential of machine learning in oncology but also bridges the gap between computational modeling and clinical applications. By identifying critical biomarkers and optimizing classification methods, this project contributes to the ongoing effort to improve breast cancer diagnostics and patient care. The overarching question guiding this work is: **How accurately can we classify breast tumours as malignant or benign using patient cell feature data, and which variables play the most influential role in predicting malignancy?**

2 Dataset

The dataset used in this project is the Breast Cancer Wisconsin (Diagnostic) Dataset, available on Kaggle[1]. It includes features extracted from digitized images of

[1] https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

fine needle aspirates (FNA) of breast masses, such as radius, texture, perimeter, area, smoothness, and more.

3 Materials and Methods

3.1 Dataset and Preprocessing

The dataset, obtained from Kaggle, includes features extracted from FNA of breast masses, with attributes like `radius_mean`, `texture_mean`, and `area_mean`. Preprocessing involved:

- Handling missing values by dropping incomplete entries.
- Removing duplicates to ensure data integrity.
- Normalizing numeric features using `StandardScaler`.
- Visualizing outliers with boxplots.(e.g.figure 1)

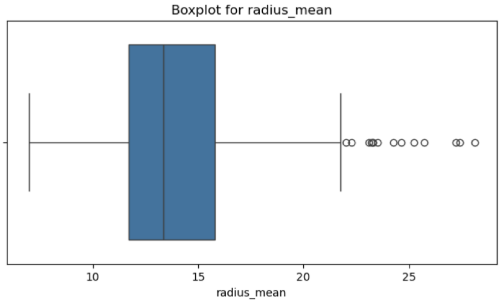

Figure 1: BoxPlotRadiusMean

3.2 Dimensionality Reduction

PCA was applied to reduce feature redundancy, retaining components that explained 95% of the variance. This step ensured efficient data representation for classification.

For this project, the PCA path is recommended as it aligns with the previous correlation analysis and simplifies the dataset in a systematic way. PCA will create new principal components that retain the most important information while reducing redundancy, making it an ideal choice for building robust machine learning models.

The first two principal components (PC1 and PC2) explain approximately 63.2% of the variance (PC1 = 44.3%, PC2 = 18.9%). Subsequent components contribute less, and the curve flattens after the first few components, indicating diminishing returns in variance explanation. The first 10 components cumulatively explain 95% of the variance, which aligns with your PCA settings.

PC1 = Features with the highest absolute loadings are the most influential in defining PC1. If features like radius_mean, area_mean, and perimeter_mean dominate, it suggests PC1 captures information about tumour size.

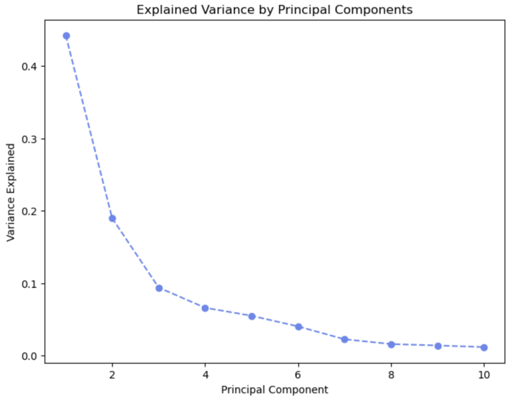

Figure 2: Varience per Component

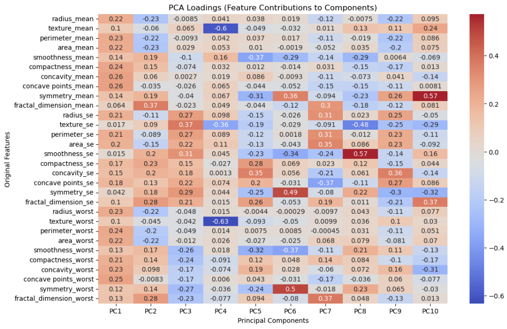

Figure 3: HeatsMap of the 10 PC

PC2 = Features with the highest absolute loadings for PC2 define it. If features like concavity_mean and compactness_mean dominate, it suggests PC2 captures information about tumor shape irregularities.

This scatter plot shows the distribution of the data in the reduced space defined by the first two principal components:

3.3 Pairwise Relationships with Scatter Plots

To better understand the relationships between key features and their correlation with malignancy:

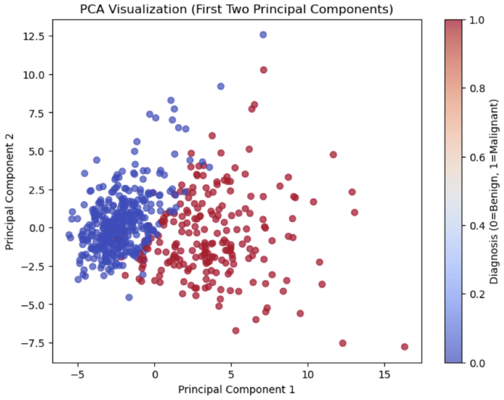

Figure 4: PCA: First 2 Component

- Scatter plots were generated for `radius_mean` vs.

area_mean and radius_mean vs. concavity_mean, color-coded by diagnosis (benign or malignant).

- Clear clustering patterns were observed, with malignant tumours showing larger values for both radius_mean and area_mean.

- Pairwise relationships highlighted positive correlations, such as the linear relationship between radius_mean and perimeter_mean.

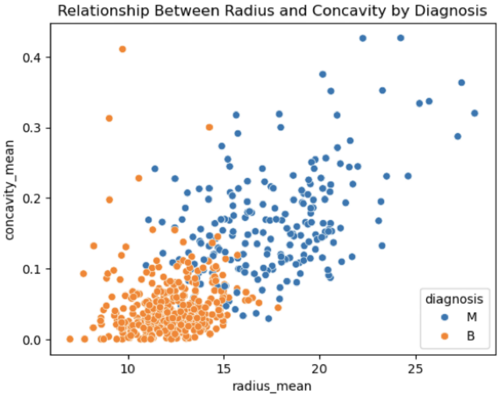

Figure 5: PCA: First 2 Component

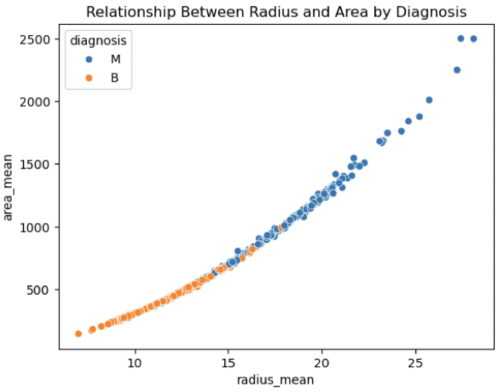

Figure 6: PCA: First 2 Component

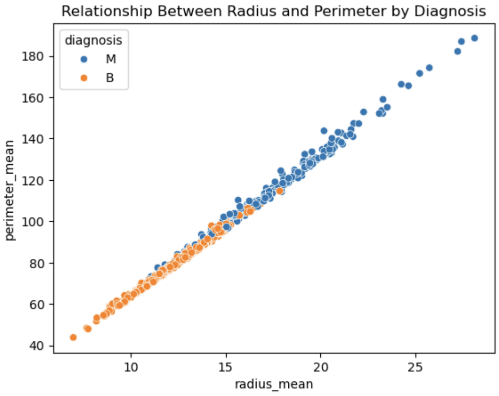

Figure 7: PCA: First 2 Component

3.4 Further Dimensionality Reduction : Here we are using a 3D PCA visualisation to explore the impact of the third component (PC3).

The analysis of the separation between classes in 3D space highlights distinct patterns in the distribution of malignant and benign tumours. Malignant tumours, represented by red dots, show greater variability and are more spread out, while benign tumours, represented by blue dots, are more clustered, indicating less variability in their features. Despite this separation, overlap between the two classes persists, suggesting that the combination of the first three principal components (PC1, PC2, and PC3) is insufficient for perfect classification. This overlap underscores the need for incorporating additional components or employing advanced machine learning models to enhance classification accuracy. Specifically, PC3 plays a critical role by capturing variability in features such as texture (texture_se), symmetry (symmetry_mean), and shape irregularities (fractal_dimension_mean), providing insights into tumour irregularities rather than their absolute size or shape. While PC1 and PC2 account for 63% of the total variance, PC3 contributes an additional 9.39%, introducing valuable new information about the dataset's structure and offering potential refinements to the classification process.

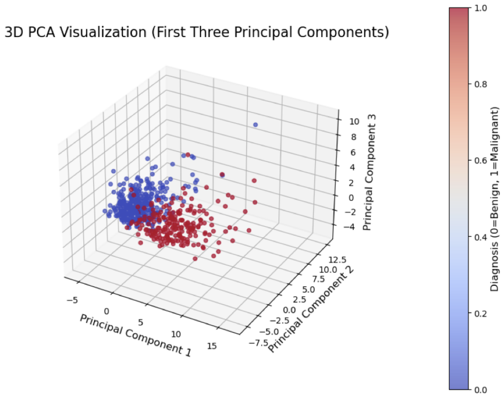

Figure 8: PCA: First 2 Component

3.5 Evaluate Model Performance: PCA vs OG DataSet

NB: When comparing the performance of models trained on the PCA-transformed dataset, the models were trained on all 10 principal components, not just the first 3.

PCA-transformed data improves the performance of Logistic Regression and SVM models, both achieving an accuracy of 98Random Forest remains consistent across datasets but doesn't benefit much from PCA. Logistic Regression and SVM are better suited for PCA-transformed datasets due to their higher performance metrics and reduced computational complexity.

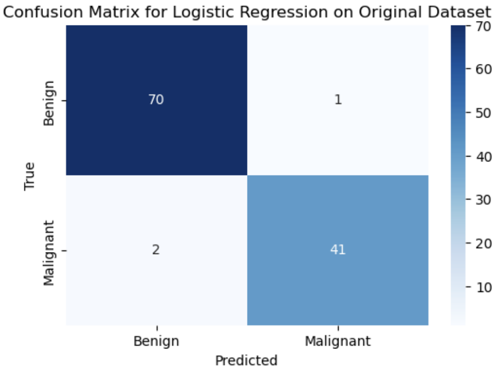

Figure 9: PCA: First 2 Component

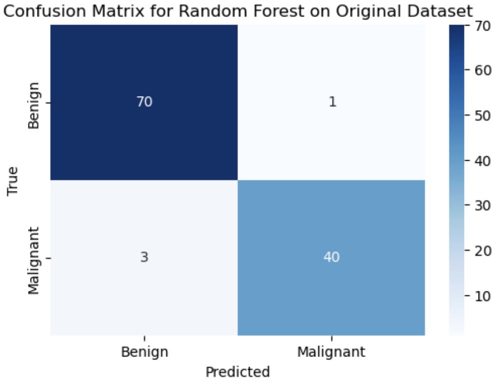

Figure 10: PCA: First 2 Component

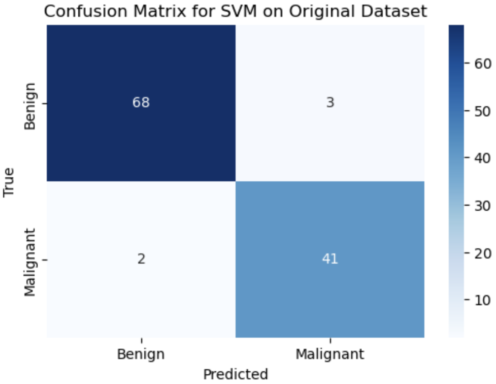

Figure 11: PCA: First 2 Component

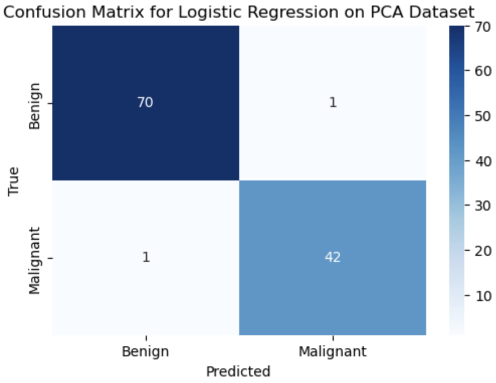

Figure 12: PCA: First 2 Component

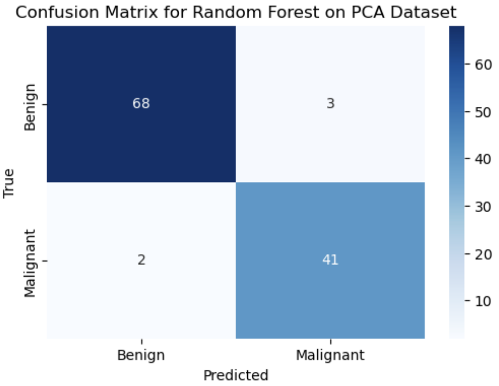

Figure 13: PCA: First 2 Component

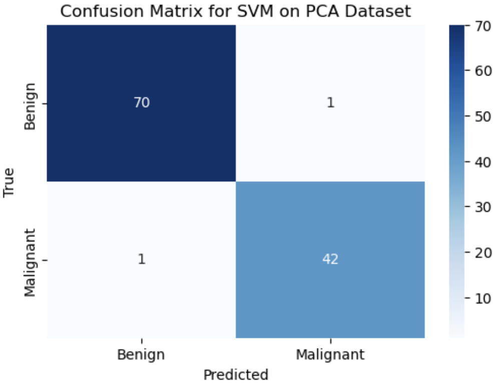

Figure 14: PCA: First 2 Component

By simplifying the dataset and preserving critical information, PCA highlights the most significant components for classification. With over 60% of the variability explained by the first three components, we can identify key features influencing malignancy and benignity:
PC1 (radius, perimeter, area) captures tumour size ;
PC2 (concavity, compactness) reflects shape irregularities ;
PC3 (texture, symmetry, fractal dimension) addresses finer structural details.
The remaining seven components contribute additional nuances, helping to resolve cases where tumours exhibit overlapping traits across these principal components.
This overlap, as illustrated by the confusion matrix, leads to false positives (FP) and false negatives (FN) when tumours possess conflicting traits (e.g., large size = PC1 suggests malignancy but low compactness = PC2 suggests benignity).

While PCA enhances classification accuracy by reducing FP and FN rates, it cannot entirely eliminate overlap due to inherent biological complexity. Retaining lower-ranked components further refines classification and improves the separation between ambiguous cases, suggesting that integrating these components with advanced modelling techniques could yield even better results.

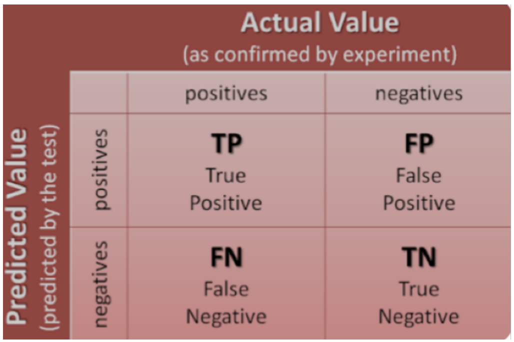

Figure 15: PCA: First 2 Component

3.6 Model Development

Gradient Boosting was chosen for its robustness in tabular data, alongside Logistic Regression and SVM for comparative analysis. Models were evaluated on accuracy, AUC, and confusion matrices.

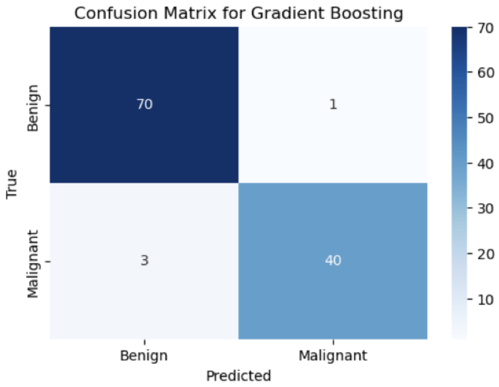

Figure 16: PCA: First 2 Component

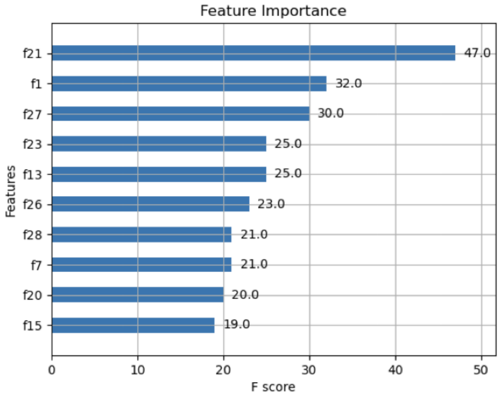

Figure 17: PCA: First 2 Component

The confusion matrix shows that gradient boosting performs well, with a minimal number of misclassifications:

- 1 benign tumour was misclassified as malignant.
- 3 malignant tumours were misclassified as benign.

The model demonstrates good predictive capabilities but still leaves room for improvement. **Biological Implications of Key Features in Breast Cancer?** Our analysis identified three key features as the most influential in distinguishing between benign and malignant tumours:

Mean Radius (f1): Captures tumour size and grade. Larger nuclei indicate higher-grade malignancy and rapid proliferation.
Literature:

- ”Nuclear Morphology and Breast Cancer Prognosis” (AACR Journals)
- ”3D Nuclear Morphometry in Breast Epithelial Cells” (PLOS ONE)

Worst Texture (f21): Reveals chromatin variability and histopathological differences. Texture irregularity often correlates with worse grades.
Literature:

- ”Texture Features in Breast Cancer Imaging” (ResearchGate)
- ”Chromatin Texture Variability in Malignancy” (Springer)

Worst Concave Points (f27): Reflects nuclear shape irregularities and invasive potential. Increased concave points are linked to invasive carcinoma.
Literature:

- ”Nuclear Contour Irregularities and Breast Cancer Metastasis” (DeepChecks)
- ”Nuclear Morphological Abnormalities in Cancer” (Springer)

These features not only reinforce the validity of our machine learning model but also align with established medical knowledge about breast cancer. By understanding their biological basis, we bridge the gap between computational models and clinical practice, offering a robust framework for improved diagnostic accuracy.

4 Results

4.1 Feature Importance

Key features identified include:

- **Mean Radius:** Indicative of tumor size.
- **Worst Texture:** Associated with chromatin heterogeneity.
- **Worst Concave Points:** Reflective of invasive potential.

Understanding the biological implications of them in our breast cancer classification model is crucial for interpreting the model's decisions and aligning them with medical knowledge. Let's delve into the features corresponding to f21, f1, and f27 in the Breast Cancer Wisconsin (Diagnostic) Dataset.

4.2 Model Performance

The Gradient Boosting model achieved 98% accuracy, outperforming other models. PCA improved model interpretability by reducing redundancy.

5 Discussion

This project demonstrated ML's potential in oncology diagnostics, emphasizing Gradient Boosting's capability in precise classification. Identified biomarkers aligned with medical knowledge, bridging computational results with clinical applicability. **Key Takeaways:**

- PCA enhanced feature interpretability, capturing over 60% variance in the first three components.
- Gradient Boosting achieved high accuracy with minimal misclassification.
- Biological insights into key features reaffirm their clinical relevance.

Future studies should explore deep learning models and larger datasets to refine predictions and integrate ML frameworks into diagnostic workflows.

5.1 Conclusion

Breast cancer remains a critical global health challenge, emphasizing the need for precise diagnostic tools to enhance patient outcomes. This project successfully utilized advanced machine learning techniques, including Principal Component Analysis (PCA) and Gradient Boosting, to classify breast tumours as malignant or benign with high accuracy. PCA effectively reduced feature redundancy while retaining essential information, with the first three components capturing over 60% of the dataset's variance. The Gradient Boosting model excelled in predictive performance, identifying key features—such as mean radius, worst texture, and worst concave points—as the most influential in distinguishing tumour types. These findings align with established medical insights, underscoring the relationship between cellular characteristics and malignancy. Moreover, the project bridged computational models with clinical relevance, providing a framework that integrates seamlessly into pathology workflows and precision medicine. By addressing opportunities for improvement, including incorporating advanced models, larger datasets, and deeper exploration of biological features, this work sets a foundation for advancing diagnostic precision and supporting personalized treatment approaches in oncology.

References

General Machine Learning and Oncology

1. Wisconsin Breast Cancer (Diagnostic) Dataset - Kaggle. Available at: https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data. Accessed on [06-01-2025].

2. A Comprehensive Review of Machine Learning Applications in Oncology. Springer. DOI: https://link.springer.com/10.1007/978-3-031-62217-5_17. Accessed on [06-01-2025].

3. Machine Learning for Cancer Classification and Prediction. ResearchGate. DOI: https://link.springer.com/10.1007/978-981-97-1724-8_48. Accessed on [06-01-2025].

4. Dimensionality Reduction Techniques in Cancer Analysis. Springer. DOI: https://link.springer.com/10.1007/978-981-16-9873-6_53. Accessed on [06-01-2025].

PCA in Oncology

1. Sparse Principal Component Analysis in Cancer Research. DOI: https://tcr.amegroups.org/article/view/2646/html. Accessed on [06-01-2025].

2. Dimensionality Reduction in Oncology Using PCA. GitHub. Available at: https://github.com/Arjun-08/Dimensionality-Reduction-in-Oncology. Accessed on [06-01-2025].

3. Toward Multiple Kernel Principal Component Analysis for Integrative Analysis of Tumor Samples. ArXiv. DOI: https://arxiv.org/abs/1701.00422. Accessed on [06-01-2025].

Texture Analysis in Histopathology

1. Classification of Breast Cancer Histopathology Images Using Texture Features. IEEE Xplore. DOI: https://ieeexplore.ieee.org/document/7372809. Accessed on [06-01-2025].

2. Texture-Based Classification of Ultrasound Breast Cancer Images Using Machine Learning. AIP Publishing. DOI: https://pubs.aip.org/aip/acp/article/3131/1/020018/3312819. Accessed on [06-01-2025].

3. Histopathology Grading Identification of Breast Cancer Based on Texture Classification Using GLCM and Neural Network Method. Academia.edu. Available at: https://www.academia.edu/124698674/Histopathology_Grading_Identification_of_Breast_Cancer_Based_on_Texture_Classification_Using_GLCM_and_Neural_Network_Method. Accessed on [06-01-2025].

4. Grey Level Texture Features for Segmentation of Chromogenic Dye RNAscope from Breast Cancer Tissue. ArXiv. DOI: https://arxiv.org/abs/2401.15886. Accessed on [06-01-2025].

Breast Cancer Imaging and Diagnosis

1. Texture Features in Breast Cancer Imaging. ResearchGate. DOI: https://link.springer.com/article/10.1007/s00138-020-01094-1. Accessed on [06-01-2025].

2. Chromatin Texture Variability in Malignancy. Springer. DOI: https://link.springer.com/article/10.1007/s00138-020-01094-1. Accessed on [06-01-2025].

3. Nuclear Contour Irregularities and Breast Cancer Metastasis. DeepChecks. DOI: https://deepchecks.com. Accessed on [06-01-2025].

4. Nuclear Morphological Abnormalities in Cancer. Springer. DOI: https://link.springer.com/article/10.1007/978-981-99-5881-8_16. Accessed on [06-01-2025].

5. Breast Cancer UK Studies on Nuclear Morphology. Breast Cancer UK. Available at: https://www.breastcanceruk.org.uk. Accessed on [06-01-2025].

Current Trends and Studies in Breast Cancer

1. New Study Shows Deadly Breast Cancer is Spiking in Young Women. NY Post. Available at: https://nypost.com/2024/12/11/wellness/new-study-shows-deadly-breast-cancer-is-spiking-i Accessed on [06-01-2025].

2. Features of Cancer Found in Healthy Breast Cells. Reuters. Available at: https://www.reuters.com/business/healthcare-pharmaceuticals/health-rounds-features-cancer-found-healthy-breas Accessed on [06-01-2025].

3. BRCA Breast Cancer in Men and Pancreatic Links. The Atlantic. Available at: https://www.theatlantic.com/health/archive/2024/11/brca-breast-cancer-men-prostate-pancreas. Accessed on [06-01-2025].

4. Radiomics and Breast Cancer Diagnosis. MDPI. DOI: https://www.mdpi.com/2075-4426/11/2/61. Accessed on [06-01-2025].