# **COMPREHENSIVE MORPHOLOGICAL ANALYSIS FOR BREAST CANCER DIAGNOSIS**

**Introduction:**

Breast cancer is one of the most common cancers seen in women after skin cancer and is a global health care concern. Diagnosis of this cancer into benign and malignant is crucial for improved patient outcomes and early detection is crucial for effective treatment. Breast Cancer can affect various parts of the breasts like mammary glands and ducts leading to lumps or masses. This cancer type can be seen both in men and women but predominantly seen in women.

**Objective:**

The objective of this paper is to study the data set and use different statistical analytical tools and machine learning techniques to predict the malignancy of breast cancer in this dataset. This study will provide the clinicians with precise and efficient means of identifying potential cancer cases.

● The objectives of this study include developing a predictive model

● Contributing to the ongoing efforts to improve the prediction of breast cancer

● Addressing the current diagnostic methods

**Motivation:**

The motivation for this study was mainly based on the need to improve existing breast cancer screening models especially in terms of accuracy and efficiency. Current diagnostic methods, although effective, may benefit from the integration of advanced machine learning techniques to provide nuanced understanding of tumor features and enhance predictive capabilities developed traditional diagnostic methods such as mammography and physical examination have become increasingly important in the diagnosis of breast cancer. However, these techniques sometimes present challenges to achieve optimal sensitivity and specificity leading to instances of false positives and false negatives. This study motivates to overcome these challenges and uses tools like machine learning to refine the methods of breast cancer detection. The complex nature of breast cancer requires careful evaluation of clinical, morphologic, and molecular features.

**Description of the specific problem and data :**

The specific problem addressed in this study is the accurate classification of breast tumors as malignant or benign. This data set contains a variety of features including clinical, morphological, and other mass characteristics like smoothness, radius, and texture of the mass and these patterns and data together can provide valuable insights. These features are extracted from fine needle aspiration methods of breast cancer biopsies providing a dataset for training and testing models. This program explores the challenges of breast cancer detection, examining relationships between variables and effect on classification accuracy. The complexity of this problem stems from breast tumor heterogeneity which can manifest in different ways in individuals. To address this challenge this study focuses on predictive models that can distinguish subtle patterns embedded in data by using advanced machine learning techniques. Clinical characteristics include patient demographics, medical history, and medical context. Physical features capture the structural properties of tumor cells while molecular features delve into genetics associated with each biopsy. This multidimensional dataset extracted from a variety of breast cancer biopsies provides a comprehensive picture of the disease allowing microscopic analysis in factors influencing tumor distribution.

**Methods of Data Analysis:**

Our approach of data analysis for this study focuses on machine learning algorithms like logistic regression, support vector machines (SVM) but are not limited to them. The reason behind these choices lies in the nonlinear relationships of the breast cancer dataset.

We will also be using tools and workflows like Orange and Stata

**Stata-**

Stata is a statistical software to be used for in depth statistical analysis and robustness of our model. It tests model performance and generates detailed statistical summaries of results.

We have built a Statistical Model with Stata using the data set and have chosen Logistic Regression model for our data as it is suitable for our data type. We have opted this as the outcome of my dataset is binary and logistic regression model is usually opted for binary or categorical datasets. My dependent variable or diagnosis in Breast cancer and in this context of data set 0 is “malignant” and 1 is “benign” so this model can estimate the probability of an individual having breast cancer.

Logistic regression

Number of obs = 569

LR chi2(7) = 655.74

Prob > chi2 = 0.0000

Log likelihood = -47.847653 Pseudo R2 = 0.8727

The LR chi-square test, which shows the presence of at least one predictor variable with the result associations and a value of 0.0000 and 655.74 suggests that the logistic regression model appears to be statistically significant. Higher values suggest a better match, and a log likelihood value of -47.847653 demonstrates how well the model predicts the observed data. With a pseudo-R-squared value of 0.8727, the model provides a strong overall fit and explains a significant percentage of the variation in the dependent variable.

**Orange Workflow-**

Orange workflow is a visual design tool that makes it easy to build and analyze machine learning models. It allows easy visual representation of data processing and model building steps this interface facilitates the implementation of various machine learning algorithms which aids in interactive development. The SVM model along with random forest and logistic regression exhibited very high accuracy, precision, recall, and F1 scores, suggesting excellent performance in classifying instances, especially in the context of identifying malignant cases in breast cancer diagnosis. The high precision implies that the model is reliable when it predicts a case as malignant, minimizing false positives. The high recall indicates that the model is effective in capturing a significant proportion of the actual malignant cases, demonstrating high sensitivity. However, while these models show outstanding performance, it's essential to consider potential overfitting, especially as this dataset used for evaluation is not very vast. Further validation on independent datasets and cross-validation techniques could help ensure the generalizability of the model.

Validating the model generalizability can be done by collecting an independent dataset, applying identical preprocessing as the original, predicting with the trained model, calculating consistent performance metrics (accuracy, precision, recall, F1 score), and comparing results to affirm robustness across different datasets. K-Fold Cross-Validation can be done by dividing the dataset into k folds, iteratively training and evaluating the model, calculating average performance, and addressing overfitting concerns.

Data visualization:

In our data analysis and machine learning exploration, we leveraged the versatility of Orange for insightful data visualization. Specifically, we employed scatter plots to visually depict the intricate relationships and distributions among diverse morphological features characterizing breast tumors. The scatter plots served as a powerful tool to uncover patterns and trends, enhancing our understanding of the dataset. This graphical representation not only facilitated model development but also provided a comprehensive overview of the complex interplay between variables.

**Results:**

The application of machine learning models to breast cancer database has provided insights that contribute to ongoing efforts to improve diagnostic accuracy.

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No** | **Feature** | **Mean (Malignant)** | **Standard Error (Malignant)** |
| **1.** | Radius | 12 to 23 | 0.2 to 1 |
| **2.** | Texture | 16 to 28 | 0.4 to 1.9 |
| **3.** | Perimeter | 77 to 144 | 1.4 to 8 |
| **4.** | Concavity | 0.02 to 0.25 | 0.012 to 0.1 |
| **5.** | Area | 430 to 1400 | 13 to 150 |
| **6.** | Smoothness | 0.8 to 0.1 | 0.002 to 0.010 |
| **7.** | Compactness | 0.05 to 0.2 | 0.008 to 0.061 |

By using data analysis techniques, the above comprehensive findings on malignant breast tumors have been found which can serve as a foundation for developing predictive models to ascertain the diagnosis and malignancy of the disease. By integrating the diverse morphological features such as size, shape, texture, concavity, smoothness, and compactness, machine learning algorithms can be trained to recognize intricate patterns indicative of malignancy. This predictive approach holds immense potential for refining diagnostic procedures, enhancing the accuracy of breast cancer assessments, and facilitating early detection, thereby significantly improving patient outcomes, and guiding tailored treatment strategies based on the nuanced characteristics revealed through in-depth data analysis.

**Unsolved issues:**

Unexplored Feature Interactions: Investigating how different features interact with each other could unveil hidden patterns and provide a more comprehensive understanding of malignant tumor characteristics. Lack of Genetic and Molecular Insights: Incorporating genetic data could uncover specific mutations or biomarkers associated with malignancy, enhancing the precision of diagnostic models, and contributing to targeted therapies.

**Interesting findings:**

An intriguing finding is the inverse relationship between smoothness and area in malignant tumors. Tumors with larger areas tend to exhibit lower smoothness. This observation could prompt further investigation into the biological implications of this relationship and its potential significance in tumor aggressiveness.

**Limitations of the study:**

The study relies solely on mean and standard error values to characterize malignant tumors. This approach may oversimplify the complexity of tumor characteristics, overlooking potential nuances and outliers that could provide valuable insights.

The analysis lacks temporal information, and therefore, it doesn't account for changes in tumor features over time. Breast cancer is dynamic, and understanding its evolution could contribute to more accurate prognostic models.

The study may be based on a limited sample size, potentially impacting the generalizability of the findings.

**Comparison with prior studies:**

This study provides quantitative insights into specific morphological features for breast cancer diagnosis. Some of the previous studies address specific aspects of breast cancer heterogeneity, ERBB2-positive cancer, and ultrasound image classification, demonstrating the multifaceted nature of breast cancer research. This study could complement these by offering detailed insights into the quantitative variability of certain morphological features, potentially aiding in the broader understanding and classification of breast tumors.

**Conclusion:**

In conclusion, our study using advanced computer techniques has shown positive strides in accurately detecting breast cancer. The developed predictive models exhibit promise in distinguishing between benign and malignant tumors based on various tumor characteristics. This computational approach, utilizing features like size and shape, holds significant potential for refining diagnostic processes and improving the precision of breast cancer assessments. While these findings are encouraging, there's still more to explore, such as understanding how different features interact and delving into the genetic aspects. The incorporation of powerful visualization tools like Stata and Orange workflow enhancing the credibility of our results. This research contributes to the ongoing enhancement of breast cancer detection through the application of sophisticated computational tools.

**REFERENCES**

· Aleskandarany, M. A., Vandenberghe, M., Marchió, C., Ellis, I. O., Sapino, A., & Rakha, E. A. (2018). Tumour Heterogeneity of breast cancer: From morphology to Personalised Medicine. *Pathobiology*, *85*(1–2), 23–34.<https://doi.org/10.1159/000477851>

· Bychkov, D., Linder, N., Tiulpin, A., Kücükel, H., Lundin, M., Nordling, S., Sihto, H., Isola, J., Lehtimäki, T., Kellokumpu‐Lehtinen, P., Von Smitten, K., Joensuu, H., & Lundin, J. (2021). Deep learning identifies morphological features in breast cancer predictive of cancer ERBB2 status and trastuzumab treatment efficacy. *Scientific Reports*, *11*(1).<https://doi.org/10.1038/s41598-021-83102-6>

· Dahlke, D. V., Yoshikawa, A., McAdams, M., Malatok, S., & McAdams, M. (2021). Analysis of Healthcare Team Communication Needs Among Younger vs Older Breast Cancer Survivors: A Web-based Survey (Preprint). *JMIR Cancer*.<https://doi.org/10.2196/31118>

· Denisov, E. V., Skryabin, N. A., Gerashchenko, T. S., Таширева, Л. А., Wilhelm, J., Buldakov, M. A., Слепцов, А. А., Lebedev, I. N., Вторушин, С. В., Zavyalova, М. V., Чердынцева, Н. В., & Перельмутер, В. М. (2017). Clinically relevant morphological structures in breast cancer represent transcriptionally distinct tumor cell populations with varied degrees of epithelial-mesenchymal transition and CD44+CD24- stemness. *Oncotarget*, *8*(37), 61163–61180.<https://doi.org/10.18632/oncotarget.18022>

· Hou, I., Lin, H., Shen, S., Chang, K., Tai, H., Tsai, A., & Dykes, P. C. (2020). Quality of life of women after a first diagnosis of breast cancer using a Self-Management Support MHealth app in Taiwan: randomized controlled trial. *Jmir Mhealth and Uhealth*, *8*(3), e17084.<https://doi.org/10.2196/17084>

· Jongerius, C., Russo, S., Mazzocco, K., & Pravettoni, G. (2019). Research-Tested Mobile Apps for Breast Cancer Care: Systematic review. *Jmir Mhealth and Uhealth*, *7*(2), e10930.<https://doi.org/10.2196/10930>

· Maksimenko, J., Rodrigues, P. P., Nakazawa-Miklaševiča, M., Pinto, D., Miklaševičs, E., Trofimovičs, G., Gardovskis, J., Cardoso, F., & Cardoso, M. (2022). Effectiveness of secondary Risk–Reducing strategies in patients with unilateral breast cancer with pathogenic variants of BRCA1 and BRCA2 subjected to Breast-Conserving Surgery: Evidence-Based Simulation study. *JMIR Formative Research*, *6*(12), e37144.<https://doi.org/10.2196/37144>

· Singleton, A., Raeside, R., Partridge, S. R., Hayes, M., Maka, K., Hyun, K., Thiagalingam, A., Chow, C. K., Sherman, K. A., Elder, E., & Redfern, J. (2021). Co-designing a Lifestyle-Focused text message intervention for women after breast cancer Treatment: Mixed Methods study. *Journal of Medical Internet Research*, *23*(6), e27076.<https://doi.org/10.2196/27076>

· Wei, M., Du, Y., Wu, X., Q, S., Zhu, J., Zheng, L., Lv, G., & Zhuang, J. (2020). A benign and malignant breast tumor classification method via efficiently combining texture and morphological features on ultrasound images. *Computational and Mathematical Methods in Medicine*, *2020*, 1–12. https://doi.org/10.1155/2020/5894010