**Breast Cancer Classification: A Machine Learning Approach with** **Feature Extraction and SVM**

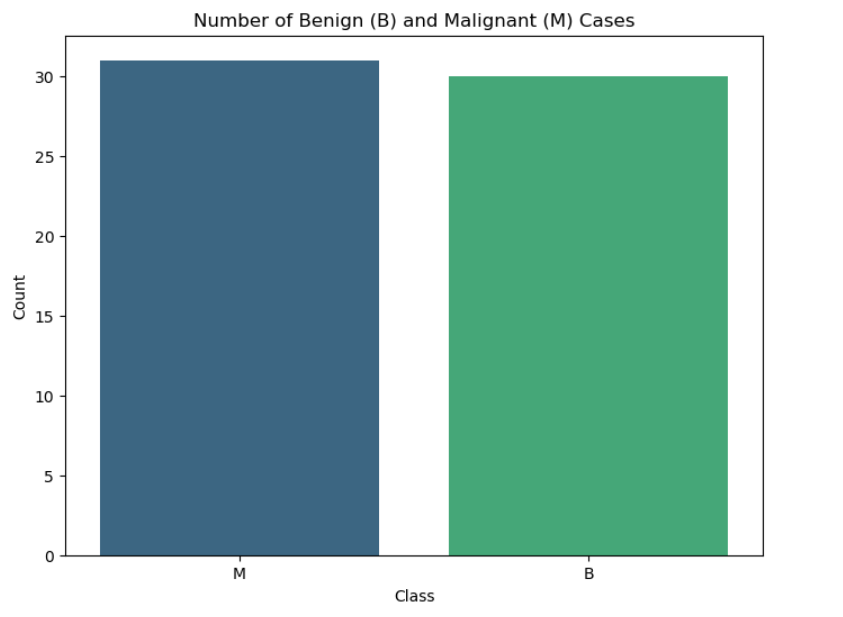
**Introduction:**

Breast cancer is the most prevalent cancer among women globally. Early detection is critical to improving patient outcomes. The idea of this research is to create a model for recognizing cancerous cells in medical photographs. Our goal is to extract significant features from these photos, choose the most important ones, construct a predictive model, and assess its effectiveness. We developed a model that included picture preprocessing, feature extraction, feature selection, SVM classification, and model evaluation.

**Data Acquisition:**

Two folders containing 31 benign and 30 malignant breast tissue images were used.

The specific file paths are mentioned in the code (benign\_folder and malignant\_folder).



**Image Preprocessing:**

The images were loaded using OpenCV (cv2) and preprocessed using grayscale (cv2.cvtColor) conversion and Gaussian blur (cv2.GaussianBlur) to reduce noise and enhance feature extraction accuracy.

**Feature Extraction:**

Shape and texture information is taken from the preprocessed photos. Area and perimeter are shape features that were calculated using contours (cv2.findContours, cv2.contourArea, cv2.arcLength), whereas the Gray Level Co-occurrence Matrix (GLCM) was utilized to capture textural patterns (feature.greycomatrix).acquire texture features such as contrast, dissimilarity, homogeneity, energy, and correlation.

**Feature Selection & Engineering:**

Each image's features were saved in a dictionary, together with a class name ('B' for benign, 'M' for malignant).

Features and labels were gathered from both folders and integrated into a single list.

A pandas DataFrame (df) was formed to store all features and labels.

This data frame has been saved as a CSV file (extracted\_features.csv) for future study.

The most informative characteristics were identified using recursive feature elimination (RFE) with an SVM classifier. Recursive Feature Elimination (RFE) is a machine learning feature selection technique that automatically selects a subset of the most relevant features (variables) for predicting the target variable.

"In addition to its capacity to choose relevant features, Recursive Feature Elimination (RFE) provides other benefits. First, it is a simple and systematic methodology that automates the feature selection process, eliminating the need for manual trial and error. Second, by iteratively evaluating feature importance and deleting irrelevant features, RFE reduces overfitting and enhances the generalization performance of machine learning models. Third, RFE can efficiently handle datasets with a large number of features, making it an excellent choice for high-dimensional data processing jobs. Finally, RFE sheds light on the significance of particular characteristics, which aids model interpretation and improves comprehension of underlying data patterns."

RFE reduces overfitting, improves model interpretability, and increases computational efficiency by focusing on the model's most informative characteristics. RFE chose the top four features based on their contribution to classification accuracy (RFE and SVC).

**Normalization:**

Min-Max scaling was applied to normalize all features between 0 and 1 (MinMaxScaler), ensuring that the scale of different features does not bias the model.

The use of Min-Max normalization for scaling features is justified for a variety of reasons. For starters, it preserves the original distribution of the data, guaranteeing that relative relationships between data points are maintained. Second, it is less sensitive to outliers than other normalizing algorithms, which is especially useful in medical datasets that are prone to outliers. Third, normalization to a common scale improves the SVM model's interpretability, making it easier to comprehend the importance of features. Fourth, Min-Max normalization allows for compatibility with SVM, which is sensitive to feature scales. Finally, its ease of use and computational efficiency make it ideal for huge datasets and iterative model tweaking operations. Overall, Min-Max normalization meets the needs of SVM classification while preserving data integrity, resilience, interpretability, and computational economy.

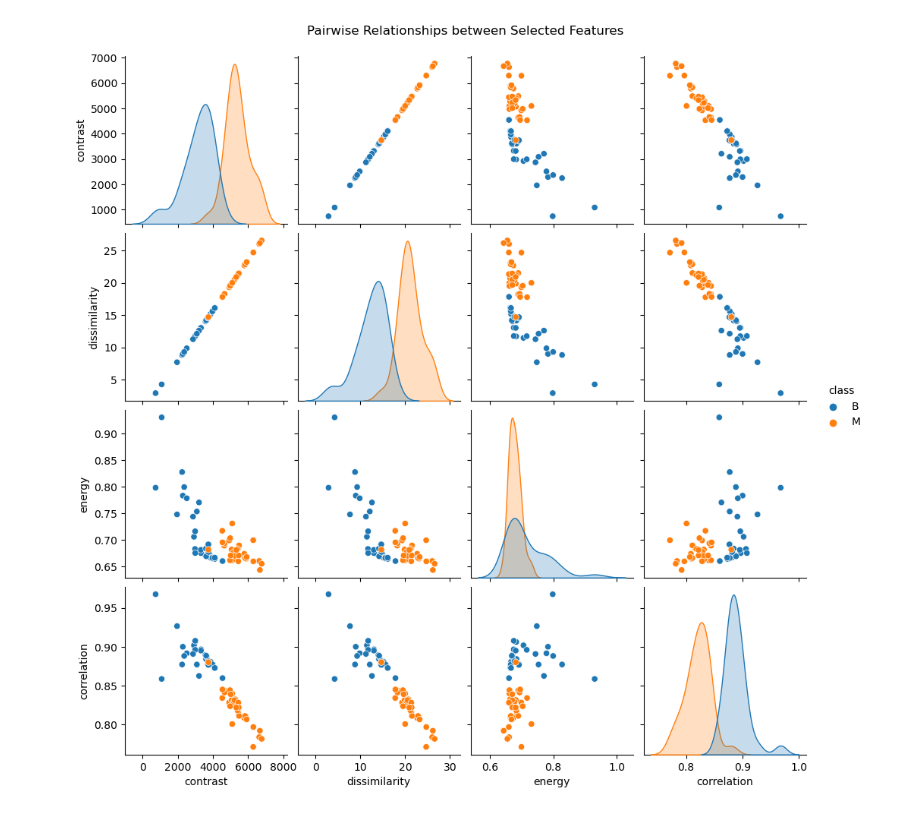
**Model Training and Evaluation:**

A linear Support Vector Machine (SVM) classifier was trained with normalized and carefully chosen features.

A linear kernel was utilized for the SVM classifier due to the potentially high number of features extracted from medical images. The default C value (regularization parameter) is used but might require tuning for better performance depending on the data. Finally, In the feature selection part using Recursive Feature Elimination (RFE), the number of features to select (n\_features\_to\_select) is set to 4.

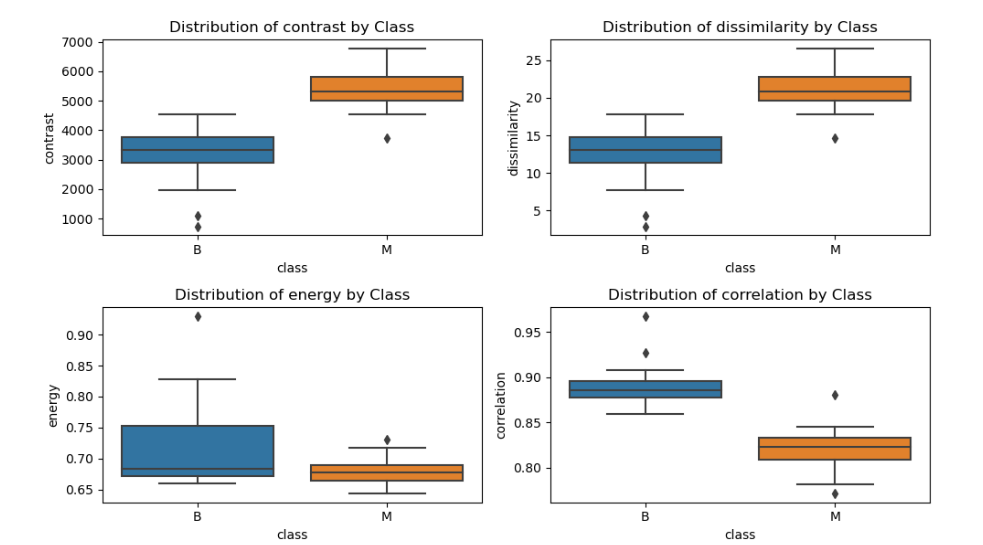
Model performance was extensively examined using a comprehensive assessment technique that included a confusion matrix and a classification report. The latter provides a more sophisticated view of the model's precision, recall, F1 score, and accuracy metrics. Notably, the SVM algorithm was chosen for its ability to manage high-dimensional information and identify complicated decision boundaries. The resulting model was meticulously scrutinized, with a special emphasis on its ability to reliably distinguish between benign and malignant cases, as demonstrated by the evaluation criteria.

**Relationships between pairs of features:**



The pairplot focuses on relationships between pairs of selected features.

The pair plots provide some initial clues about potential relationships between features and how they might differ based on the class (benign vs. malignant).



* The box plot shows the distribution of the top four features (contrast, energy, correlation, and dissimilarity) for benign and malignant breast tissue samples.
* The box for the benign class (blue) seems slightly lower than the malignant class (orange), suggesting that benign cells might have generally lower contrast compared to malignant cells.
* The dissimilarity distribution for the malignant class (orange) might be slightly wider than for the benign class (blue), potentially indicating a larger variation in intensity patterns within malignant cell nuclei.
* Both benign and malignant samples have a range of contrast values, with some overlap between the distributions.

**Results and Conclusion:**

The RFE method identified the following top 4 features for cancer detection:

* Energy
* Correlation
* Contrast
* Dissimilarity

The classification report presents a classification model's performance on a dataset with two classes (B and M). In summary, the model performs well across all measures, with precision, recall, and an F1-score of 0.97 for both classes, showing robust classification ability. Furthermore, the high accuracy of 97% indicates that the model can efficiently distinguish between the two classes.

The selected criteria, which include energy, correlation, contrast, and dissimilarity, have a high predictive ability for detecting malignant cells.

The SVM classifier, trained on these selected features, performs admirably in distinguishing benign and malignant cases.

Visual study of feature correlations and distributions adds to the significance of the selected features in cancer diagnosis.

**References:**Wang, H., & Zhang, Y. (2020). Feature Selection Based on Hybrid Optimization for Enhanced Brain-Computer Interface Classification. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 28(2), 480-490. <https://doi.org/10.1109/TNSRE.2020.2963774>

Source of BreaKHis: Breast Cancer Histopathological Database by Spanhol, F., Oliveira, L. S.,

Petitjean, C. and Heutte, L. is licensed under a Creative Commons Attribution 4.0 International

License.

https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/