Comparative Analysis of Machine Learning Models in Breast Cancer Detection

Bahadir Bagci
bahadirbagci@gmail.com
https://www.linkedin.com/in/bahadirbagci

March 14, 2024

https://sophisticated21.github.io

I. Introduction

IA. The Pivotal Role of Breast Cancer Research in Medical Advancements

As per the World Health Organization's report in 2023, breast cancer is defined by the unchecked proliferation of irregular cells within the breast, resulting in the development of tumors. If left unattended, these tumors have the capacity to spread to different areas of the body, potentially leading to a life-threatening situation.

Breast cancer researchs are critical for advancing our understanding of the disease and enhancing patient outcomes. It plays a pivotal role in sharing current research findings, improving patient accrual, and deepening our knowledge of the basic mechanisms underlying breast cancer (Piccart & Winer, 2004). The expansion of breast cancer researchs, aided by communication networks and advocacy organizations, underscores the importance of early detection and treatment (Garfinkel & Stellman, 1997).

IB. The Integration of Machine Learning in Breast Cancer Research: A New Frontier

Machine learning is increasingly integral in breast cancer researchs, enhancing classification accuracy and improving disease prediction. For instance, machine learning approaches, including pre-trained networks, have been pivotal in improving classification accuracy in breast cancer histopathology images (Sharma & Mehra, 2020). Machine learning also aids in early-stage breast cancer prediction, with some algorithms achieving up to 100% accuracy (Shilpa et al., 2022). These advancements highlight the role of machine learning in enhancing disease prediction accuracy, sensitivity, and overall diagnostic performance in breast cancer research.

IC. Objective And Reseach Questions

Our objective is to evaluate the performance and make comparisons among four different machine learning models, along with various hyperparameters, to determine their suitability for our dataset. We aim to answer the following three fundamental questions:

1. Among Logistic Regression, SVM (Support Vector Machine), KNN (K-Nearest Neighbors), and Random Forests, which model is more effective in breast cancer detection?

2. In cases with multiple outliers, which machine learning model demonstrates superior accuracy in breast cancer detection?

3. Which features play a more significant role in this detection process?

Through this analysis, we aim to uncover insights that can enhance the accuracy and reliability of breast cancer detection using machine learning techniques.

II. Methodology

In this project, we conducted a comprehensive analysis of the breast cancer dataset, which was created by professors from the University of Wisconsin. Our analysis involved the following steps:

1. **Data Preprocessing**
    - The dataset used in this study was created by Dr. William H. Wolberg from the General Surgery Department, W. Nick Street from the Computer Sciences Department, and Olvi L. Mangasarian from the Computer Sciences Department, all affiliated with the University of Wisconsin. The dataset was designed to provide insights into breast cancer diagnosis based on features derived from digitized images of fine needle aspirates (FNAs) of breast masses.
    - The features in the dataset were computed from FNA images using advanced image processing techniques. This involved extracting relevant information from images to characterize the cell nuclei visible in the samples.
    - The dataset includes a total of 32 attributes, which encompass an ID number, diagnosis labels (M for malignant and B for benign), and 30 real-valued features related to cell nucleus characteristics. These features were carefully selected after an exhaustive search in the feature space, considering different combinations of 1-4 features and 1-3 separating planes.
    - We began by cleaning the dataset using Python libraries such as Pandas and NumPy to handle missing data, outliers, and any other data quality issues.

2. **Exploratory Data Analysis**
    - We conducted exploratory data analysis (EDA) to gain insights into the dataset's characteristics and distribution of features.
    - We meticulously examined specific features in the breast cancer dataset due to their pivotal role in improving the diagnostic accuracy and distinction between benign and malignant tumors. These features were selected based on insights from reputable sources such as

(Allada et al., 2021) and (Kopans, 1986), which highlighted their significance in cancer diagnosis.

- Essential Parameters:
  - Radius, Texture, Perimeter, Area, Smoothness, Concavity, Compactness
  - Importance: These parameters play a crucial role in enhancing the precision of breast cancer diagnosis. (Allada et al., 2021); (Kopans, 1986).
- Additional Key Features:
  - Concave Points and Symmetry
  - Role: These features contribute to improving the accuracy of classifying breast masses. (Analysis of Breast Cancer dataset, 2020); (Mashudi et al., 2021).
- Fractal Dimension:
  - Significance: The fractal dimension is indispensable for cytologic diagnosis and characterizing mammographic masses.
  - Effectiveness: It complements other shape factors, leading to enhanced classification accuracy. (Einstein et al., 1998); (Shanmugavadivu et al., 2016).

3. **Oversampling**
   - To address class imbalance in the dataset, we performed oversampling to ensure a balanced representation of benign and malignant cases.
4. **Model Selection and Hyperparameter Tuning**
   - We selected four different machine learning techniques and used grid search to find the best hyperparameters for each technique.
5. **Performance Evaluation**
   - We evaluated the performance of the selected machine learning models on both the test set and extreme cases.
6. **Results and Visualization**
   - We summarized the results of our analysis and presented them in a clear and interpretable manner.

By following these steps, we aimed to provide valuable insights into the breast cancer dataset and assess the effectiveness of different machine learning algorithms in classifying malignant and benign cases. Our analysis was conducted using Python's Pandas, NumPy, Matplotlib, Seaborn libraries, and the scikit-learn library for machine learning tasks.

III. Findings

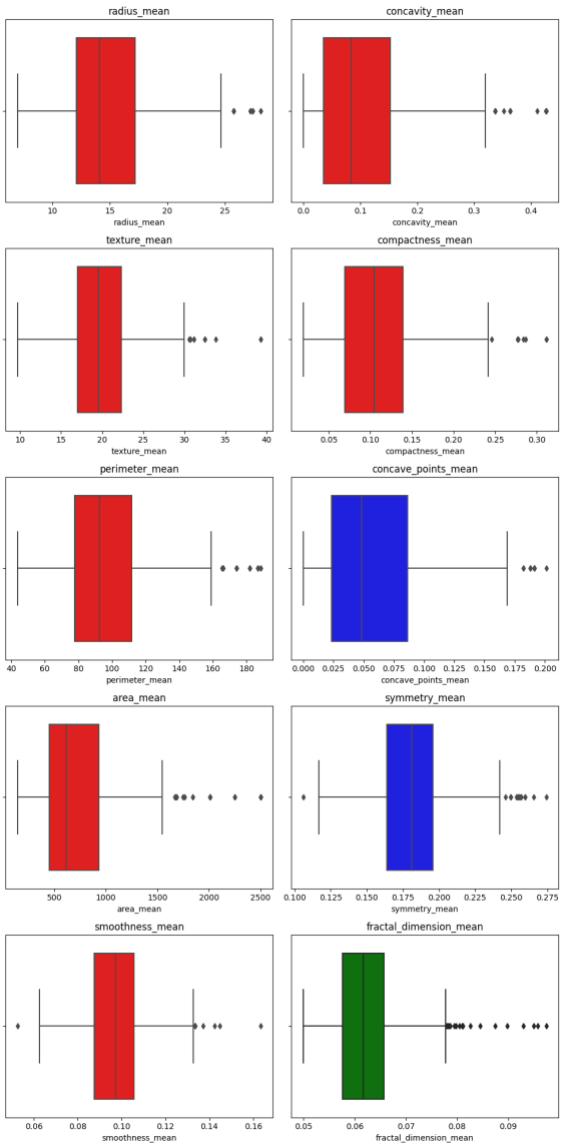


Image 1: Boxplots

Red Boxplots (Essential Features):

- **Radius_mean, Texture_mean, Perimeter_mean:** High median value with a relatively wide IQR and presence of many outliers.
- **Area_mean:** Displays a wide range of values with several outliers, highlighting significant differences in tumor areas.
- **Smoothness_mean:** Less variability than size-related features.
- **Concavity_mean:** Wide IQR and many outliers indicate varied tumor concavity.
- **Compactness_mean:** High variability and several outliers suggest differences in tumor density.

Blue Boxplot (Additional Key Feature):

- **Concave_points_mean:** Notably high variability and numerous outliers, indicating that the number of concave points varies significantly across tumors.
- **Symmetry_mean:** Shows a narrower, suggesting more consistency in tumor symmetry.

Green Boxplot (Feature with Specific Importance):

- **Fractal_dimension_mean:** Also demonstrates a narrower IQR, indicating less variability in this measure of tumor roughness or complexity.

Clinical Relevance:
The essential features (in red) like size and concavity appear to have the most variability and could potentially be the most informative for distinguishing between different types of breast tumors, which aligns with their designation as "essential" in the analysis.

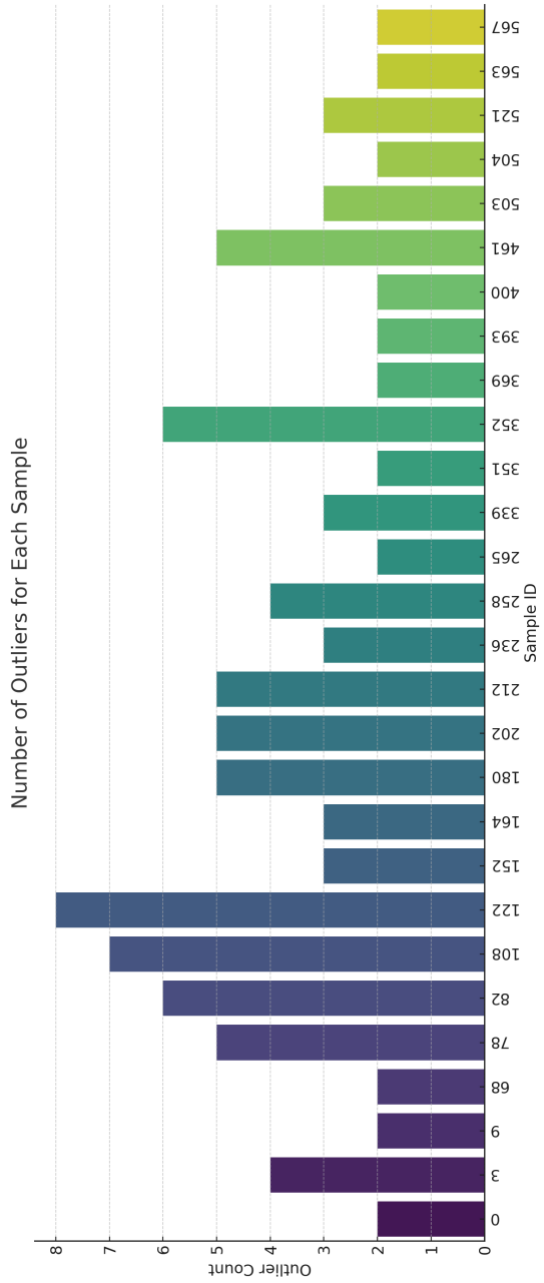


Image 2: The Number of Outliers for Each Sample

Number of Outliers for Each Sample

'The Number of Outliers for Each Sample' visual shows the number of cases in our dataset that have multiple outliers across the selected features, and the count of outliers in each of these cases:

- The first number (28) indicates that there are a total of 28 distinct cases in our dataset. These cases have outliers in at least two of the selected features.
- The following list shows the total number of outliers in each case. For example, the first case (index 0) has two outliers, the case with index 258 has four outliers, and so on.

In summary, this analysis reveals that certain cases in our dataset have multiple outliers across different features, indicating that these cases might be significantly different from others and may require special attention. We saved these cases separately to later assess how well our machine learning model predicts them.

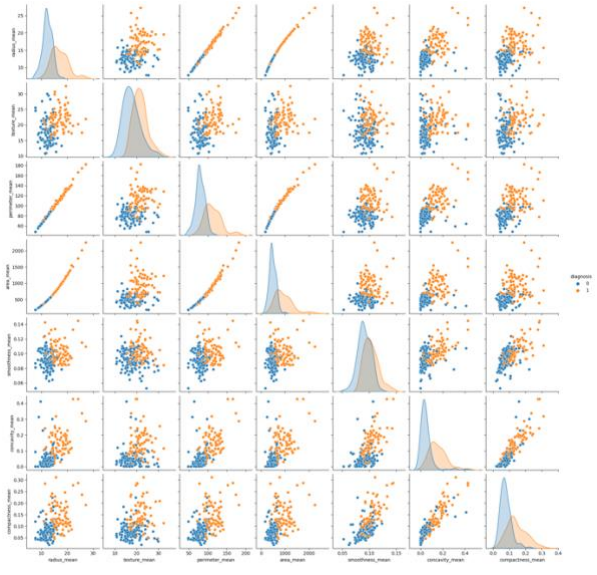


Image 3: Pairplot 1(Essential Features)

Observations From The Pairplot 1

Size-Related Features: Features such as radius_mean, perimeter_mean, and area_mean display a distinct separation between benign and malignant cases. Malignant tumors tend to have higher values for these features, which is indicative of larger and potentially more aggressive tumors.

Feature Correlations: There's a strong positive correlation between features that are geometrically related, like radius_mean and perimeter_mean. This suggests that as the size of the tumor increases, so does its perimeter and area, which aligns with expected geometric principles.

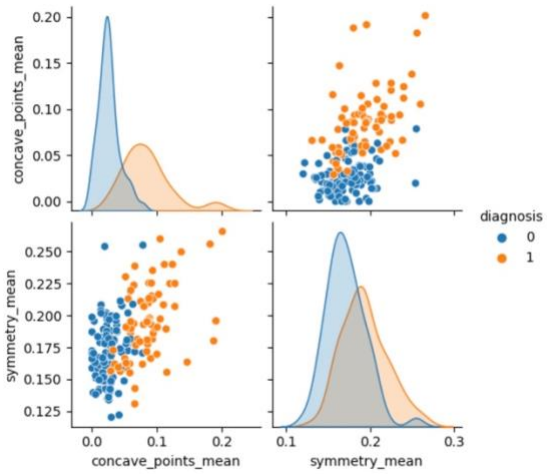
Texture and Smoothness: The texture (potentially represented by texture_mean) does not seem to distinguish as clearly between benign and malignant as the size-related features do, given the greater overlap in the scatter plots. Similarly, smoothness_mean shows some overlap between the classes but also hints that **malignant tumors might have a tendency towards higher values.**

Shape Features: Concavity_mean and compactness_mean demonstrate that **malignant tumors often have more irregular shapes and are less uniform**, as indicated by higher values in these features for the malignant class

Distribution Patterns: Looking at the density plots on the diagonal, malignant tumors (orange) have a wider spread in most features, implying a higher variance in the malignant class compared to the benign class.

Potential Outliers: There are several data points that stand out from the main clusters, especially in the area_mean feature, suggesting the presence of outliers which could represent unusual cases. Hence, we measured our ML models on this unusual cases.

Clinical Relevance: Clinically, this visualization underscores the importance of size, shape, and texture characteristics in differentiating between benign and malignant breast tumors. **Larger, less uniform, and more irregular tumors are more likely to be malignant.**



Observations From The Pairplot 2

Concave Points Distribution: Benign tumors show lower concave_points_mean, while malignant ones have a broader distribution, suggesting **more concave points typically indicate malignancy.**

Symmetry Variation: Symmetry_mean overlaps for both tumor types, with malignant ones displaying a broader spread, hinting at higher variability.

Concave vs. Symmetry Dynamics: A non-linear relationship is observed, with malignant tumors generally exhibiting higher concave points, without a strong linear correlation with symmetry_mean.

Aggressiveness Indicator: Malignant tumors show greater variation in concave_points_mean, potentially marking tumor aggressiveness.

Clinical Implications: Concavity and symmetry metrics are crucial for tumor analysis. Elevated concave points are often associated with malignancy, as reflected in the data.

Analysis of Heatmap

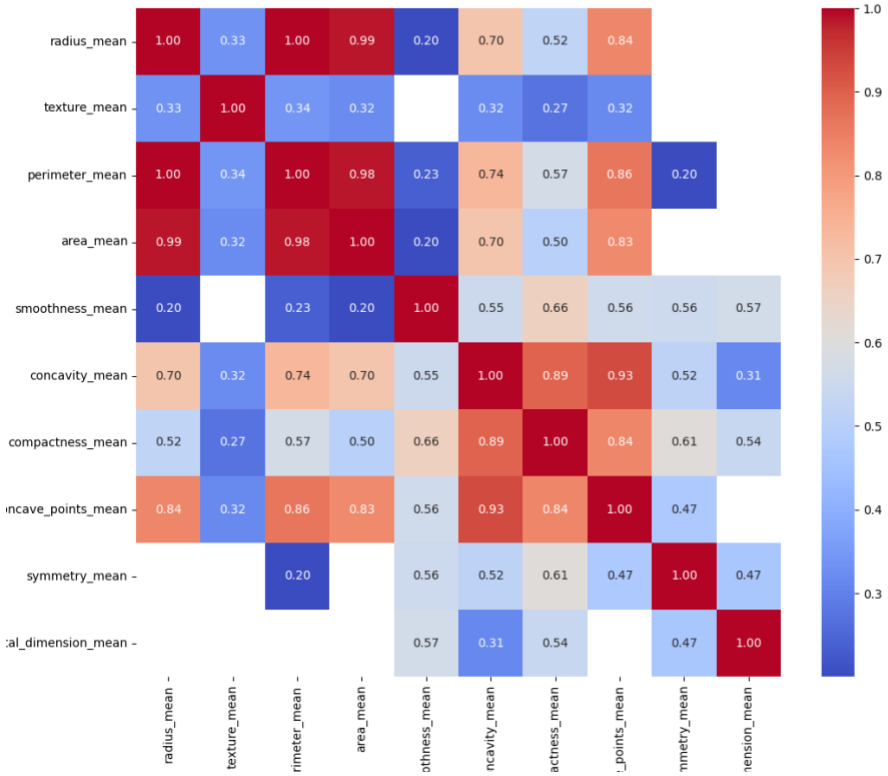


Image 4: Heatmap

The heatmap visualization indicates strong correlations among the selected features. High correlation means that some features are strongly related and may contain redundant information. This could lead to multicollinearity issues in machine learning models.

To address this, there are several approaches could be considered

Feature Selection: We could choose to remove some of the highly correlated features to train our model with fewer but more effective predictors.

Principal Component Analysis (PCA): This technique reduces the dimensionality of our data, decreasing multicollinearity and potentially improving our model's generalization.

Regularization: Techniques such as Ridge (L2 regularization) or Lasso (L1 regularization) can mitigate multicollinearity by penalizing the model for complexity and pushing coefficients of redundant features towards zero.

We proceed with implementing Principal Component Analysis to enhance our model's performance.

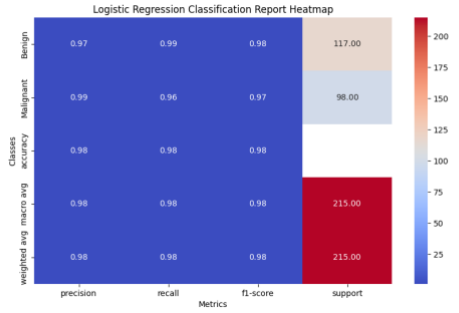


Image 5: Logistic Regression(Test set)

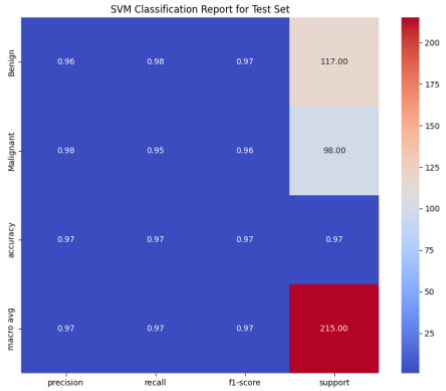


Image 6: SVM(Test set)

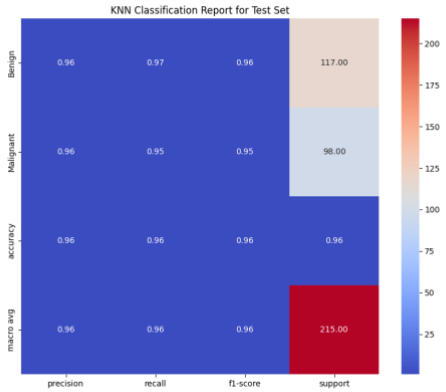


Image 7: KNN(Test set)

Image 8: Random Forest(Test set)

Comparison of Different ML Models on Test Set

The heatmaps provided represent the classification reports for different machine learning models, showcasing their performance metrics on a test set. The color intensity in each cell corresponds to the performance score for each metric (precision, recall, f1-score, support) for the Benign and Malignant classes, as well as overall accuracy, macro average, and weighted average scores.

Comparing Performances:

- **Highest Precision**: Logistic Regression(C: 0.088, penalty: l2), for the Malignant class(0.99).
- **Highest Recall**: Logistic Regression for both classes(0.99 and 0.96)
- **Highest F1-Score**: Logistic Regression for both classes.(0.98 and 0.97)
- **Support**: The number of instances for each class remains the same across all models.
- **Highest Overall Accuracy**: Logistic Regression(0.98).

In summary, Logistic Regression slightly outperforms the other models in terms of overall accuracy and recall for the Malignant class, which is crucial in medical diagnostics to reduce the risk of missing a cancer diagnosis. The SVM follows closely, with a minor trade-off in recall for the Malignant class. KNN maintains a balanced performance but with slightly lower accuracy, while the Random Forest has comparable precision but marginally lower recall and f1-scores. The choice between these models may depend on the specific clinical context and the relative importance of each performance metric.

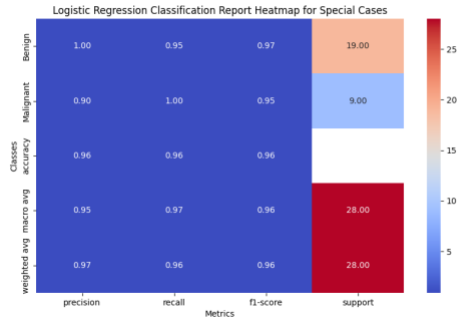


Image 9: Logistic Regression(Special cases)

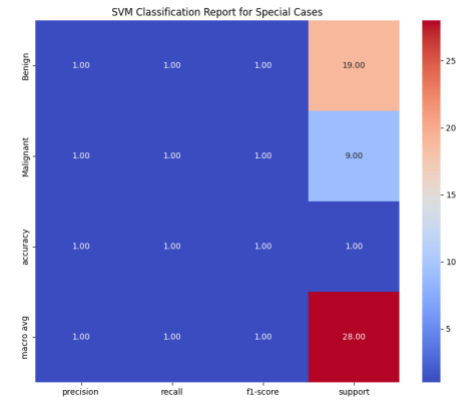


Image 10: SVM(Special Cases)

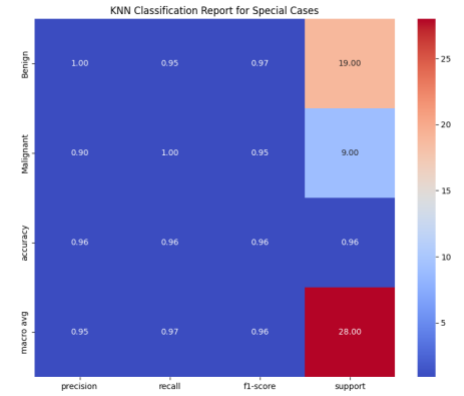


Image 11: KNN(Special cases)

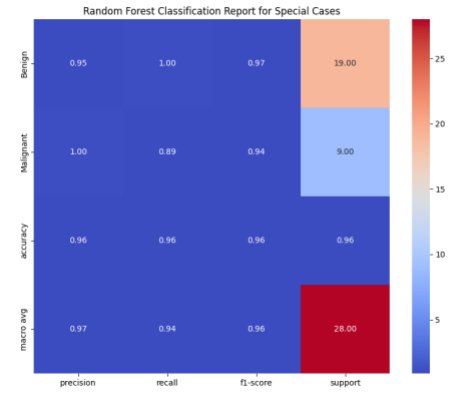


Image 12: Random Forest(Special cases)

Comparison of Different ML Models on Special Cases

- **SVM(with C: 545.55, gamma: 0.012, kernel: rfb) stands out with perfect scores across all metrics, indicating that it correctly classified all special cases.**
- **Logistic Regression** and **KNN(with 3 neighbors)** show identical precision, recall, and f1-scores for the Malignant class, indicating a high sensitivity (no false negatives).
- **Random Forest(max_depth: 30, min_samples_split: 2, n_estimators: 100** has a slightly lower recall for the Malignant class (0.89) compared to Logistic Regression and KNN, suggesting it might have missed one Malignant case (a potential false negative).
- The **support** indicates that there were more Benign cases (19) than Malignant (9) in the special cases set.
- The **accuracy** of SVM is the highest, at 1.00, indicating perfect performance. Logistic Regression, KNN, and Random Forest share the same accuracy of 0.96, which is also very high.

In practical terms, the **SVM** model provided the best performance on this set of special cases, potentially making it the most reliable for this specific scenario. However, for real-world applications, one would also consider factors like model interpretability, computational cost, and performance across different datasets before choosing the best model.

IV. Conclusion

In this report, we have conducted a thorough analysis of a breast cancer dataset, focusing on the delineation of tumor characteristics and the performance of various machine learning models. The critical findings and their implications are summarized as follows:

Key Feature Insights:

- Size-related features, including the mean radius, perimeter, and area, are pivotal in distinguishing between benign and malignant tumors, with larger values typically indicating malignancy.
- Texture and smoothness, while informative, offer less clear differentiation between tumor types compared to size-related features.
- Shape-related attributes, specifically concavity and compactness, highlight that malignant tumors tend to have more irregular and denser structures.

Outlier Analysis:

- The identification of cases with multiple outliers across different features suggests the presence of unique or severe tumor presentations. These cases were isolated for specialized analysis, acknowledging their potential complexity and need for advanced diagnostic approaches.

Model Performance Evaluation:

- Logistic Regression demonstrated high precision and recall rates on the test set, indicating its effectiveness in correctly identifying malignant cases, which is critical to minimizing false negatives in medical diagnosis.
- SVM showed exemplary performance on special cases, achieving perfect precision and recall. This suggests SVM's superior capability in handling complex or atypical tumor presentations.
- KNN and Random Forest models also delivered commendable performance, though slightly trailing behind Logistic Regression and SVM in certain metrics.

Clinical Relevance:

- The analysis underlines the clinical importance of accurately measuring and interpreting tumor features, such as size, shape, and texture, for effective diagnosis and treatment planning.
- The variability in feature presentation, particularly in cases with multiple outliers, emphasizes the necessity for personalized diagnostic strategies.

The comprehensive evaluation reveals that while all the tested models perform well, Logistic Regression and SVM stand out in their respective areas of the test set and special cases. These insights not only contribute to the advancement of machine learning applications in breast cancer research but also underline the need for integrating these analytical tools within clinical practices for enhanced patient care. The choice of model in a clinical setting should, therefore, be informed by the specific diagnostic context, prioritizing accuracy, sensitivity, and the ability to handle diverse presentations of breast tumors.

Our code and notebook can be found at: “https://github.com/sophisticated21/Breast-Cancer-ML.git”

References

1. Allada, A., Rao, G., Chitturi, P., Chindu, H., Prasad, M., & Tatineni, P. (2021). Breast Cancer Prediction using Deep Learning Techniques. 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 306-311. https://doi.org/10.1109/ICAIS50930.2021.9395793
2. Einstein, A., Wu, H., & Gil, J. (1998). Self-Affinity and Lacunarity of Chromatin Texture in Benign and Malignant Breast Epithelial Cell Nuclei. Physical Review Letters, 80, 397-400. https://doi.org/10.1103/PHYSREVLETT.80.397
3. Garfinkel, L., & Stellman, S. (1997). Breast cancer and Women & Health. Women & Health, 26(1), 1-5. https://doi.org/10.1300/J013V26N01_01
4. Kopans, D. (1986). Breast cancer detection. The Western journal of medicine, 144(1), 73-76. https://doi.org/10.5580/1916
5. Mashudi, N., Rossli, S., Ahmad, N., & Noor, N. (2021). Breast Cancer Classification: Features Investigation Using Machine Learning Approaches. International Journal of Integrated Engineering. https://doi.org/10.30880/ijie.2021.13.05.012
6. Piccart, M., & Winer, E. (2004). Introducing Breast Cancer Research's updates on clinical trials. Breast Cancer Research, 6, 164. https://doi.org/10.1186/bcr810
7. Sharma, S., & Mehra, R. (2020). Conventional Machine Learning and Deep Learning Approach for Multi-Classification of Breast Cancer Histopathology Images—a Comparative Insight. Journal of Digital Imaging, 33, 632-654. https://doi.org/10.1007/s10278-019-00307-y
8. Shanmugavadivu, P., Sivakumar, V., & Sudhir, R. (2016). Fractal dimension-bound spatio-temporal analysis of digital mammograms. The European Physical Journal Special Topics, 225, 137-146. https://doi.org/10.1140/EPJST/E2016-02615-X
9. Shilpa, K., Adilakshmi, T., & Chitra, K. (2022). Applying Machine Learning Techniques To Predict Breast Cancer. 2022 Second International Conference on Interdisciplinary Cyber Physical Systems (ICPS), 17-21. https://doi.org/10.1109/ICPS55917.2022.00011
10. (2020). Analysis of Breast Cancer dataset using Supervised Machine Learning Classifiers. International Journal of Recent Technology and Engineering. https://doi.org/10.35940/ijrte.f1030.0386s20