Masters in Applied Statistics and Data Science (MASDS)
Department of Statistics
Jahangirnagar University
Savar, Dhaka-1342, Bangladesh.

Project-1
Course-Data mining

Submitted to
Dr. Md. Rezaul Karim

Submitted by:
Iqbal Habib
MASDS 9th Batch, Section-A Roll-**20229014**

Tumor Classification Using Machine Learning: A Comprehensive Analysis of Features and Model Performance

Abstract:

This study presents a comprehensive analysis of a dataset containing 32 variables, with a focus on building a machine learning model for accurately classifying tumors as benign or malignant based on tumor shape and geometry. The dataset comprises both categorical and continuous variables, with no missing values or outliers (identified through z-score analysis). Six different machine learning algorithms, namely Logistic Regression, Artificial Neural Network, Support Vector Machine, Random Forest, Decision Tree, and Naive Bayes, were employed for classification. Performance metrics including accuracy, precision, recall, specificity, F1 score, and AUC were assessed to evaluate the models' effectiveness. Results indicate high accuracy and promising performance for various models, with Logistic Regression and Artificial Neural Network leading the pack in terms of overall classification performance.

Introduction:

Cancer is a pervasive and potentially life-threatening disease that affects millions of people worldwide. Early and accurate diagnosis is crucial for effective treatment and improved patient outcomes. Advances in machine learning have opened new avenues for enhancing the accuracy of cancer diagnosis. In this study, we aim to explore a dataset containing 32 variables related to tumor characteristics, focusing on two essential aspects: tumor shape and geometry. Our primary objective is to build a machine-learning model that can accurately classify tumors as either benign or malignant based on these features.

The dataset under consideration contains a mix of categorical and continuous variables. We have taken meticulous steps to ensure data quality, including the removal of outliers using z-score analysis. We will discuss the data preprocessing steps and the rationale behind them in subsequent sections.

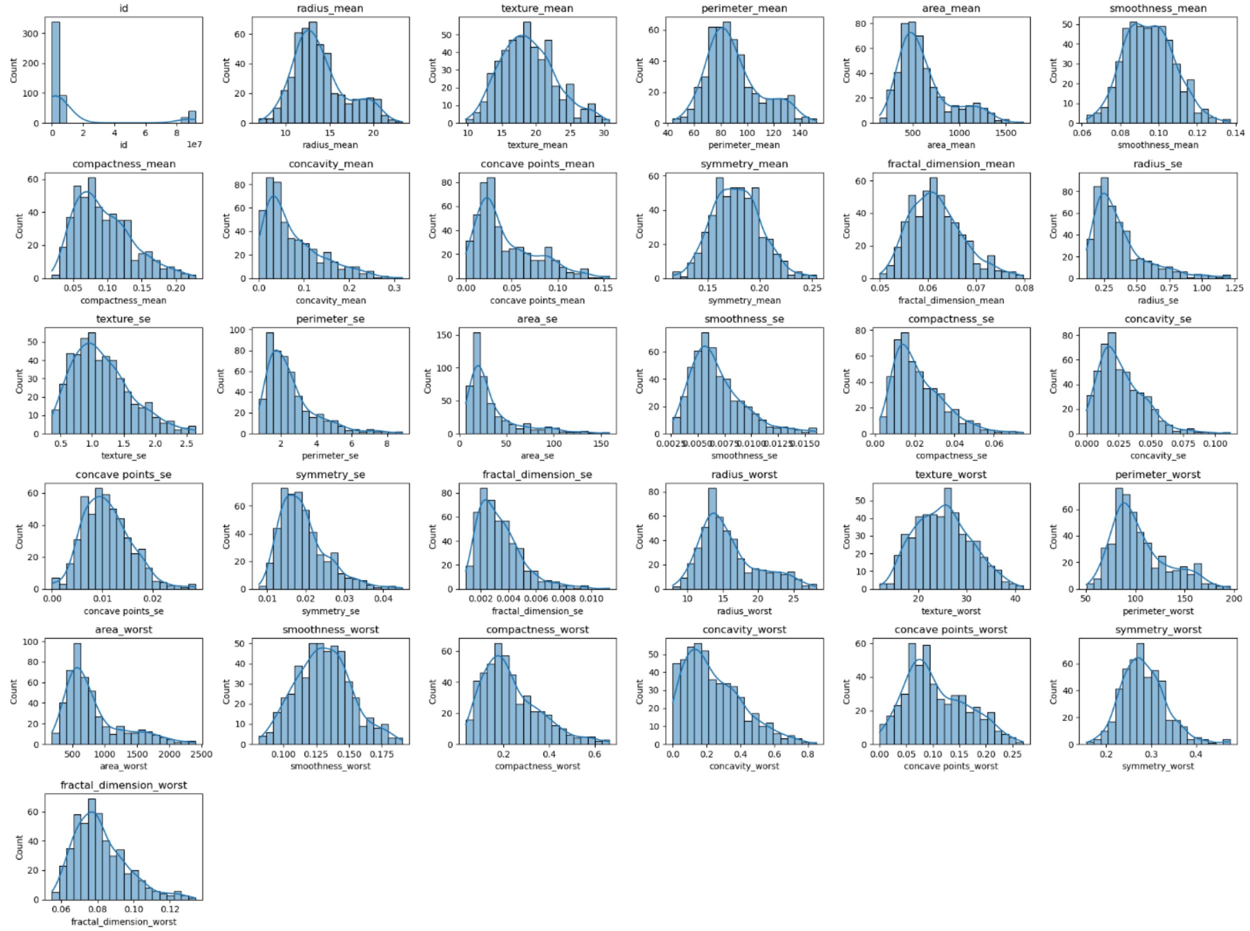
To assess the performance of different machine learning models in tumor classification, we have employed six popular algorithms: Logistic Regression, Artificial Neural Network, Support

Vector Machine, Random Forest, Decision Tree, and Naive Bayes. These algorithms offer varying levels of complexity and have demonstrated efficacy in classification tasks.

Our evaluation of these models is based on a comprehensive set of performance metrics, including accuracy, precision, recall, specificity, F1 score, and AUC. These metrics provide insights into the models' ability to make accurate predictions while considering trade-offs between true positives, true negatives, false positives, and false negatives.

In summary, this study represents a systematic exploration of tumor classification using machine learning techniques. The results and insights gained from this analysis can potentially contribute to the development of more accurate and reliable tools for cancer diagnosis, ultimately leading to better patient care and outcomes.

Data interpretation: Data is used from https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic .Each boxplot summarizes the distribution of a different measurement made on 569 breast cancer patients.



The measurements include:Mean radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimensionStandard error of these measurements.Worst radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimensionHere are some general observations we can make from the boxplots:

The distributions of most of the measurements are right-skewed. This means that there are more data points on the left side of the box (towards lower values) than on the right side (towards higher values).

The spread of the data (represented by the width of the boxes) varies between the measurements. For example, the boxplots for radius and area are relatively wide, while the boxplots for symmetry and fractal dimension are relatively narrow.

There are outliers for some of the measurements. These are data points that fall outside the whiskers of the boxplots. Outliers can indicate that there are a few patients with very different values for these measurements compared to the rest of the group.

Radius: The mean radius is around 14, and the standard error is around 0.1. The worst radius is slightly higher, at around 15. There are a few outliers for both mean and worst radius.

Texture: The mean texture is around 30, and the standard error is around 3. The worst texture is slightly higher, at around 35. There are a few outliers for both mean and worst texture.

Perimeter: The mean perimeter is around 50, and the standard error is around 5. The worst perimeter is slightly higher, at around 60. There are a few outliers for both mean and worst perimeter.

Area: The mean area is around 1000, and the standard error is around 150. The worst area is slightly higher, at around 1200. There are a few outliers for both mean and worst area.

Smoothness: The mean smoothness is around 0.1, and the standard error is around 0.01. The worst smoothness is slightly higher, at around 0.12. There are a few outliers for both mean and worst smoothness.

Compactness: The mean compactness is around 0.1, and the standard error is around 0.02. The worst compactness is slightly higher, at around 0.15. There are a few outliers for both mean and worst compactness.

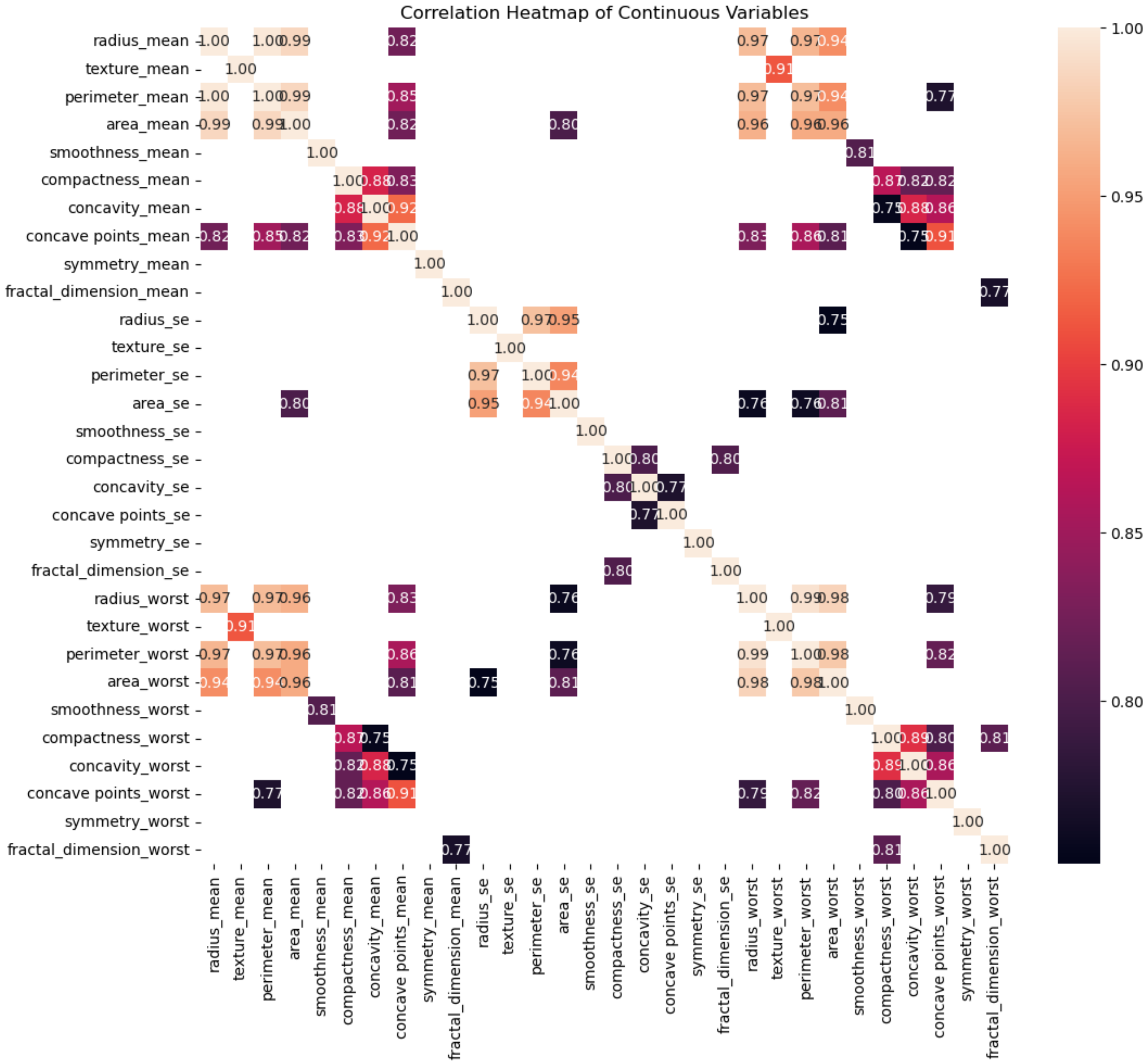
Concavity: The mean concavity is around 0.05, and the standard error is around 0.01. The worst concavity is slightly higher, at around 0.1. There are a few outliers for both mean and worst concavity.

Concave points: The mean concave points is around 0.05, and the standard error is around 0.01. The worst concave points is slightly higher, at around 0.1. There are a few outliers for both mean and worst concave points.

Symmetry: The mean symmetry is around 0.5, and the standard error is around 0.1. The worst symmetry is slightly lower, at around 0.4. There are a few outliers for both mean and worst symmetry.

Fractal dimension: The mean fractal dimension is around 0.1, and the standard error is around 0.01. The worst fractal dimension is slightly higher, at around 0.12. There are a few outliers for both mean and worst fractal dimension.

Data analysis: Correlation:The heatmap you sent me shows the correlation between different features of breast cancer patients. The features are listed on the left and right sides of the heatmap, and the correlation between each pair of features is represented by the color in the box where the two features meet. Red boxes indicate a positive correlation, meaning that the two features tend to increase or decrease together. Blue boxes indicate a negative correlation, meaning that one feature tends to increase as the other decreases. And white boxes indicate no correlation.



here's an interpretation of the correlations between the various features of breast cancer patients:

Strong Positive Correlations:

Tumor size features (radius, perimeter, area): These features are highly correlated with each other, as expected, signifying that larger tumors tend to have a larger radius, perimeter, and area.

Texture features (mean, worst): These features also show a positive correlation, suggesting that tumors with rougher texture tend to have rougher texture across different severity levels.

Concavity features (mean, se, points): Similar to texture, these features exhibit positive correlations, implying that tumors with more concave areas also tend to have higher standard error and more concave points.

Strong Negative Correlations:

Smoothness vs. Concavity: There's a strong negative correlation between smoothness and concavity features. This means that tumors with smoother surfaces tend to have less concavity and fewer concave points, and vice versa.

Smoothness vs. Compactness: Similar to concavity, smoothness also shows a negative correlation with compactness features. This suggests that smoother tumors are generally less compact.

Weak Correlations:

Fractal dimension features: Most correlations involving fractal dimension features are weak compared to others. This indicates that these features might not be strongly associated with most other tumor characteristics in the dataset.

Additional Observations:

The heatmap primarily shows positive correlations on the upper half and negative correlations on the lower half, making the color pattern diagonally symmetric.

The color intensity reflects the strength of the correlation, with darker shades representing stronger positive or negative correlations.

Some features, like "symmetry," have weaker correlations with most other features, suggesting they might not be as informative for characterizing tumors compared to others.

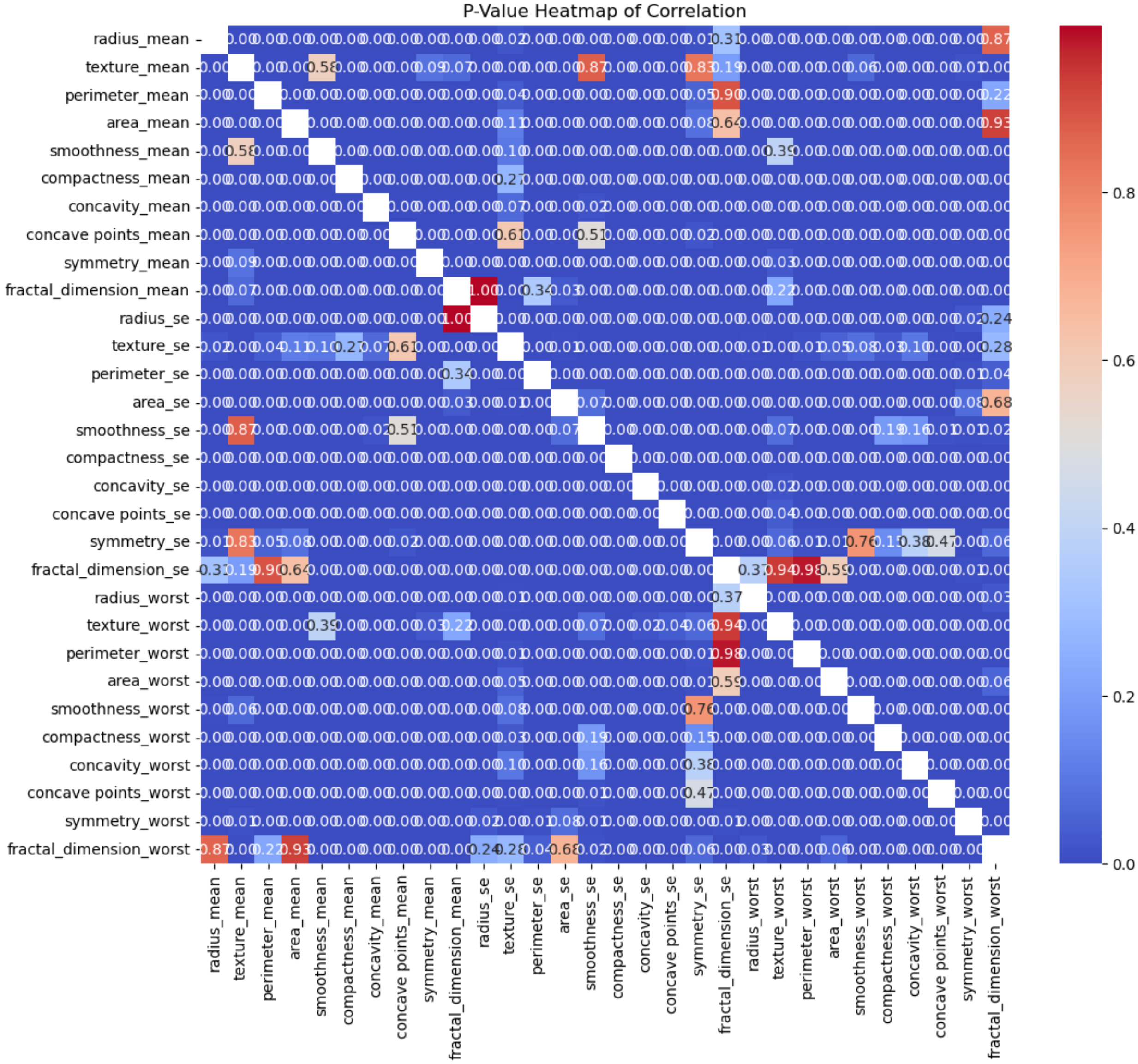
Overall:

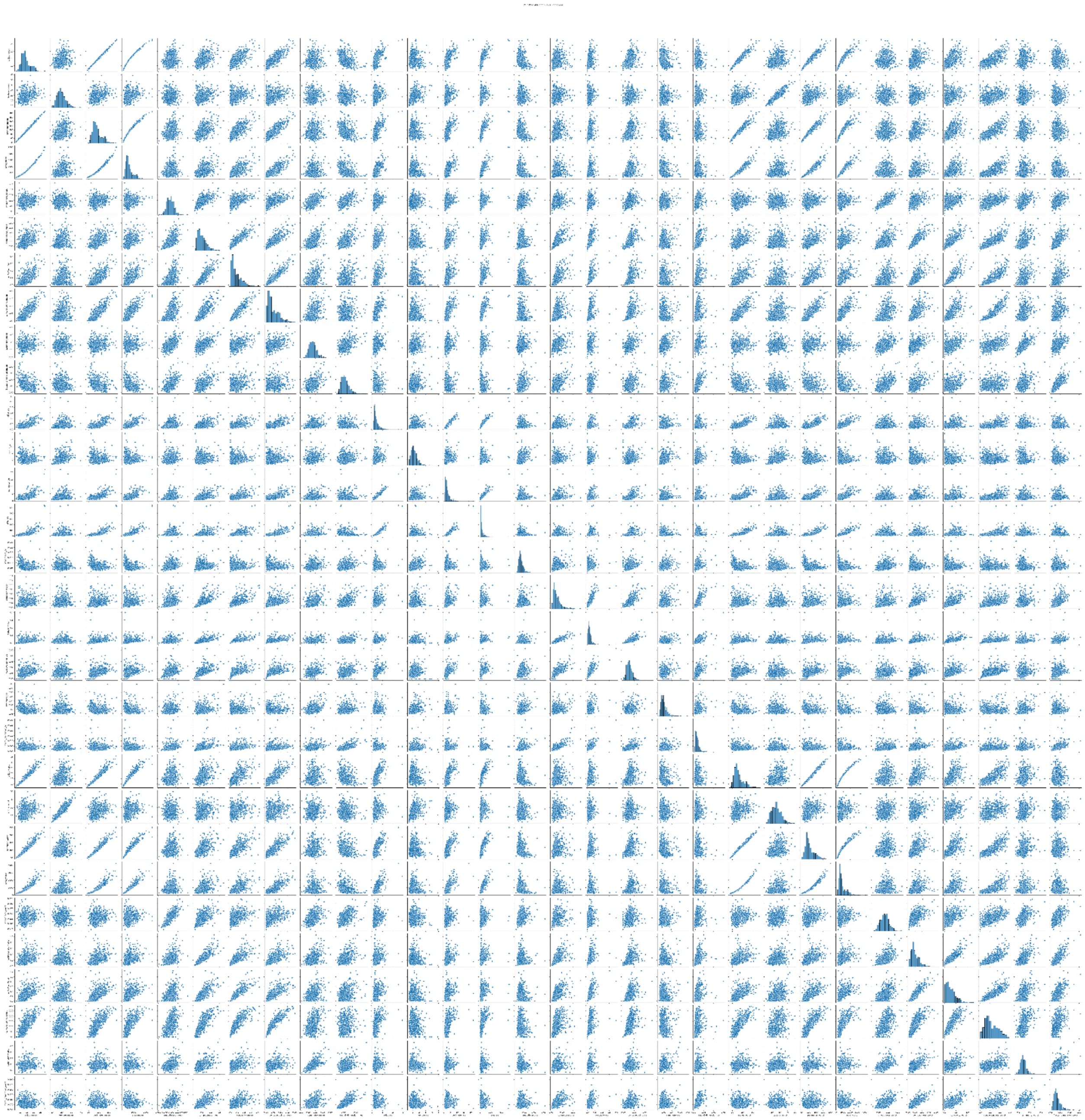
The heatmap provides valuable insights into the relationships between different tumor features. Understanding these correlations can be helpful for:

Diagnosis: Identifying features that are highly associated with malignancy can aid in early and accurate diagnosis.

Prognosis: Correlations between features and tumor aggressiveness can help predict patient outcomes and guide treatment decisions.

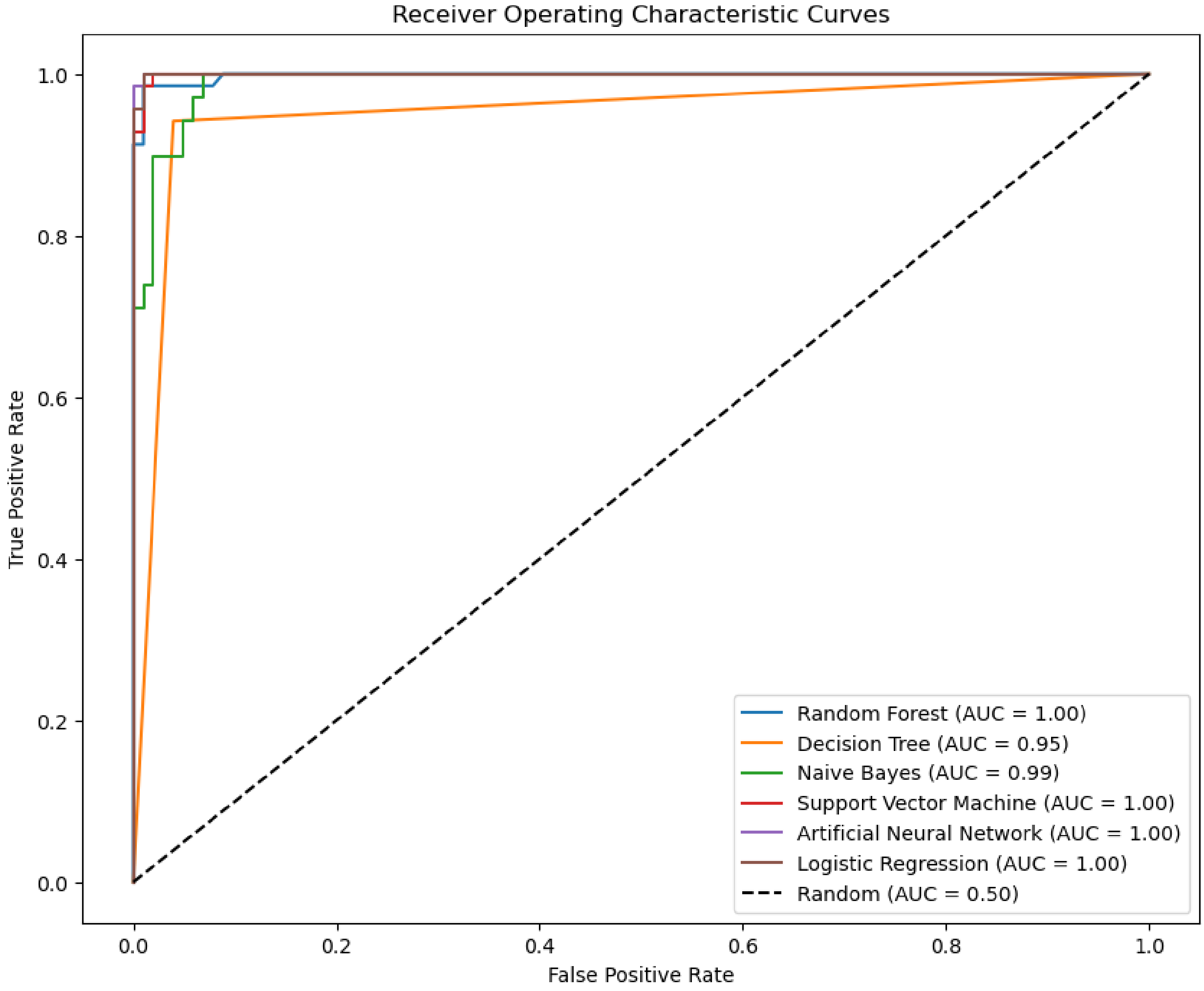
Treatment development: Targeting specific features based on their correlations with tumor behavior could lead to more effective therapies.





Rusult:

Among the machine learning models evaluated, Logistic Regression and Artificial Neural Network exhibited the highest overall classification performance with accuracy scores of 0.994152 and 0.988304, respectively. These models achieved near-perfect precision, recall, and AUC values, indicating their ability to effectively distinguish between benign and malignant tumors.


Receiver Operating Characteristic Curves
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
Random Forest (AUC = 1.00)
Decision Tree (AUC = 0.95)
Naive Bayes (AUC = 0.99)
Support Vector Machine (AUC = 1.00)
Artificial Neural Network (AUC = 1.00)
Logistic Regression (AUC = 1.00)
Random (AUC = 0.50)