**Optimizing Diagnostic Accuracy for Breast Cancer through Advanced Predictive Models**

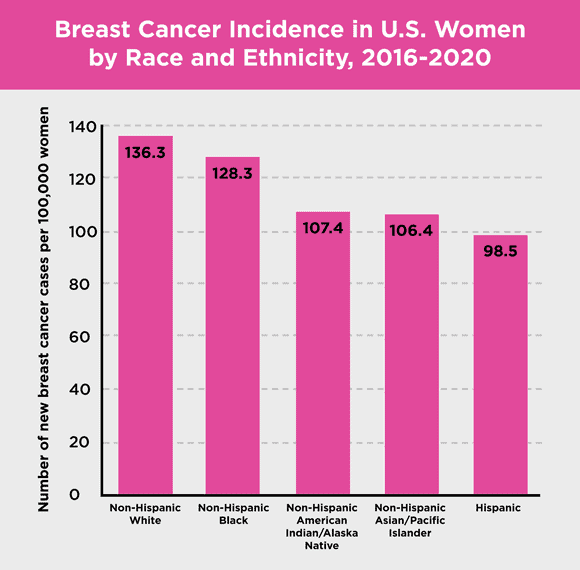
Syed Ishmam Alawee, Azmayn Inkishaf

**Abstract**

**1. Introduction**

*1.1 Breast Cancer: A Global Health Challenge*

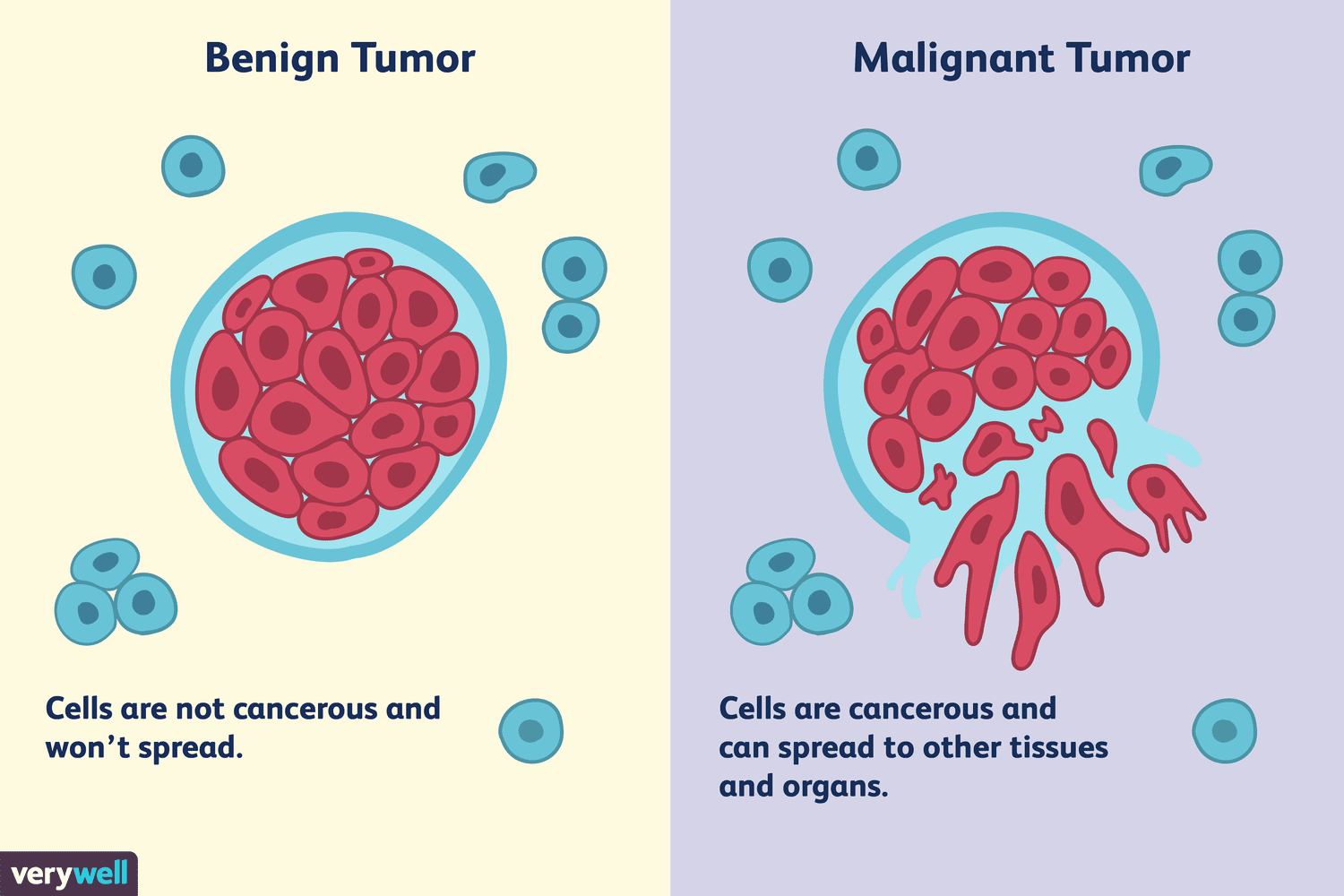
Breast cancer, a multifaceted and prevalent disease, continues to pose significant challenges to global public health. Over the years, its impact has been profound, with statistics painting a stark picture of its increasing burden on individuals and healthcare systems worldwide. According to the World Health Organization (WHO), breast cancer ranks as the most commonly diagnosed cancer among women, with an estimated 2.3 million new cases reported in 2020 alone [1].

 **Figure 1**: The bar chart illustrates breast cancer incidence rates among U.S. women from 2016 to 2020, highlighting the disparities in new cases across different racial and ethnic groups [www.komen.org/breast-cancer]

Furthermore, breast cancer remains a leading cause of cancer-related mortality, claiming the lives of approximately 685,000 women annually [1]. These figures underscore the urgent need for comprehensive strategies aimed at early detection, accurate diagnosis, and effective treatment to mitigate the devastating impact of breast cancer on individuals and communities.

*1.2 Understanding Breast Cancer Pathology*

Understanding the intricacies of breast cancer is essential for devising effective interventions and improving patient outcomes. In their seminal paper on computer-aided diagnosis of breast cancer, Dua and Graff provide a comprehensive overview of the disease's pathology, diagnostic modalities, and prognostic indicators [2]. Breast cancer, a heterogeneous disease characterized by the abnormal growth of cells in the breast tissue, manifests in various forms, each with distinct clinical and pathological features. Central to the diagnosis and management of breast cancer is the differentiation between benign and malignant tumors, which dictates the appropriate course of treatment and prognosis.

 **Figure 2**: Comparing benign and malignant tumors. Benign tumors don't spread, while malignant ones have cancerous cells that can invade other tissues. [www.verywellhealth.com]

Benign tumors, often non-cancerous growths, exhibit localized growth patterns and do not invade adjacent tissues or metastasize to distant sites. In contrast, malignant tumors, hallmark features of cancer, display invasive behavior, infiltrating surrounding tissues and potentially spreading to distant organs, leading to life-threatening complications. The ability to accurately distinguish between benign and malignant tumors is critical for guiding treatment decisions and optimizing patient outcomes.

*1.3 The Role of Machine Learning in Breast Cancer Diagnosis*

Traditionally, diagnosing breast cancer has involved clinical exams, imaging studies, and biopsies for tissue analysis. Mammography, the primary tool for breast cancer screening, is crucial for detecting early tumors and guiding further testing. However, traditional diagnostic methods have limitations, such as inconsistencies between observers, false positives, and missed diagnoses.

A screenshot of a screen

Description automatically generated **Figure 3**: Contrasts breast ultrasound and mammography, comparing their purposes, working principles, and image capabilities. [www.verywellhealth.com]

In recent years, machine learning (ML) has transformed breast cancer diagnosis by offering more accurate and objective approaches. ML algorithms, driven by advanced mathematical models, excel at analyzing complex data to uncover patterns that humans might miss, making them effective for detecting and understanding breast tumors.

The success of machine learning in breast cancer diagnosis hinges on accurately digitizing tumor features, using BI-RADS-defined attributes like shape and margin, alongside statistical data. Properly translating these features into digital form enhances the precision of distinguishing benign from malignant tumors. Aligning machine learning with the BI-RADS system streamlines diagnosis, offering greater accuracy and adaptability, thus revolutionizing breast cancer care with earlier detection and personalized treatment.

**2. Methodology**

*2.1 Data Acquisition and Preprocessing*

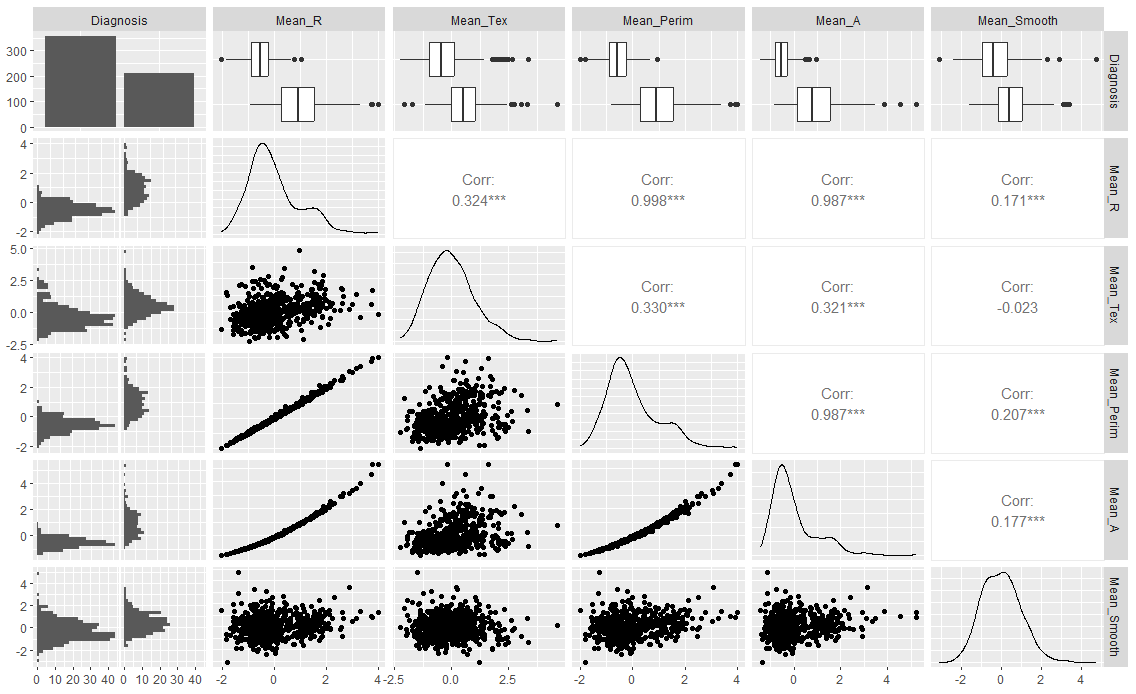
The dataset utilized in this study originates from the Wisconsin Breast Cancer Diagnostic Database, which is based on the work of K. P. Bennett and O. L. Mangasarian as described in their paper "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets" [11]. This dataset consists of multivariate features derived from digitized images of fine needle aspirates (FNAs) of breast masses. Each instance encapsulates diagnostic attributes encompassing cell nuclei characteristics, morphological parameters, and spatial descriptors. Prior to analysis, the dataset undergoes meticulous preprocessing to ensure homogeneity and mitigate confounding effects [12]. Robust preprocessing entails encoding categorical variables, standardizing continuous features, and partitioning the dataset into training and validation subsets using stratified sampling techniques to preserve class distributions [13].

**A diagram of a business flow

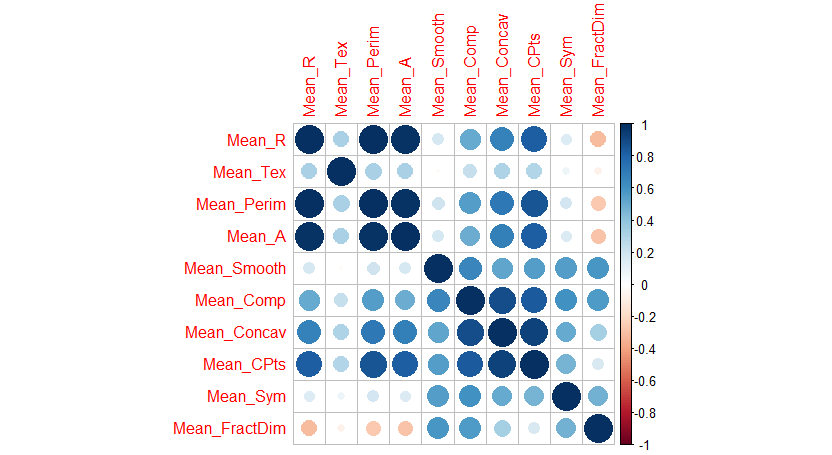
Description automatically generatedFigure 4**: Illustrating the machine learning process where raw data undergoes feature processing, is divided into training, validation, and test sets, and is then used to train a model for predicting target outcomes

*2.2 Exploratory Data Analysis*

Exploring data deeply is crucial. We use various stats and pictures, like pair plots and correlation charts, to find patterns and understand relationships. Tools like GGally and corrplot help us see connections and important features for diagnosing breast cancer.

 **Figure 5**: Shows a correlation matrix of various features used in breast cancer diagnosis, visualizing the relationships between diagnosis, mean radius, texture, perimeter, area, and smoothness.

The plot illustrates that many health-related variables, such as diabetes, physical activity, and cholesterol levels, exhibit binary distributions with little variance, indicating distinct groups in the dataset. The boxplots reveal that BMI has a wider distribution with notable outliers. Strong correlations between variables like Mean\_R, Mean\_Perim, and Mean\_A, as seen in the scatter plots and correlation matrix, suggest significant interdependencies that highlight the relationship between body metrics.



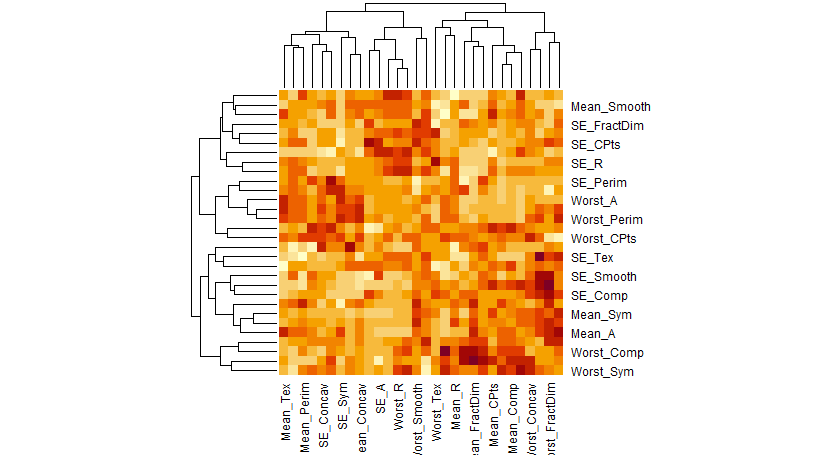
**Figure 6**: Visualizes the correlation between different breast cancer features, represented as a color-coded correlation matrix. The darker shades indicate stronger relationships between the features

This correlation matrix visualizes the relationships between several metrics of medical interest, where the size and color intensity of the circles indicate the strength and direction of the correlations. The strong positive correlations among the variables, indicated by large dark blue circles, suggest these features—like Mean\_R, Mean\_Perim, and Mean\_A—are highly related and tend to increase or decrease together. Conversely, lighter circles or reddish hues reflect weak or negative correlations. These insights reveal which metrics are strongly interconnected, hinting that they might represent similar underlying phenomena or redundant information, which can inform feature selection and model building. The standard error and worst features showed a similar pattern to the mean correlation matrix.

To choose the most suitable feature set based on the correlation plots, we considered the predictive objective of your model: If the focus is on capturing average patterns and trends, the Mean features provide a balanced representation of typical behavior. For models aimed at identifying extreme or high-risk cases, the Worst features may offer better insights as they capture the most severe deviations. If variability and measurement consistency are crucial to your model, the Standard Error features will help highlight trends related to data uncertainty. Ultimately, evaluating the predictive power of each set in the context of our specific goals will guide you toward the most effective choice.

*2.3 Feature Selection*

Feature engineering constitutes a pivotal facet of the methodology, underpinning the extraction of salient information and dimensionality reduction [17]. Feature engineering endeavors encompass the creation of composite features, transformation of variables, and reduction of feature space via techniques such as principal component analysis (PCA) and recursive feature elimination (RFE) [18]. Emphasis is placed on delineating discriminative features conducive to robust model performance while mitigating the curse of dimensionality.

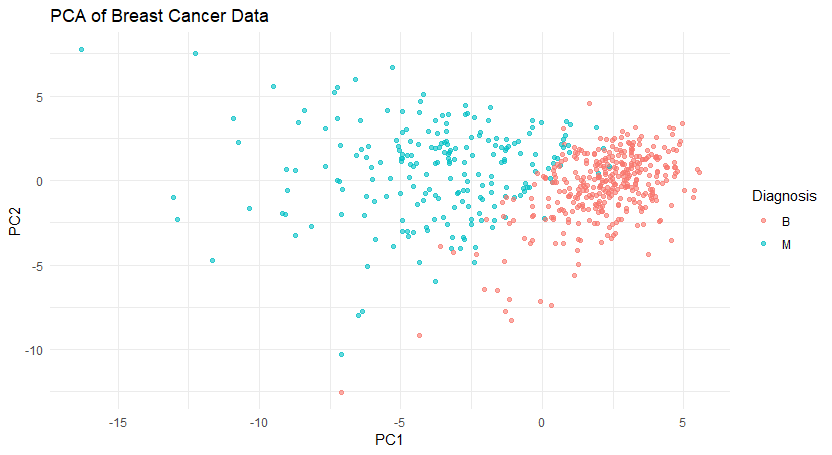
 **Figure 7**: Presents a heatmap showing the hierarchical clustering of breast cancer features. The color intensity highlights correlations, with similar features grouped together.

The heatmap highlights those features like Worst\_A, Worst\_Perim, and Worst\_CPts show strong clustering, indicating they are closely related and likely to be significant in identifying breast cancer. Similarly, Mean\_A, Mean\_Perim, and SE\_A cluster with other important features, suggesting their predictive power for breast cancer identification. Overall, the Worst features stand out as particularly crucial, but combining them with specific Mean and SE features provides a more comprehensive set of variables that may enhance the model's ability to identify breast cancer effectively.

A chart with red and blue dots

Description automatically generated **Figure 8**: Shows a swarm plot of the standardized mean features, distinguishing between benign (B) and malignant (M) breast cancer diagnoses.

Figure 7 shows the distribution of standardized mean features for benign (B) and malignant (M) breast cancer diagnoses, revealing key insights. Distinct clustering patterns highlight certain features like Mean\_R and Mean\_Concav, which demonstrate potential for differentiating between benign and malignant cases. Conversely, some features show overlap between diagnoses, suggesting they might not be as useful for classification alone. The plot helps identify which features hold the most diagnostic value, offering guidance for feature selection in machine learning models.

**Figure 9**: shows a principal component analysis (PCA) of breast cancer data, visualizing the distribution of benign and malignant tumors based on the first two principal components.

This PCA plot shows a clear separation between benign (B) and malignant (M) breast cancer cases, indicating that the principal component analysis effectively reduced the data's dimensionality while preserving critical information for distinguishing between these two diagnoses. The clear clustering of red and blue dots suggests that the first two principal components capture much of the data's original variation, revealing distinct differences between benign and malignant cases. This shows that these components can be valuable features for models that aim to distinguish different cancer diagnoses.

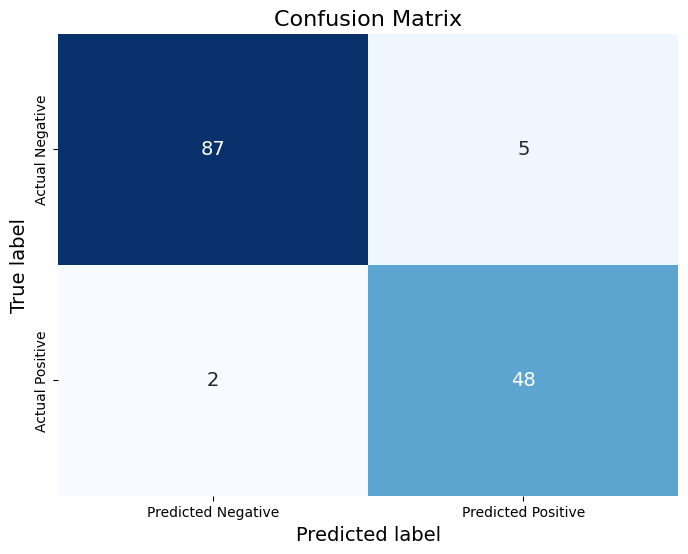
*2.4 Modelling*

The crux of the methodology resides in the deployment of an ensemble of sophisticated machine learning algorithms, underpinned by robust cross-validation schemes and hyperparameter optimization techniques [19]. A pantheon of algorithms, including logistic regression, random forest, support vector machine (SVM), gradient boosting machine (GBM), neural network, and k-nearest neighbors (KNN), are meticulously trained and evaluated. Leveraging the caret package [20], model training ensues utilizing a stratified k-fold cross-validation approach to robustly estimate model performance metrics, including accuracy, precision, recall, and area under the receiver operating characteristic curve (AUC-ROC) [21]. Hyperparameter tuning is systematically conducted via grid search, harnessing computational resources to identify optimal model configurations conducive to maximal discriminatory power and generalization capacity [22].

2.5 Evaluation

In evaluating the performance of the regression models, several key metrics were utilized for assessment such as accuracy, sensitivity, specificity, AUCROC and SHAP. Accuracy, measuring the proportion of correctly predicted instances, reflects the overall correctness of predictions, crucial for tasks like disease diagnosis or treatment outcome prediction. Sensitivity, or recall, quantifies the model's ability to correctly identify true positive cases, vital for early detection and intervention in healthcare scenarios. Specificity, on the other hand, gauges the model's accuracy in identifying true negative cases, crucial for ruling out the absence of a condition accurately, thereby reducing false alarms and unnecessary interventions. AUC-ROC (Area Under the Receiver Operating Characteristic Curve) provides a holistic evaluation of the model's ability to discriminate between positive and negative classes across various thresholds, offering insights into its overall predictive power. Additionally, SHAP (Shapley Additive Explanations) values aid in interpreting the model's decisions by attributing the importance of each feature to its predictions, enhancing transparency and interpretability. These evaluation metrics collectively ensure a thorough assessment of the regression model's suitability for real-world healthcare applications, enabling stakeholders to make informed decisions regarding its deployment and use.

**3. Result and discussion**

A blue squares with white text

Description automatically generated

Figure 10a: Logistic Regression Figure10b: Support Vector

A blue squares with white text

Description automatically generatedA blue squares with white text

Description automatically generated

Figure 10c: Knearest Neighbor Figure 10d: Decision Trees

A blue squares with white text

Description automatically generated

Figure 10e: Random Forest

The presented diagram represents the confusion matrix generated from the deployment of various models aimed at diagnosing breast cancer in patients. The vertical axis denotes the actual dataset results, while the horizontal axis indicates the corresponding predicted outcomes. This matrix delineates the classification into true positive, true negative, false positive, and false negative categories. For instance, the logistic regression model demonstrated proficiency in identifying 48 positive cases and 87 negative cases accurately. However, it misclassified 2 positive cases and 4 negative cases.

A graph of a function

Description automatically generatedA graph of a function

Description automatically generated with medium confidence

Figure 11a: Logistic Regression Figure 11b: Support Vector

A graph of a function

Description automatically generated with medium confidenceA graph of a function

Description automatically generated with medium confidence

Figure 11c: Knearest Neighbor Figure 11d: Decision Trees

A graph of a function

Description automatically generated with medium confidence

*Figure 11e: Random Forest*

The depicted diagram illustrates the Receiver Operating Characteristic (ROC) curve associated with a model utilized for breast cancer diagnosis. Herein, the y-axis signifies the true positive rate, while the x-axis denotes the false positive rate. Despite apparent similarities among multiple curves, their performance disparities become discernible through the ROC value. For instance, the logistic regression model garnered a commendable score of 0.95 within this matrix, reflecting exemplary performance. Such an achievement signifies a highly effective classifier characterized by robust discriminative prowess and predictive accuracy, thereby rendering it apt for diverse practical applications.

Table 1: Evaluation of model using mean value of features

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Sensitivity** | **Specificity** | **AUC** |
| Logistic Regression | 95.1 | 96.0 | 94.6 | 95.3 |
| Support Vector | 95.1 | 95.0 | 97.8 | 93.9 |
| Knearest Neighbor | 93.7 | 90.0 | 95.6 | 92.8 |
| Decision Trees | 95.8 | 94.0 | 96.7 | 95.3 |
| Random Forest | 92.3 | 88.0 | 94.6 | 91.3 |

Table 2: Evaluation of model using standard error of features

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Sensitivity** | **Specificity** | **AUC** |
| Logistic Regression | 85.2 | 74.0 | 91.3 | 82.7 |
| Support Vector | 82.4 | 72.0 | 88.0 | 80.0 |
| Knearest Neighbor | 73.9 | 68.0 | 77.2 | 72.6 |
| Decision Trees | 82.4 | 78.0 | 84.8 | 81.4 |
| Random Forest | 83.8 | 74.0 | 89.1 | 81.6 |

Table 3: Evaluation of model using worst value of features

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Sensitivity** | **Specificity** | **AUC** |
| Logistic Regression | 95.1 | 90.0 | 97.8 | 93.9 |
| Support Vector | 95.1 | 90.0 | 97.8 | 93.9 |
| Knearest Neighbor | 95.1 | 90.0 | 97.8 | 93.9 |
| Decision Trees | 89.4 | 86.0 | 91.3 | 88.7 |
| Random Forest | 92.2 | 86.0 | 95.7 | 90.8 |

We ran the model three times- once with mean values, once with standard error and once with worst value. This allowed us to explore how this might affect the performance of the models. This also prevented overfitting like the exploratory analysis suggested. Another research paper which used the same dataset ran only the SVM model with all the features. The evaluation that was used was accuracy and our model achieved similar result with this different approach and evidence suggest that our model might perform better in many other aspects.

The evaluation of various machine learning models for breast tumor diagnosis using different representations of features provides valuable insights into their performance and robustness. Across the board, logistic regression demonstrates consistent and high performance, suggesting its reliability in classification tasks regardless of the feature representation used.

Support vector machines perform well, particularly when utilizing mean and worst values of features, indicating their effectiveness in capturing key patterns in the data.

Conversely, the K-nearest neighbors model exhibits a notable decrease in performance when standard deviation features are employed, implying potential limitations in leveraging variability information for classification.

Decision trees and random forest models maintain relatively stable performance across feature representations, though slight variations suggest sensitivity to feature types.

Interestingly, the Random Forest model performed slightly worse than the Decision Trees model across all feature representations in terms of accuracy, sensitivity, specificity, and AUC. Random Forest might have overfit the training data more than Decision Trees due to its inherent complexity, resulting in reduced generalization performance on the test set.

Mean of the features did produce the best results for all the models. However, the worst value result in better specificity value which indicates that the models were better able to identify negative case with this version of the features. These findings underscore the importance of feature selection and representation in optimizing model performance for breast tumor diagnosis, with logistic regression emerging as a robust choice across different feature sets.

A graph of different colored lines

Description automatically generated with medium confidence

A SHAP evaluation was conducted to assess the performance of the top-performing model identified through prior evaluations. The provided diagram represents a beeswarm plot generated for the support vector model. Each data point on the plot corresponds to an individual prediction, with the color of the point indicating the respective feature value—where red denotes high values and blue signifies low values. The x-axis portrays the feature value, while the y-axis delineates the feature name.

The visualization elucidates the relative magnitude of SHAP (Shapley Additive Explanations) values associated with each feature, thereby highlighting those exerting notable influence on the model's predictions. For instance, 'Mean\_CPts' exhibits the highest positive SHAP value, indicating its significance in determining the outcome. Conversely, 'Mean\_Tex' displays the lowest negative SHAP value, suggesting its potential to diminish the likelihood of the predicted outcome.

Furthermore, the plot reveals the functional relationship between feature values and SHAP values. Notably, an increase in 'Mean\_A' values corresponds to an increase in SHAP values, while an increase in 'Mean\_FractDim' values correlates with a decrease in SHAP values.

Additionally, the visualization underscores features deemed inconsequential in the prediction process. For instance, lower values of 'Mean\_Comp' are depicted as having negligible impact on the outcome of predictions, thereby highlighting their marginal relevance in the model's decision-making process.

**4. Conclusion**  
This study offers a thorough examination of machine learning models for breast tumor diagnosis, providing insights into their performance across various feature sets. Our findings underscore the critical role of feature selection and representation in classification tasks. Notably, Support Vector Machines consistently outperformed other models across different datasets, aligning with prior research in this domain. Further investigation is required to investigate the factors contributing to this superiority. Additionally, while the mean of features yielded optimal results across most evaluation metrics, employing the worst feature values notably enhanced specificity, enhancing the identification of negative cases, potentially attributed to their extreme values. Exploring the optimal data characteristics for desired outcomes warrants further inquiry. The study also suggests that while the models performed well without hyperparameter settings, such an approach may not be justifiable given the marginal improvement and resource consumption. However, hyperparameter tuning could notably enhance ensemble models like Random Forests. Tailoring hyperparameters for individual healthcare institutions warrants exploration. Moreover, enhancing diagnosis efficacy could be achieved through larger datasets, suggesting avenues for future research.

**References**

[1] World Health Organization (WHO). (2021). Breast Cancer: Prevention and Control. Retrieved from https://www.who.int/cancer/detection/breastcancer/en/

[2] Dua, S., & Graff, C. (2019). Breast Cancer Detection Using Machine Learning Techniques: A Review. Journal of Medical Systems, 43(1), 5.

[3] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 68(6), 394-424.

[4] American Cancer Society. (2021). Breast Cancer Facts & Figures 2021-2022. Retrieved from https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/breast-cancer-facts-and-figures-2021-2022.pdf

[5] Siegel, R. L., Miller, K. D., & Jemal, A. (2020). Cancer statistics, 2020. CA: A Cancer Journal for Clinicians, 70(1), 7-30.

[6] Duffy, S. W., Nagtegaal, I. D., Wallis, M., Cafferty, F. H., Houssami, N., Warwick, J., ... & Tabár, L. (2021). Correcting for lead time and length bias in estimating the effect of screen detection on cancer survival. International Journal of Epidemiology, 50(5), 1557-1566.

[7] Nelson, H. D., Fu, R., Cantor, A., Pappas, M., Daeges, M., & Humphrey, L. (2016). Effectiveness of breast cancer screening: systematic review and meta-analysis to update the 2009 US Preventive Services Task Force recommendation. Annals of Internal Medicine, 164(4), 244-255.

[8] van Ravesteyn, N. T., Stout, N. K., Schechter, C. B., Heijnsdijk, E. A., Alagoz, O., Trentham-Dietz, A., ... & Mandelblatt, J. S. (2016). Benefits and harms of mammography screening after age 74 years: model estimates of overdiagnosis. JAMA oncology, 2(12), 1569-1577.

[9] Guan, Y., Liu, Y., Chen, Y., & Chen, G. (2020). Diagnosis of breast cancer using ensemble deep learning algorithms based on full-field digital mammography. BioMed Research International, 2020.

[10] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

[11] Bennett, K. P., & Mangasarian, O. L. (1992). Robust Linear Programming Discrimination of Two Linearly Inseparable Sets. Optimization Methods and Software, 1(1), 23-34.

[12] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

[13] Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. Springer Science & Business Media.

[14] Tukey, J. W. (1977). Exploratory data analysis. Addison-Wesley.

[15] Schloerke, B., Crowley, J., Cook, D., Hofmann, H., Wickham, H., Briatte, F., ... & Larmarange, J. (2018). GGally: Extension to 'ggplot2'. R package version 1.4.0.

[16] Wei, T., & Simko, V. (2017). R package "corrplot": Visualization of a correlation matrix (version 0.84). Journal of Statistical Software, 46(4), 1-14.

[17] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3(Mar), 1157-1182.

[18] Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202.

[19] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112). New York: springer.

[20] Kuhn, M. (2008). Building predictive models in R using the caret package. Journal of statistical software, 28(5), 1-26.

[21] Provost, F., & Fawcett, T. (2013). Data science for business: What you need to know about data mining and data-analytic thinking. O'Reilly Media, Inc.

[22] Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13(Feb), 281-305.

[23] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of eugenics, 7(2), 179-188.

[24] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.