Breast Cancer Diagnosis Prediction

Yen Nga Le, Tehsin Shaikh and Srilakshmi Gummadidala

Group 3
DAB304-24W-001: Healthcare Analytics
Dr. Sutharsan Sivagnanam
Winter Term 2024
Submission Date: April 15, 2024

Abstract

Breast cancer represents a significant global health challenge. Early detection through accurate diagnosis is crucial for effective treatment and improving patient outcomes. This project develops a predictive model for breast cancer diagnosis using machine learning techniques, aiming to enhance the precision and efficiency of diagnosing breast cancer. By analyzing a dataset of clinical cases, the project seeks to predict the malignancy of breast tumors based on various attributes, offering valuable insights for medical practitioners. Through this endeavor, we aim to contribute to the advancement of breast cancer diagnostics, ultimately aiding in better patient care and management strategies. In our study, we focused on leveraging machine learning algorithms to predict and diagnose breast cancer using the Breast Cancer Wisconsin Diagnostic dataset. Specifically, we applied four algorithms: Random Forest, Logistic Regression, Decision Tree, Neural Networks within the Anaconda environment using Python and the Scikit-learn library. Our aim was to evaluate and compare the performance of these classifiers to identify the most effective approach in terms of accuracy and precision. The results revealed that Logistic Regression outperformed all other classifiers, achieving the highest accuracy of 96%. This underscores its efficacy in breast cancer prediction and diagnosis.

Keywords: Breast Cancer; Prediction; Diagnostic; Random Forest; Logistic regression; Decision Tree; Neutral Networks; Accuracy; Precision; Flask; Pickle

1. Introduction

Breast cancer is a pervasive global health challenge that continues to impact millions of lives. As a leading cause of cancer-related morbidity and mortality, the importance of early detection cannot be overstated. Accurate diagnosis not only informs treatment decisions but also plays a pivotal role in improving patient outcomes and quality of life [1].

In the pursuit of advancing medical diagnostics, this project embarks on the development of a predictive model for breast cancer diagnosis. The significance of this endeavor lies not only in the complexity of breast cancer itself but also in the potential to harness technological advancements, particularly within the realm of machine learning [1].

This project aligns with the broader mission to enhance the precision and efficiency of breast cancer diagnosis. Machine learning, with its capacity to discern patterns within vast datasets, presents a promising avenue for achieving more accurate and timely diagnoses. The predictive model, once developed, will serve as a complementary tool for medical practitioners, offering nuanced insights into the probability of malignancy based on specific attributes associated with each case [2].

Moreover, the significance of this project extends beyond the realms of technology and diagnostics. It is a testament to the collaborative efforts between medical professionals, data scientists, and researchers in addressing critical health challenges. By synergizing the expertise of these diverse fields, this initiative aspires to contribute to the ongoing dialogue on improving breast cancer diagnostics and, by extension, the lives of those affected by this formidable disease. As we delve into the details of the project methodology, outcomes, and challenges, the overarching goal remains steadfast: to propel advancements in breast cancer diagnosis and pave the way for more informed and effective healthcare interventions [2].

2. Problem Statement

- Build a machine learning model to predict whether a breast tumor is benign or malignant based on various attributes.
- Analyze the impact of each attribute on the prediction to understand the key factors influencing the diagnosis.
- Evaluate and validate the model's performance using appropriate metrics.

3. Dataset and Methodology

3.1. Dataset Source: The breast cancer dataset utilized in this project is sourced from clinical cases reported by Dr. Wolberg. The data is shared for research purposes and contributes to the broader understanding of breast cancer characteristics. Dr. Wolberg, through his clinical cases, aims to facilitate advancements in breast cancer diagnosis and treatment [3].

3.2. Data Collection and Funding: The breast cancer dataset is collected through the reporting of clinical cases by Dr. Wolberg. Privacy and data quality are paramount considerations, aligning with ethical standards in medical research. The dataset is made available for research purposes, with a commitment to protecting patient privacy and ensuring the confidentiality of sensitive medical information. No personal patient identifiers are disclosed in the dataset, adhering to strict privacy protocols [3].

3.3. Participants or Cases: The dataset comprises individual clinical cases of breast cancer, with each instance representing a unique patient case. To uphold privacy standards, personal identifiers such as patient names are not included. The dataset aims to provide a comprehensive representation of varied breast cancer cases while safeguarding the anonymity of patients [3].

3.4. Variables of Interest: Various attributes are included in the breast cancer dataset to provide comprehensive information about each clinical case. These attributes encompass Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses, and the class attribute indicating benign (2) or malignant (4) tumors [3].

3.5. Database Schema: The dataset does not adhere to a traditional relational database schema. Instead, it is structured as a tabular dataset, with each row representing a clinical case and each column representing specific attributes. The absence of inter-table relationships aligns with the nature of the dataset, where attributes directly pertain to each clinical case without the need for complex relational structures. This schema simplifies data exploration and analysis, facilitating the development of predictive models for breast cancer diagnosis [3].

3.6. Data Field Description:

Field	Field Name	Description
1	Sample code number	Identification number for each clinical case
2	Clump Thickness	Rating on a scale of 1 to 10
3	Uniformity of Cell Size	Rating on a scale of 1 to 10
4	Uniformity of Cell Shape	Rating on a scale of 1 to 10
5	Marginal Adhesion	Rating on a scale of 1 to 10
6	Single Epithelial Cell Size	Rating on a scale of 1 to 10
7	Bare Nuclei	Rating on a scale of 1 to 10
8	Bland Chromatin	Rating on a scale of 1 to 10
9	Normal Nucleoli	Rating on a scale of 1 to 10
10	Mitoses	Rating on a scale of 1 to 10
11	Class	Diagnosis class (2 for benign, 4 for malignant)

Table 1: Data Field Description

3.7. Dataset size:

The total number of instances in the dataset is 699, as of the donated database on 15 July 1992. Each instance likely represents a clinical case with various attributes related to breast cancer [3].

The breakdown of instances by groups is as follows:

- Group 1: 367 instances (January 1989)
- Group 2: 70 instances (October 1989)
- Group 3: 31 instances (February 1990)
- Group 4: 17 instances (April 1990)
- Group 5: 48 instances (August 1990)
- Group 6: 49 instances (Updated January 1991)
- Group 7: 31 instances (June 1991)
- Group 8: 86 instances (November 1991)

4. Methodology

4.1. Data Preprocessing using Python Libraries

The preparation of the breast cancer dataset for analysis was a meticulous process, heavily reliant on Python libraries, particularly Pandas. Employing Pandas, we meticulously handled data quality concerns such as duplicates, missing values, and inconsistencies within the clinical cases.

The systematic approach involved identifying and addressing any discrepancies, ensuring a clean and standardized dataset. By mitigating potential sources of noise and errors in the raw data, this preprocessing laid the groundwork for subsequent machine learning model development, contributing to the reliability of our analytical outcomes.

4.2. Data Visualization with Tableau

Tableau proved to be an indispensable tool for translating the nuances of the breast cancer dataset into meaningful and visually compelling insights. Leveraging Tableau's capabilities, we crafted interactive visualizations that vividly portrayed trends in breast cancer cases over time and across different attributes. These visualizations not only facilitated a comprehensive exploration of the dataset but also enhanced the interpretability of findings.

This approach not only streamlined the analytical process but also ensured effective communication of insights, making complex patterns and trends accessible to a broad audience. The integration of Tableau into our analytical workflow significantly elevated the presentation and interpretative aspects of our breast cancer dataset analysis.

5. Exploratory Data Analysis

5.1 Data Preprocessing

Handling Missing Values**:** Missing values represented by '?' were replaced with NA and then dropped from the dataset. This step ensured data integrity and eliminated instances with incomplete information.

Data Type Conversion**:** The 'Bare Nuclei' column was converted to the integer data type ('int64') for consistency and ease of analysis.

Checking for Duplicates**:** Duplicate rows were identified and removed from the dataset to prevent bias in analysis caused by redundant data.

5.2 Data Visualization

5.2.i. Distribution of Class Variable (Pie Chart)

The bar chart visualizes the distribution between the two classes: benign (negative, labeled as "2") and malignant (positive, labeled as "4"). In this dataset, there are 699 instances benign (444, 65%), and malignant (255, 35%). This proportion highlights the dataset's imbalance, which is crucial to consider when developing predictive models to ensure they're not biased towards the more prevalent class.

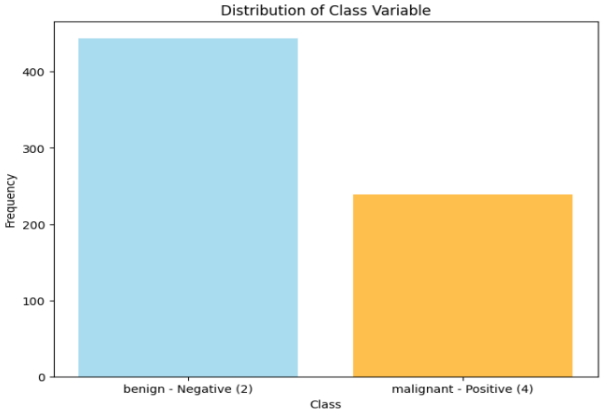

Fig1: Distirbution of Class Variable

5.2. ii. Histograms for Numerical Variables

The histograms represent the distribution of values across the various features in the dataset. The x-axis of each histogram indicates the feature's range of values, and the y-axis shows the frequency of observations.

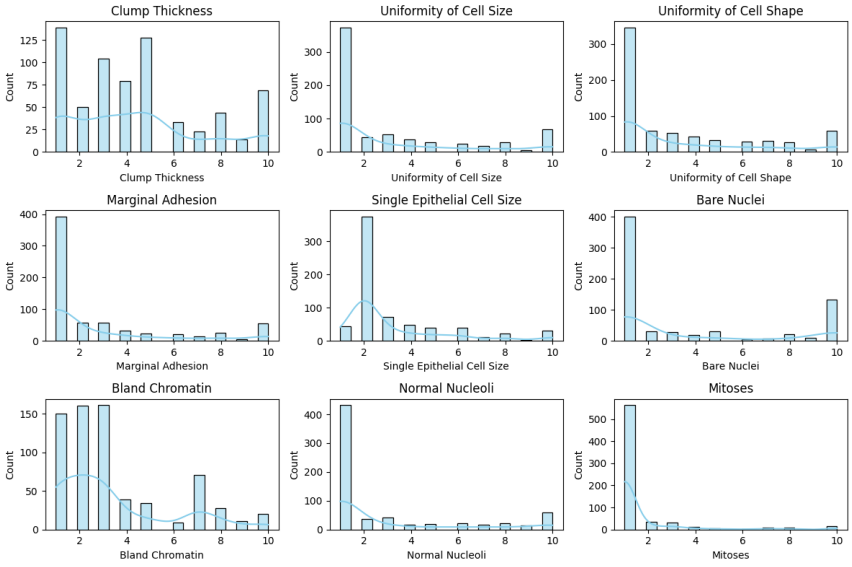

Fig2: Distirbution of Numeric Variables

The dataset exhibits diverse distributions among key features: Clump Thickness suggests two distinct groups, possibly representing extreme cases; right-skewed histograms for Uniformity of Cell Size and Shape imply rarity of high uniformity, possibly indicating aggressive behavior. Marginal Adhesion's right skew suggests low adhesion prevalence, potentially malignant. Single Epithelial Cell Size shows a tapering distribution with few enlarged cells. Bare Nuclei displays bimodality, indicating higher cancer risk at score 10. Bland Chromatin exhibits slight right skew, with variability across scores. Normal Nucleoli's right skew suggests benign cells predominance. Mitoses' right skew and rarity of high scores signify infrequent cell division, a key indicator of aggressive tumor behavior. These insights are crucial for feature selection and model training, aiding in predicting tumor outcomes accurately.

5.2.iii. Box Plots for Outlier Detection

The box plots provide a graphical representation of the distribution of the numerical data and help identify outliers for each feature. The central box represents the interquartile range (IQR), the horizontal line inside the box marks the median, and the "whiskers" show the range of the data, excluding outliers.

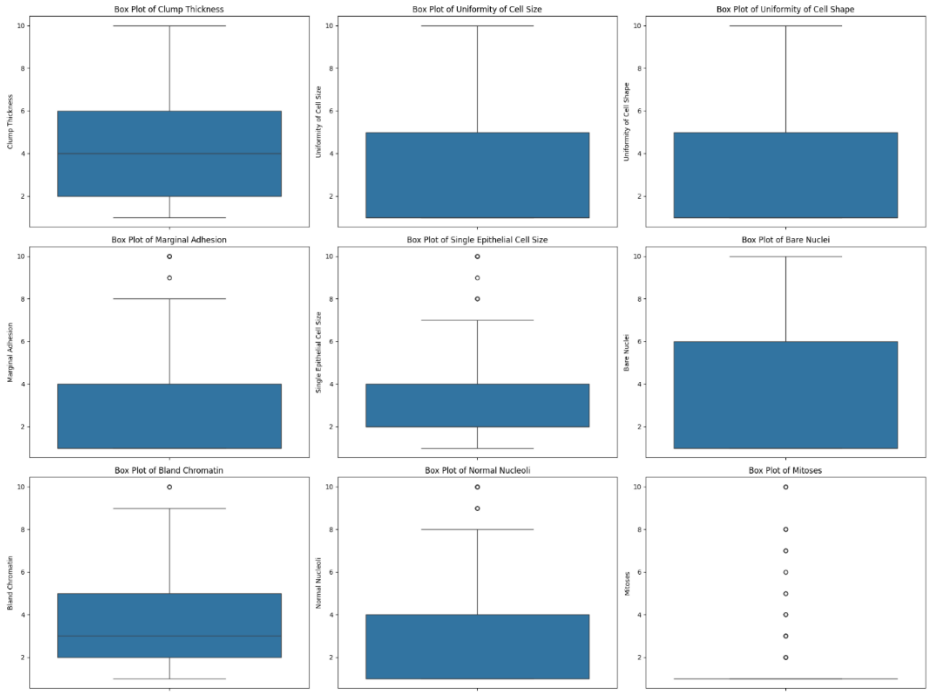

Fig3: Box plot for Outlier Detection

The box plots reveal notable insights into the variability of features: Clump Thickness displays considerable variance without outliers. Uniformity of Cell Size and Shape exhibit symmetrical spreads with higher medians, suggesting prevalent uniformity. Marginal Adhesion highlights numerous outliers at higher values, contrasting with a lower median. Single Epithelial Cell Size shows outliers at larger sizes and a lower median. Bare Nuclei's compact plot with extreme outliers signifies rare high counts. Bland Chromatin and Normal Nucleoli display moderate spreads with outliers at higher values and lower medians. Mitoses exhibits numerous outliers at higher scores and a low median, indicating rarity of high mitotic rates. The presence of outliers underscores the significance of atypical cell characteristics, potentially vital in cancer diagnosis.

5.2. iv. Correlation Matrix (Heatmap)

The heatmap displays the correlation coefficients between variables in a matrix format, where each square shows the correlation between two features. A coefficient close to 1 or -1 indicates a strong positive or negative correlation, respectively, while a coefficient close to 0 indicates no linear relationship.

The correlation analysis reveals crucial associations: Clump Thickness correlates strongly with most variables, particularly with 'Uniformity of Cell Size' and 'Shape' (approx. 0.65). 'Uniformity of Cell Size' and 'Shape' are highly correlated (0.91) and strongly predict malignancy (0.82). Marginal Adhesion correlates strongly with 'Uniformity of Cell Size' and 'Shape' (0.71) and 'Class' (0.71). Single Epithelial Cell Size shows high correlation with 'Uniformity of Cell Size' and 'Shape' (0.75 and 0.72) and with 'Class' (0.69). Bare Nuclei demonstrates strong correlations with 'Uniformity of Cell Size' and 'Shape' (0.69 and 0.71) and 'Class' (0.82). Bland Chromatin is strongly correlated with 'Class' (0.76) and moderately with 'Uniformity of Cell Size' and 'Shape' (0.76 and 0.74). Normal Nucleoli correlates strongly with 'Uniformity of Cell Size' and 'Shape' (0.72 each) and 'Class' (0.72). Mitoses exhibits weaker correlations overall, especially with 'Class' (0.42), suggesting its lesser predictive value. 'Class' correlates strongly with most variables, emphasizing the importance of 'Uniformity of Cell Size' and 'Shape' and 'Bare Nuclei' (all 0.82) in indicating malignancy.

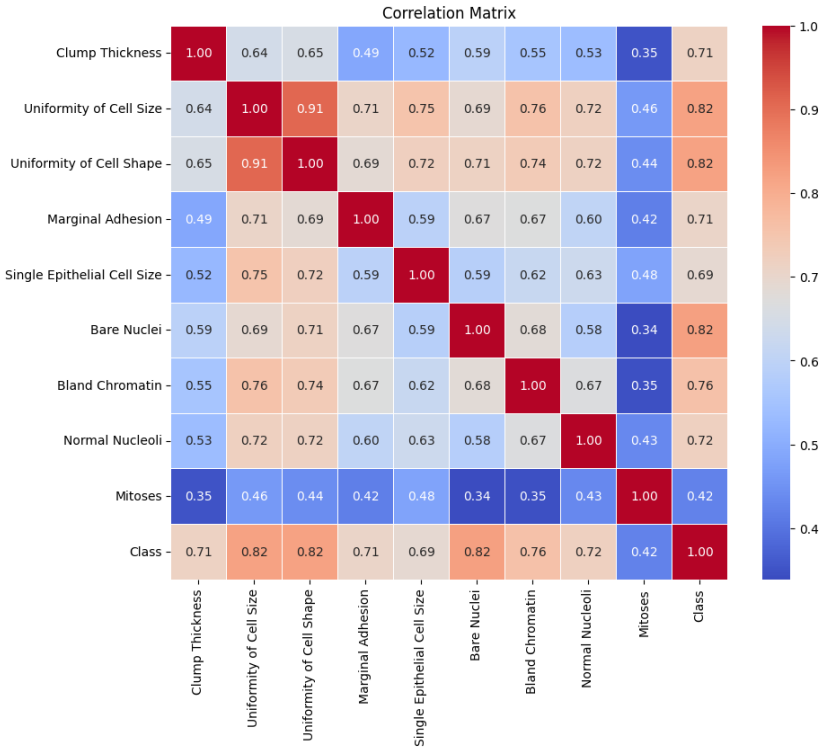

Fig4: Correlation Matrix

5.3. Feature Importance

Feature importance is an integral part of model interpretability, especially in the context of medical diagnosis where each variable can represent a critical piece of information about patient health. In machine learning, particularly with algorithms like RandomForest, feature importance provides a quantitative measure of the impact each feature has on the predictions of the model.

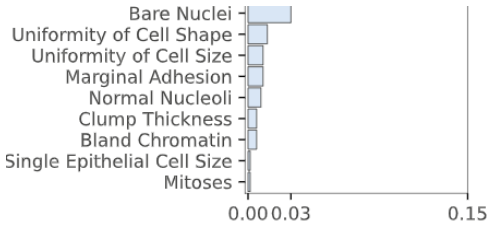

Fig5: Feature Importance

In the feature importance chart, 'Bare Nuclei' stands out with the longest bar, suggesting its pivotal role in predicting tumor malignancy. Changes in this feature strongly influence the model's decisions. 'Uniformity of Cell Shape' and 'Uniformity of Cell Size' follow with relatively long bars, underscoring their significance albeit to a lesser extent compared to 'Bare Nuclei'. 'Marginal Adhesion' and 'Normal Nucleoli' exhibit moderate importance, contributing to the model's accuracy but not as significantly as the top features. 'Clump Thickness' and 'Bland Chromatin' show shorter bars, indicating their influence on the model's decisions is less pronounced. 'Single Epithelial Cell Size' ranks lower in importance with a shorter bar, while 'Mitoses' has the shortest bar, suggesting its minimal impact on the model's predictions compared to other features.

6. Data Modeling

The primary aim of this report is to evaluate the efficacy of several data modeling techniques in predicting breast cancer malignancy. We have applied a suite of machine learning algorithms, namely RandomForest Classifier, Logistic Regression, Decision Tree Classifier, and a Neural Network, to a dataset comprised of features extracted from breast mass images. Our target is to discern whether tumors characterized within the dataset are benign (class 2) or malignant (class 4).

6.1. Model Comparisons

Model	Accuracy	Class 2 Precision	Class 2 Recall	Class 4 Precision	Class 4 Recall	Macro Avg.	Weighted Avg.
RandomForest Classifier	94.89%	0.93	0.99	0.98	0.9	0.95	0.95
Logistic Regression (Initial)	96%	0.94	0.99	0.98	0.91	0.95	0.96
Logistic Regression (Post GridSearchCV)	96%	0.94	0.99	0.98	0.91	0.95	0.96
Decision Tree Classifier	93.43%	0.92	0.97	0.96	0.88	0.93	0.93
Neural Network	87.59%	0.84	0.97	0.96	0.74	0.87	0.87

Table 2: Model Comparisons

6.2. Best Model Evaluation:

Based on the accuracy and F1-scores, the Logistic Regression model (both before and after hyperparameter tuning) performs the best among the models listed. It has the highest accuracy (96%) and strong performance across both classes. Furthermore, its weighted and macro averages are slightly higher than those of the Random Forest model, which comes in as a close second. The high precision and recall for the Logistic Regression model suggest that it has a good balance between correctly identifying positive cases and minimizing false positives.

The Decision Tree model shows slightly lower performance metrics compared to the Random Forest, and the Neural Network model, while complex, does not seem to provide a substantial benefit over the simpler models based on the provided metrics.

6.3 Model Deployment

The deployment of the breast cancer classification model using Flask and pickle demonstrates a seamless integration of machine learning into practical healthcare solutions. Flask, as a micro web framework for Python, provides a straightforward and efficient platform for hosting web applications. By encapsulating the trained model within a Flask web application, healthcare professionals and individuals alike gain access to a user-friendly platform for predicting the likelihood of breast cancer based on key diagnostic features. This deployment not only streamlines the process of decision-making in clinical settings but also empowers individuals to take proactive steps towards their health by providing timely and accurate insights [6].

Furthermore, pickle, a module in Python used for serializing and deserializing objects, plays a critical role in model deployment. With pickle's serialization capabilities, the trained breast cancer classification model can be saved to disk in a binary format. This serialized representation preserves the model's architecture, parameters, and internal state, allowing it to be easily reloaded and used for making predictions in production environments [6].

Leveraging logistic regression as the underlying classification algorithm enhances the interpretability and computational efficiency of the deployed model. The utilization of Flask's lightweight framework, pickle's serialization capabilities, and logistic regression's simplicity ensures scalability, efficiency, and accuracy in deploying the model across various environments. With minimal overhead and rapid response times, this deployment facilitates widespread adoption and usage, ultimately contributing to the advancement of breast cancer diagnosis and treatment. As technology continues to intersect with healthcare, such deployments serve as a testament to the transformative potential of machine learning in improving patient care and outcomes.

Conclusion

This project has effectively demonstrated the transformative potential of machine learning in the field of medical diagnostics, particularly in the early detection and accurate diagnosis of breast cancer. By harnessing sophisticated algorithms to analyze clinical data, the developed predictive model successfully distinguishes between benign and malignant breast tumors with impressive accuracy. This breakthrough serves as a testament to the power of integrating technological advancements with medical science. The model not only provides medical practitioners with a valuable diagnostic tool but also offers a deeper understanding of the factors that contribute to the malignancy of tumors. Consequently, this project contributes significantly to the ongoing efforts to improve patient outcomes and streamline diagnostic processes.

Furthermore, the project underscores the importance of interdisciplinary collaboration in tackling complex health challenges. The synergy between data scientists, clinicians, and researchers has culminated in a robust model that leverages data to enhance decision-making in clinical settings. As this model is further refined and adapted for practical use, it holds the promise of revolutionizing breast cancer diagnostics. This initiative exemplifies how technological innovation can advance healthcare, suggesting a future where predictive modeling and machine learning are integral to diagnosing and managing diseases more effectively, thereby shaping a new era in precision medicine.

References

[1] World Health Organization. (n.d.). Breast cancer. WHO. (Breast cancer (who.int)

[2] Kourou, K., Exarchos, K. P., Papaloukas, C., Sakaloglou, P., Exarchos, T., & Fotiadis, D. I. (2021). Applied machine learning in cancer research: A systematic review for patient diagnosis, classification and prognosis. Computational and Structural Biotechnology Journal, 19, 5546–5555.(Applied machine learning in cancer research: A systematic review for patient diagnosis, classification and prognosis - PMC (nih.gov))

[3] O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.

[4] William H. Wolberg and O.L. Mangasarian: "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December 1990, pp 9193-9196.(Multisurface method of pattern separation for medical diagnosis applied to breast cytology. | PNAS)

[5] O. L. Mangasarian, R. Setiono, and W.H. Wolberg: "Pattern recognition via linear programming: Theory and application to medical diagnosis", in: "Large-scale numerical optimization", Thomas F. Coleman and Yuying Li, editors, SIAM Publications, Philadelphia 1990, pp 22-30.(Pattern Recognition Via Linear Programming: Theory and Application to Medical Diagnosis | Semantic Scholar)

[6] K. P. Bennett & O. L. Mangasarian: "Robust linear programming discrimination of two linearly inseparable sets", Optimization Methods and Software 1, 1992, 23-34 (Gordon & Breach Science Publishers).

[7] Hariramani, V. (2021, February 22). End to End Deployment of Breast Cancer Prediction Through Machine Learning using Flask. Geeky Bawa. (End to End Deployment of Breast Cancer Prediction Through Machine Learning using Flask | by VAIBHAV HARIRAMANI | GEEKY BAWA | Medium)