Breast Cancer Diagnosis Using The Breast Cancer Wisconsin State Dataset

Introduction/Background

Advances in data-driven medical research have facilitated the application of machine learning (ML) in healthcare, especially for early detection and diagnosis of diseases. This project utilizes the Breast Cancer Wisconsin State dataset, a dataset for breast cancer diagnosis that provides vital information on tumor characteristics from biopsies. It enables a binary classification task to differentiate between malignant and benign tumors based on diagnostic features derived from cell nuclei measurements.

Dataset

The Breast Cancer Wisconsin State dataset contains several hundred samples, each with features such as clump thickness, uniformity of cell size, mitosis presence, and other diagnostic features. Each sample is labeled as either benign or malignant, forming a balanced dataset suitable for training ML models. The dataset, available on Kaggle, is structured to allow for straightforward feature extraction and model training without redundancy or bias issues. Its comprehensive feature set aids in understanding nuanced characteristics essential for accurate cancer detection.

Dataset Link:

Breast Cancer Wisconsin State Dataset

Problem Definition

Problem

The challenge in this project is to automate the diagnosis of breast cancer using the dataset's attributes, which will mimic the diagnostic process pathologists perform. Manual interpretation of biopsy data is time-intensive and can be prone to errors, potentially impacting patient outcomes.

Motivation

Developing an automated, ML-based diagnostic system could significantly reduce the workload on healthcare professionals, providing accurate and timely breast cancer diagnoses. Leveraging this dataset aligns with the growing trend of using ML to improve clinical care, and could ultimately contribute to more effective early intervention strategies for breast cancer patients.

Literature Review

- **Esteva et al. (2017):** Emphasizes the potential of deep learning to enhance diagnostic accuracy and efficiency in healthcare, supporting the project's objective to automate cancer detection with minimal human intervention. By integrating SVMs with deep learning approaches, researchers can leverage the strengths of predictive analysis to improve overall performance and address prevalent challenges currently in medical imaging data.
- **V. Romeo et al. (2022):** Underscores the effectiveness of combining multiple diagnostic features and machine learning algorithms, to enhance the accuracy of cancer classification.
- **I. Madakkatel et al.(2023):** The study's use of logistic regression to identify cancer risk showcases the model's capability to handle complex datasets and extract meaningful insights, directly aligning with our goal. Addresses comprehensive data preprocessing techniques, providing valuable information regarding feature selection processes.

Methods

Data Preprocessing

- **Data Cleaning:** Duplicate rows are removed to ensure unique entries, and rows with null values are deleted.
- **Noise Reduction:** Outliers in diagnostic attributes are identified and managed to minimize noise.
- **Feature Selection:** Visualized feature correlation through Seaborn heatmaps and drop the top correlated features (and irrelevant columns) to retain the most informative features in order to optimize model performance.

Supervised Machine Learning Algorithms

- **Support Vector Machines (SVM):** SVMs, known for their binary classification capabilities, will be employed for comparison. The core idea of this model is to find a hyperplane that best separates the two classes with the maximum margin. We chose to use SVMs because of their efficiency in cases where there is a clear margin of separation and ability to work well with high-dimensional data, like the several features of the breast cancer dataset. This is the model we have implemented for our midterm checkpoint.
- **Logistic Regression:** Logistic regression is a simple yet powerful linear model used for binary classification. It estimates the probability that a given input belongs to one of two classes. The output is then transformed using a logistic function, which maps the values to a range between 0 and 1. In medical contexts, logistic regression is particularly valuable because it helps explain the relationship between diagnostic features and the predicted outcome, making it easier to interpret. As an interpretable model, it is well-suited for predicting tumor malignancy by analyzing the contributions of various diagnostic features.
- **Random Forest:** Random Forest is an ensemble method that builds a collection of decision trees, with each tree trained on a random subset of data. The final prediction is made by averaging the outcome from all the trees. Randomness helps to reduce overfitting. This traditional ML algorithm will be used as a baseline to compare feature importance and model interpretability. By analyzing the trees, we can identify which features are most influential in determining the tumor type.

Results and Discussion

Quantitative Metrics

- **F1-Score:** This metric assesses the model’s balance between precision and recall, indicating how effectively it distinguishes between malignant and benign tumors.
- **Recall:** Measures the model’s sensitivity to true positives, crucial for ensuring malignant cases are detected.
- **Precision:** Indicates the accuracy of the model in identifying true positives, minimizing false positives.

Model Comparisons

Support Vector Machines (SVM)

- **Accuracy**: 0.9856

- **Cross-Validation Scores**: [0.9549, 0.9369, 0.9363, 0.9454, 0.9636]
- **Mean Cross-Validation Score**: 0.9475
- **F1-Score**: 0.9804

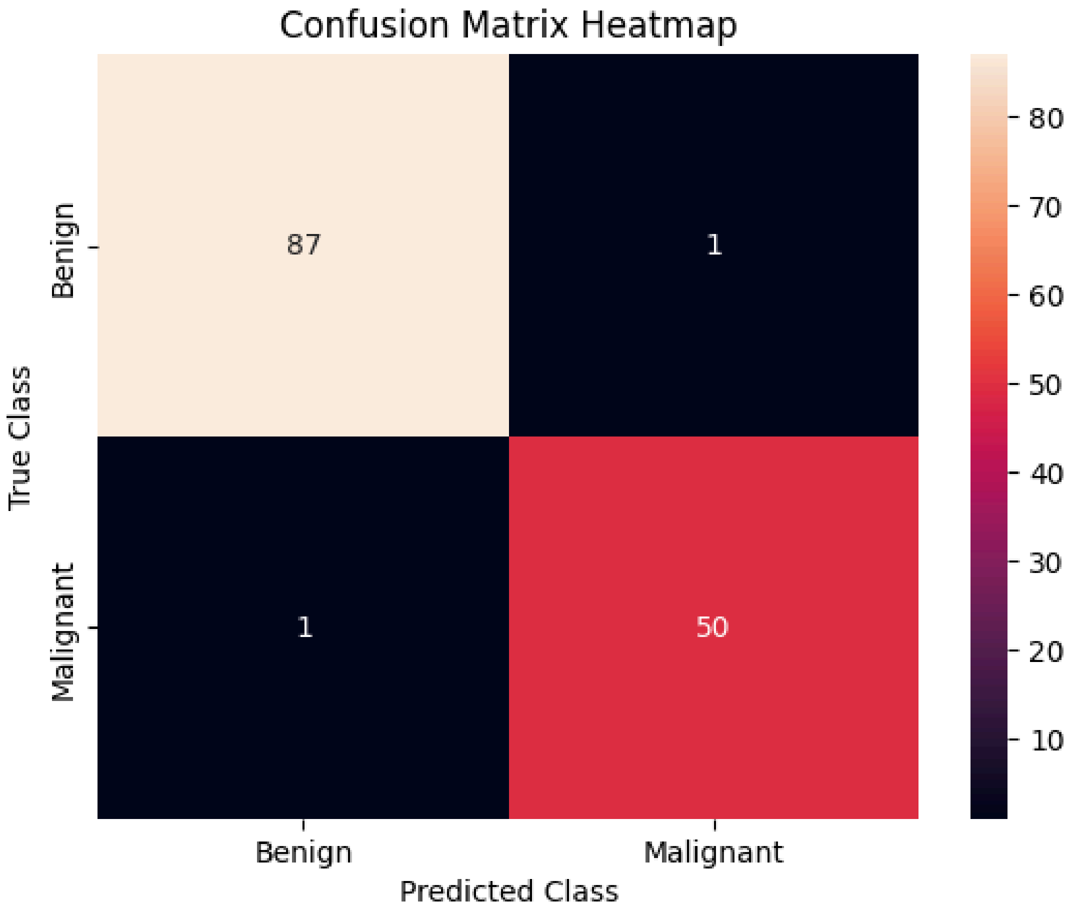
Strengths:

- SVM showed strong performance, as it has the highest accuracy among the three models.
- The method works particularly well with high-dimensional data and clearly defined class margins, which aligns well with the structure of the breast cancer dataset.

Limitations:

- SVMs are computationally expensive, especially with large datasets.
- It is harder to interpret the direct influence of individual features compared to other models.

Confusion Matrix Analysis:
The high recall shows that SVMs effectively detected malignant cases, which lowers the risk of false negatives.



Logistic Regression

- **Accuracy**: 0.9568
- **Precision**: 0.9362
- **Recall: 0**.97
- **ROC AUC score: 0**.9933
- **F1-Score**: 0.9933

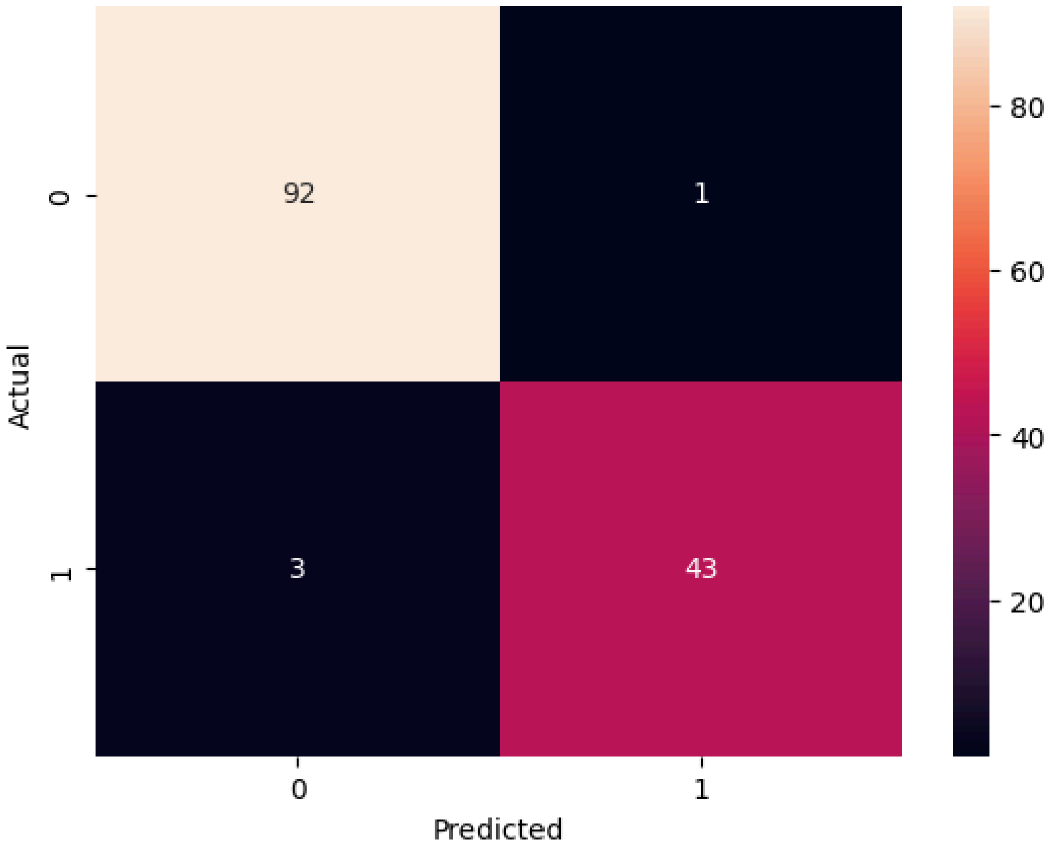
Strengths:

- Logistic Regression is simple and interpretable,which helps provide clear insights into the effects of individual features to the classification task.
- It is computationally efficient, which better suits smaller datasets or scenarios requiring quick predictions.

Limitations:

- Logistic Regression can underperform when it is compared to more complex models like Random Forest in capturing non-linear relationships within the data.
- It is sensitive to multicollinearity and needs well-preprocessed data for optimal performance.

Confusion Matrix Analysis:
The results show acceptable precision but lower recall compared to SVMs, indicating a slight trade-off in detecting malignant cases.



Random Forest

- **Accuracy**: 0.9712
- **Precision**: 0.9773
- **Recall:** 0.99
- **ROC AUC score:** 0.9556
- **F1-Score**: 0.9556

Strengths:

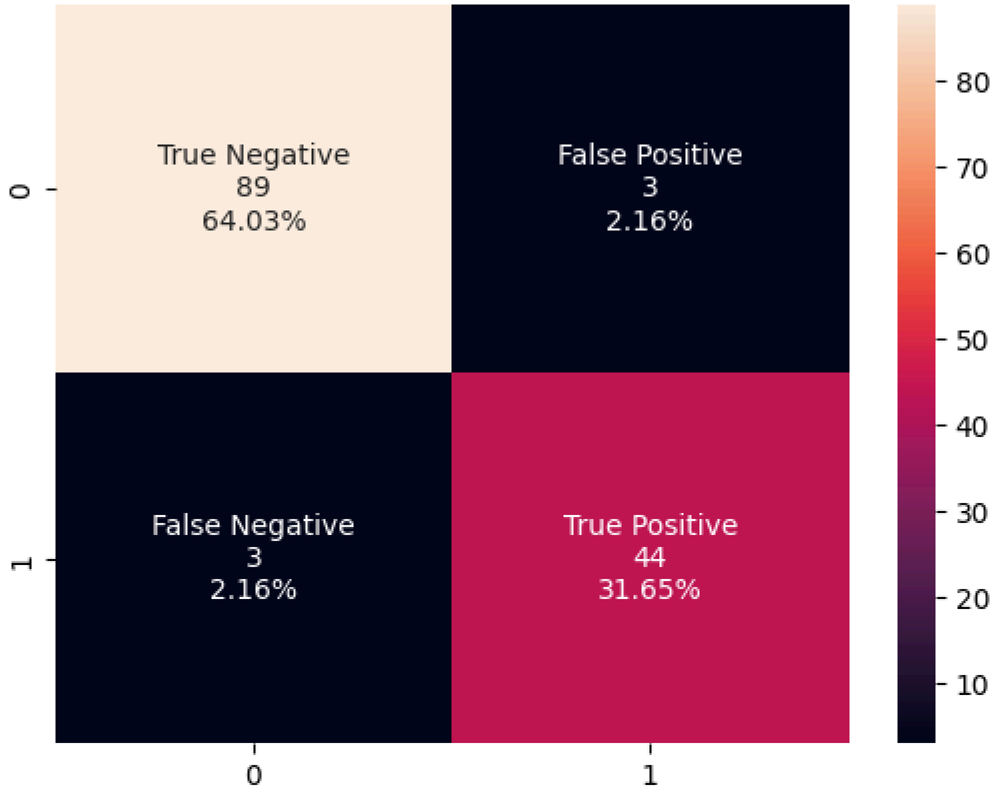
- Random Forest demonstrated robustness because it can handle noise and overfitting due to ensemble learning.
- The model's analysis on feature importance helps identify the more influential features, providing interpretability.

Limitations:

- SImilar to SVMs, Random Forest models can become computationally expensive with very large datasets.

- They might not always outperform simpler models like Logistic Regression when the dataset is relatively small.

Confusion Matrix Analysis:
Random Forest displayed balanced recall and precision, making it suitable for general-purpose tasks while still identifying important features for further analysis.



Confusion Matrix - Random Forest

Conclusions

Best Performing Model: SVM achieved the highest accuracy and recall, making it most effective for this task. Its robustness in high-dimensional data makes it a suitable choice for breast cancer diagnosis.

Trade-offs: While Logistic Regression provides interpretability, it lacks the non-linear flexibility of SVMs and Random Forest. Random Forest provides insights into feature importance but may be less computationally efficient.

Limitations: Each model has their own respective limitations, including time inefficiency, space inefficiency, or nonlinearity generalization. These issues all make each model limited for different problems or datasets.

Analysis of Algorithm/Model

The model’s effectiveness, reflected in high recall and precision, suggests strong performance in identifying cancerous cases, with misclassifications examined in the Confusion Matrix. Results highlight the importance of different features, where diagnostic factors contributed significantly to accuracy.

Next Steps

- We do not feel like there are necessary additional steps for this model.

Proposal Changes

Dataset Change

The initial dataset was too large for our Collab environment and would have required significant time and computational resources to process (even with GPU cores). Because of this we decided to switch to a smaller, more manageable dataset that could be pre-processed and trained more efficiently. This allowed us to proceed more effectively with our model training and testing.

Model Changes

We selected new models that were better suited for the smaller dataset. These models were chosen for their efficiency with reduced data, allowing us to maintain effective performance without sacrificing analysis quality or speed. Mainly instead of performing kernel image analysis using CNN techniques, we decided to switch to Logistic Regression and Random Forest Models, which are simpler but still effective at performing binary classification tasks.

References (IEEE format)

[1] A. Esteva, et al., “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, pp. 115-118, Feb. 2017. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/28778026/. [Accessed: Nov 8, 2024].

[2] V. Romeo et al., "AI-enhanced simultaneous multiparametric 18F-FDG PET/MRI for accurate breast cancer diagnosis," *Eur. J. Nucl. Med. Mol. Imaging*, vol. 49, no. 2, pp. 596–608, Jan 2022. [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=c9h&AN=154982388&site=ehost-live&scope=site. [Accessed: Nov 8, 2024].

[3] I. Madakkatel et al., "Hypothesis-free discovery of novel cancer predictors using machine learning," *Eur. J. Clin. Investig.*, vol. 53, no. 10, pp. 1–13, Oct. 2023. [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=c9h&AN=171852299&site=ehost-live&scope=site. [Accessed: Nov 8, 2024].