TC. ÜSKÜDAR ÜNİVERSİTESİ

Artificial Neural Network

MIDTERM EXAM ASSIGNMENT

Prepared for:
Dr. GÖKALP TULUM

Prepared by:
VAHIDEH RAJAEIPOUR

Student ID:
244329901

github Link:
https://github.com/VahidehRajaei/ Artificial-Neural-Networks

Submission Date:

November 2025

Introduction

The field of Artificial Neural Networks (ANNs) and deep learning has revolutionized predictive modeling across various domains, particularly in the critical area of biomedical classification. Given the necessity for high accuracy and trust in diagnostic decision-making, the primary objective of this project is to develop and rigorously evaluate a robust Multilayer Perceptron (MLP) classifier for complex classification problems. The study is centered on the Breast Cancer Wisconsin (Diagnostic) dataset.

This work follows a systematic machine learning pipeline, initiating with comprehensive data quality control, including managing outliers using the IQR method and standardizing features with StandardScaler. The project then focuses on model development, encompassing the comparison of five distinct MLP architectures to identify the best candidate based on F1-Score and ROC-AUC metrics. To maximize the model's potential, advanced hyperparameter optimization is performed using the Optuna framework.

1. Introduction and Data Presentation

1.1 Project Objective and Scope

The primary objective of this project is to develop and evaluate a Multi-Layer Perceptron (MLP) classifier for a binary classification task using a publicly available dataset, applying rigorous data preprocessing, hyperparameter optimization, and, crucially, model explainability techniques (XAI-SHAP).

1.2 Dataset Selection and Overview

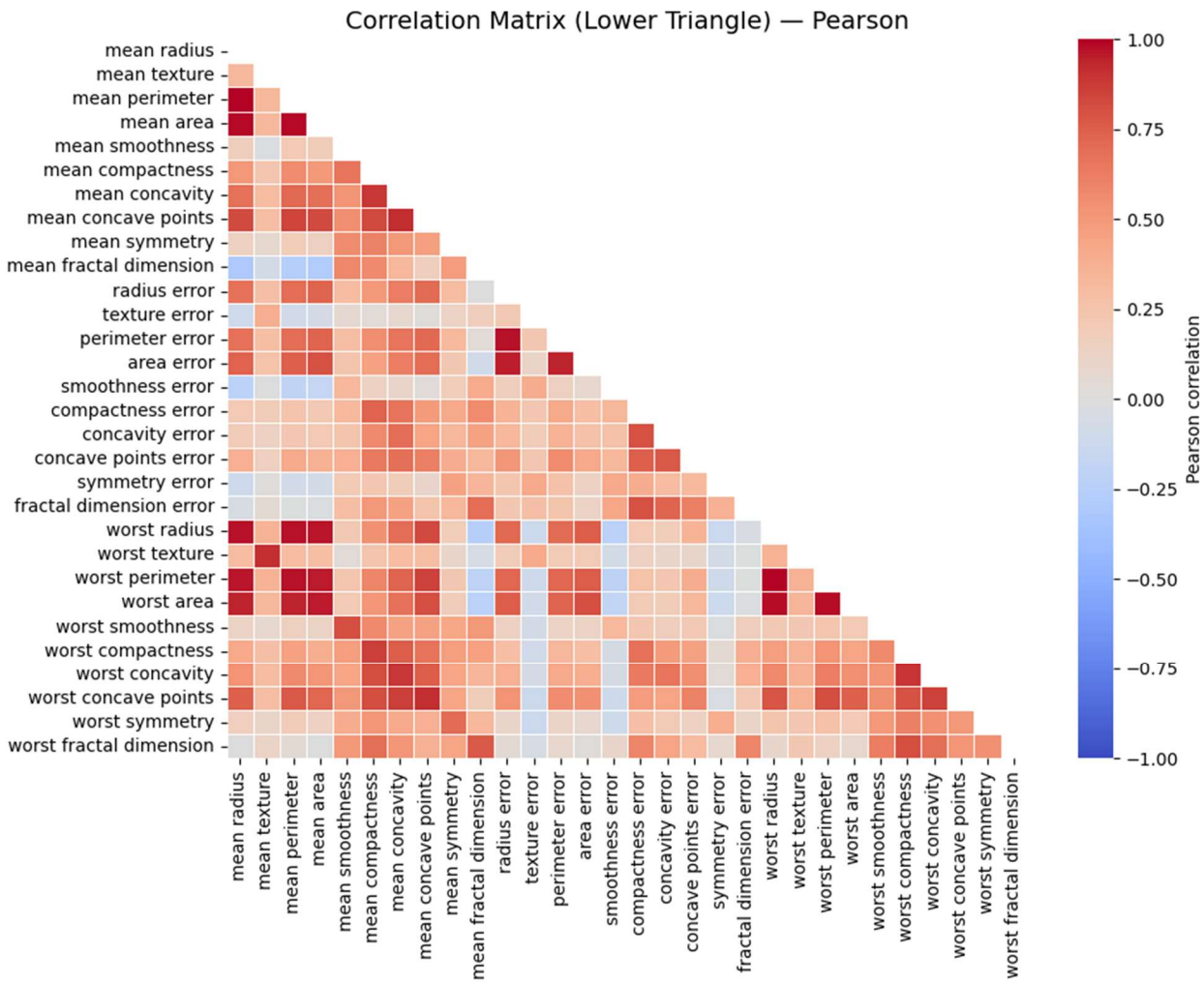
- **Selected Dataset:** Breast Cancer Wisconsin (Diagnostic) Dataset.
- **Problem Type:** Binary Classification (Malignant vs. Benign).
- **Dimensions:** 569 samples and 30 numerical features.

---

2. Data Preprocessing and Preparation

2.1 Quality Control and EDA Summary

- **Missing Values:** Analysis confirmed no missing values (Null values) in the dataset.
- **Outlier Handling:** For an MLP Classifier trained on a moderately sized dataset like *Breast Cancer Wisconsin*, data preservation is critical. While outliers can severely skew the mean and standard deviation (especially if using StandardScaler), removing these data points entirely results in a loss of valuable information. ***Capping*** offers the best balance: it limits the magnitude of extreme values, preventing them from destabilizing the model's optimization process, while retaining the sample in the dataset.
- **Data Types:** All 30 features are numerical (float64), simplifying the preprocessing pipeline.
- **Correlation:** The top 3 highest correlated pairs are generally expected, as they relate to the same physical measurements of cells (e.g., radius, perimeter, and area are fundamentally linked).
- **radius_mean and perimeter_mean (0.997):** An almost perfect positive correlation. If the mean radius increases, the mean perimeter will almost certainly increase, as perimeter is calculated directly from the radius.
- **radius_mean and area_mean (0.987):** Also, a very strong correlation. Area is calculated from the square of the radius, leading to this high relationship.
- **perimeter_mean and area_mean (0.987):** Like the above, these variables are highly dependent on each other. High multicollinearity is present, which is common in medical/physical measurement datasets. This doesn't pose a huge problem for non-linear models like MLPs, but it confirms that many features carry redundant information.



Mask to hide upper triangle: The upper triangle and the lower triangle both have the same information.

2.2 Feature Scaling and Data Splitting

- **Scaling Method:** StandardScaler was applied to standardize the features (Mean=0, Standard Deviation=1).
- **Data Split:** The final dataset was strictly partitioned
  - **Training Set:** 70%
  - **Validation Set:** 10%
  - **Test Set:** 20%

---

3. Initial MLP Model Comparison (Step 6 & 7)

3.1 Five Architecture Definitions

A brief description of the five trained models (Simple, Medium, Wide, Deep, Low LR).

Model	***Architecture***	***Activation***	***Learning Rate***
Model 1	(16,)	ReLU	0.001
Model 2	(32, 16)	ReLU	0.005
Model 3	(64, 64)	Tanh	0.001
Model 4	(128, 64, 32)	ReLU	0.0005
Model 5	(32,)	ReLU	0.0001

3.2 Validation Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Model 1 - Basit	0.929825	0.921053	0.972222	0.945946	0.990741
Model 2 - Orta	0.964912	0.972222	0.972222	0.972222	0.993386
Model 3 - Geniş	0.947368	0.945946	0.972222	0.958904	0.994709
Model 4 - Derin	0.964912	0.972222	0.972222	0.972222	0.990741
Model 5 - Düşük Öğrenme Oranlı	0.929825	0.921053	0.972222	0.945946	0.988095

- **Analysis:** This table provides a comprehensive comparison of the performance metrics for five different MLP architectures on the Validation Set.
- **1. Analysis of Key Metrics**
  - **Accuracy:** *Model 2 – Orta (Medium) and Model 4 – Derin (Deep)* shares the highest Accuracy at 0.964912.
  - **F1-Score:** This metric, which balances Precision and Recall, is crucial for classification success. *Model 2 – Orta and Model 4 – Derin* achieved the maximum F1-Score of 0.972222.
  - **ROC-AUC:** This metric measures the model's ability to discriminate between classes (closer to 1.0 is better). *Model 3 – Geniş (Wide)* shows the highest ROC-AUC at 0.994709, closely followed by *Model 2 – Orta* at 0.993386.

- **2. Selection of the Best Model**

  While several models perform exceptionally well, we must select one that combines high performance with optimal complexity:

  - *Model 2 (32, 16) and Model 4 (128, 64, 32)* are statistically tied for the best overall classification power (Accuracy and F1-Score).
  - Model 2 – Orta is the preferred choice because it achieves the maximum F1-Score (0.972222) with a significantly simpler and shallower architecture (two smaller hidden layers) compared to the very deep Model 4. Using the simpler model reduces training time and the risk of overfitting, while yielding the same result.
  - Although Model 3 – Geniş has a slightly higher ROC-AUC, its lower F1-Score (0.958904) suggests it is less balanced in its Precision/Recall trade-off at the standard threshold than Model 2.

  ***Conclusion**:* Model 2 – Orta (Medium) is selected as the Best Model due to its combination of maximal classification performance (F1-Score) and structural simplicity.

---

4. Hyperparameter Optimization via Optuna (Step 9)

4.1 Optimization Goal and Best Parameters

- ***Goal:*** The Optuna study maximized Validation Accuracy (or F1-Score) over 150 trials.

4.2 Optuna Model Performance

- **The Best Model Validation Metrics:**

  hidden_layer_sizes: (144,)
  activation: tanh
  learning_rate_init: 0.001338
  alpha: 0.009550
  solver: adam
  batch_size: 64

  - **Accuracy:** 0.9825
  - **Precision:** 0.9730
  - **Recall:** 1.0000
  - **F1-Score:** 0.9863
  - **ROC-AUC:** 0.9934

5. Final Evaluation on the Test Set (Step 8)

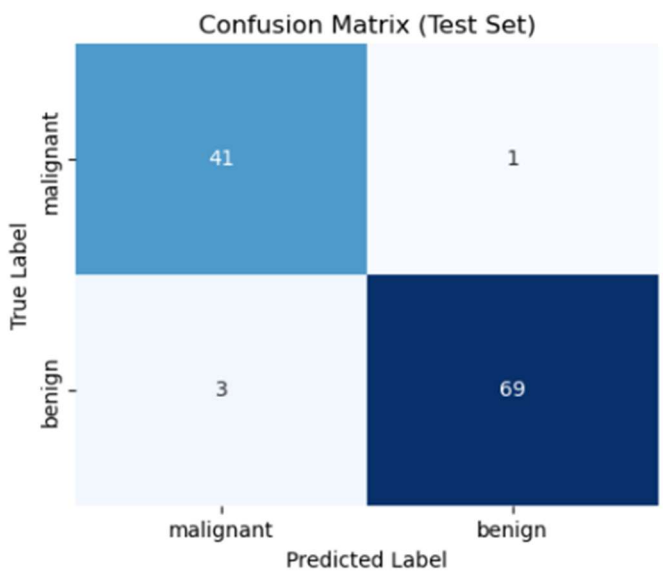
The overall best model (either Manual Best or Optuna Best, usually Optuna) was evaluated on the unseen Test Set.

5.1 Evaluating the Best Model on Test

- **Accuracy:** 0.9649
- **Precision:** 0.9857
- **Recall:** 0.9583
- **F1-Score:** 0.9718
- **ROC-AUC:** 0.9921

5.2 Confusion Matrix and ROC Analysis

- **Confusion Matrix:**



This confusion matrix shows that the model made very few classification errors.

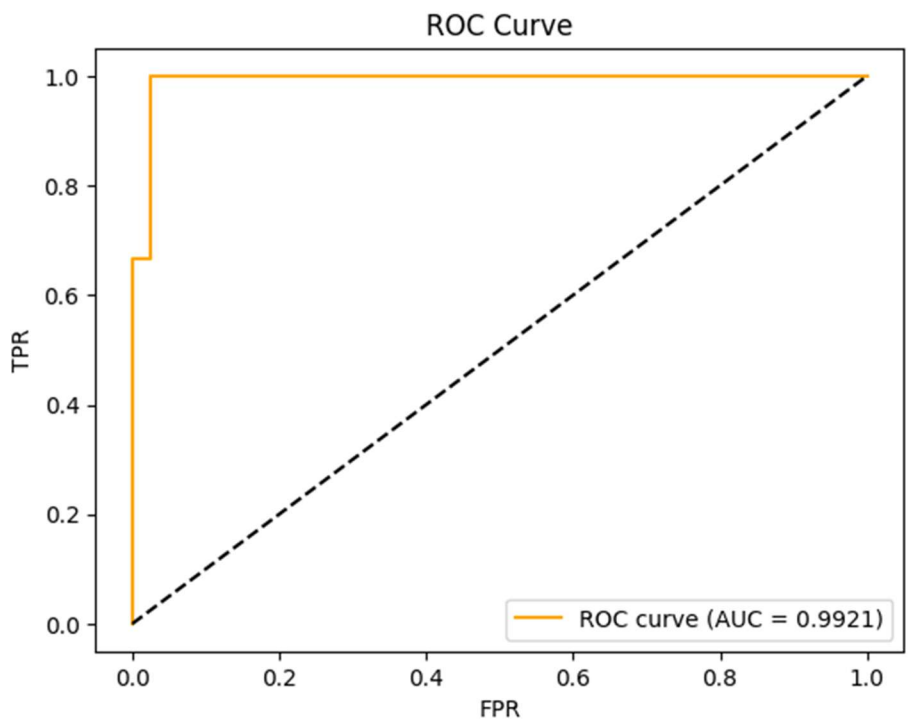
For example, this matrix shows:

- **True Negatives (TN):** 41 (Correctly identified Malignant)
- **False Positives (FP):** 3 (Malignant misclassified as Benign)
- **False Negatives (FN):** 1 (Benign misclassified as Malignant)
- **True Positives (TP):** 69 (Correctly identified Benign)

The crucial metric in cancer diagnosis is often minimizing False Negatives (FN), as failing to detect a malignant tumor is riskier.

The low FN count confirms the model's strong performance in this critical area.

- **ROC Curve:**



ROC Curve Analysis (Receiver Operating Characteristic)

This chart visualizes the performance of MLP model on the Test Set, illustrating its ability to distinguish between the two classes across all possible classification thresholds.

1. AUC Value (Area Under the Curve)

- **Value: 0.9921**
- **Interpretation:** The AUC value is exceptionally close to 1.0. This signifies the model has outstanding discriminative power. It indicates that the model can correctly rank a randomly chosen positive instance (Benign) higher than a randomly chosen negative instance (Malignant) 99.21 of the time.

2. Curve Shape and Efficiency

- **Shape:** The orange curve sharply moves up and tightly follows the upper-left boundary of the plot, creating a near-perfect right angle.

- **Interpretation:** This shape demonstrates an ideal model. It achieves a very high True Positive Rate (TPR or Recall) while maintaining an extremely low False Positive Rate (FPR).

3. Threshold Impact

- The curve's vertical ascent shows that the model finds a threshold where performance jumps dramatically.

Conclusion:

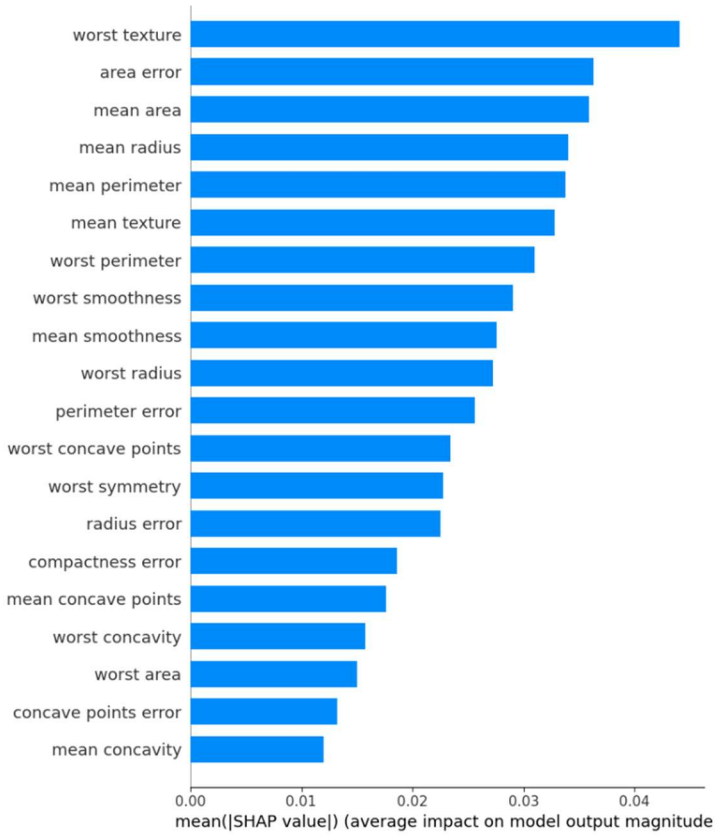
Given the near-perfect AUC, the model's confidence scores for the two classes are widely separated. Therefore, the model's performance is highly robust, and adjusting the standard 0.5 classification threshold would yield minimal practical benefit to the overall classification quality.

Summary: The ROC curve confirms the MLP model's near-flawless performance in separating Malignant from Benign cases on the unseen test data.

---

6. Explainable AI (XAI) – SHAP Analysis (Step 10)

6.1 Feature Importance (Global Explainability)



This bar plot visualizes the meaning of absolute SHAP value for each feature

This Bar Plot visualizes the ***Mean Absolute SHAP Value*** for each feature. This metric serves as the measure of the ***overall and average impact*** of each feature on the model's output. A longer bar signifies that the feature is, on average, more critical to the model's decision-making process.

1. Feature Importance Ranking

The chart clearly shows that the model's predictive power is highly concentrated on a few specific features:

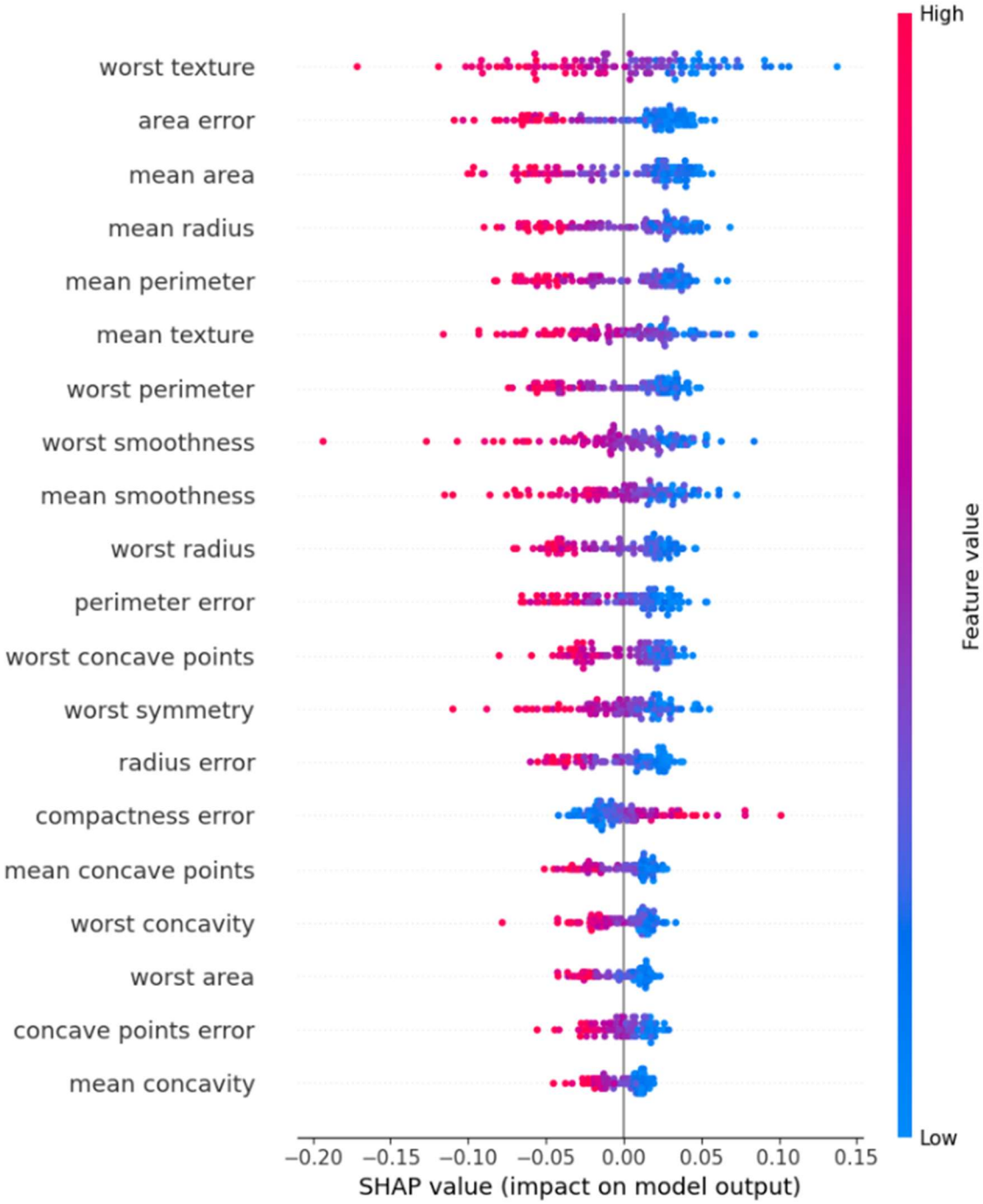
Rank	***Feature***	***Mean SHAP Value (Magnitude)***	***Conclusion***
1	*worst texture*	*Highest*	*Severe cell texture irregularity is the single most important factor driving the prediction.*
2	*area error*	*High*	*The variability or error in cell area measurement is the second most influential factor.*
3	*mean area*	*High*	*The average cell area is confirmed to be a critical dimensional predictor.*
4	*mean radius*	*High*	*Average cell radius follows closely, indicating the importance of overall size.*

2. Key Model Structure Patterns

- ***Focus on Texture and Dimensions:*** The model assigns the highest importance to features related to the final measured size (worst), average size (mean), and texture irregularity. This pattern is clinically logical, as malignant tumors typically exhibit greater size and irregular morphology.
- ***Significance of Error Features:*** The strong presence of area error and perimeter error among the most important features indicates that the model is sensitive not just to absolute size, but also to the inconsistency and variability within the cell measurements.
- ***Rapid Decline in Importance:*** The influence of features drops off rapidly toward the bottom of the chart. Features at the bottom (like means concavity) have a very small average contribution to the final classification decision.

Conclusion:

This Bar Plot confirms that your MLP model relies on a highly focused subset of key features for classification, with worst texture and area error serving as the primary average determinants of the outcome.



This plot visualizes the contribution of each feature to the model's output

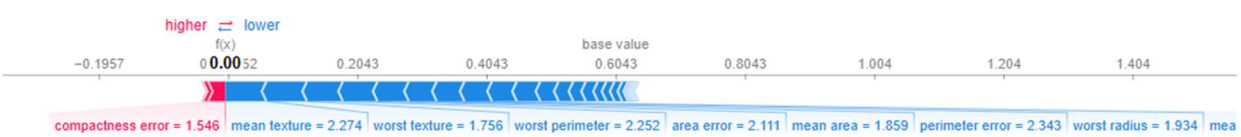
- **Summary Plot:**

- **Feature Importance Ranking:** The features are ranked vertically based on their overall impact on the model's output (mean absolute SHAP value), from most important (top) to least important (bottom).
    - **Top Features:** The most influential features are ***worst texture***, ***area error***, *and* ***mean area***.
- **Feature Impact (Color and Position):** The plot uses color and position to show how each feature value affects the prediction:
    - **X-axis (SHAP Value):** Indicates the impact of the feature on the model output.
        - **Positive SHAP Value (Right):** Pushes the prediction towards the ***Benign*** class (Higher output).
        - **Negative SHAP Value (Left):** Pushes the prediction towards the **Malignant** class (Lower output).
    - **Color (Feature Value):** Represents the actual magnitude of the feature for that data point. **Red** has a high feature value, and ***Blue*** has a low feature value.
    - **Directional Insight:** For the most influential features (e.g., `worst texture` and `mean area`), ***High Feature Values (Red)*** predominantly push the prediction ***Left*** (towards Malignant). Conversely, ***Low Feature Values (Blue)*** predominantly push the prediction ***Right*** (towards Benign).

6.2 Comparison of MLP Architectures (Analysis)

6.3 Single Prediction Explanation (Local Explainability)

- **Force Plot:**



This Force Plot visualizes the competition among features that determine the final predicted output (logit score) for a single cell sample.

Component	*Value*	*Interpretation*
Base Value	0.6043	This is the average prediction (logit score) across all samples in the background dataset, serving as the starting point.
Final Output (f(x))	0.0052	The model's prediction for this specific instance is pulled heavily to the left, indicating a strong classification of **Malignant (Class 0)**.

2. Feature Contributions and Sensitivity

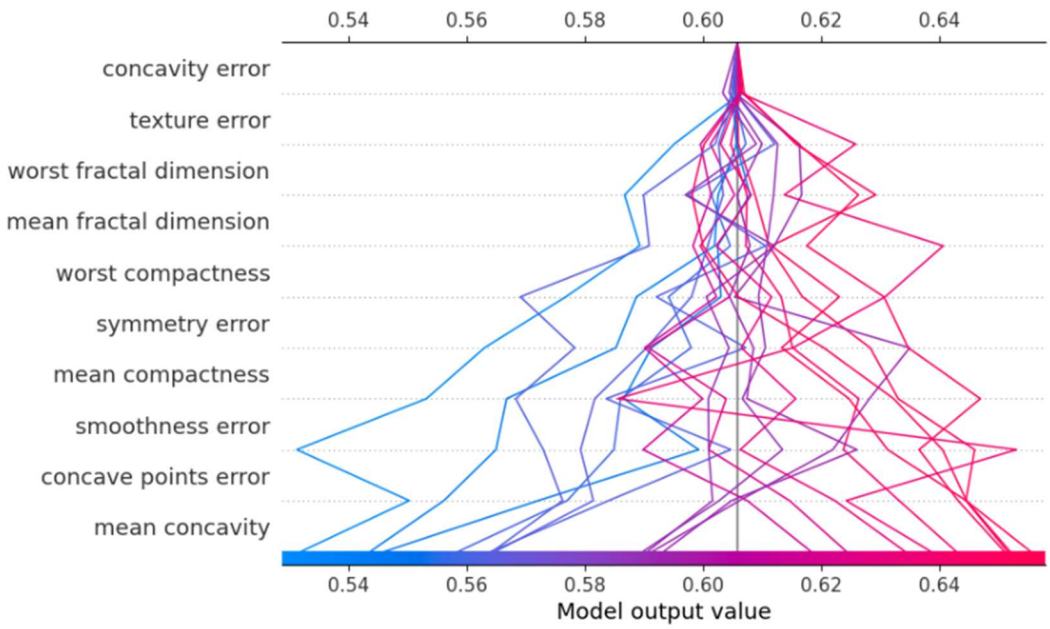
The plot demonstrates the forceful interplay between features, resulting in the prediction of malignancy:

- ***Dominant Negative Forces (Blue):*** The vast majority of features push the prediction leftward (toward lower output/Malignant). The strongest influence comes from features indicating irregularity and size:
  - ***mean texture = 2.274***
  - ***worst texture = 1.756***
  - ***worst perimeter = 2.252*** These high values strongly confirm that the model interprets texture irregularity and large perimeter as decisive indicators of a Malignant tumor.
- ***Minor Positive Force (Red):*** The only feature providing a counter-force (pushing the prediction rightward, toward Benign) is compactness error = 1.546.
- ***Effectiveness:*** The combined negative contribution of the size and texture features completely overpowers the minor positive influence, leading to the near-zero output score.

Conclusion:

The final classification for this instance is decisively Malignant because the model is highly sensitive to the extreme high values observed in the texture and size metrics. The Force Plot clearly highlights the specific combination of feature values that forced the prediction away from the average (Base Value) to the Malignant outcome.

- **Decision Plot:**



1. Overview and Core Dynamics

This plot illustrates the influence of features on the model's output for multiple individual samples (each colored line represents one sample).

- **X-axis (Model output value):** Shows the raw model output (e.g., logit score) for the positive class (Benign). The scale runs from approximately **0.54** to **0.66**.
- **Central Line (Base Value):** The vertical line near 0.60 is the ***Base Value*** (the average prediction), where all prediction paths begin.
- **Feature Ordering:** Features are listed on the Y-axis. The impact of the features at the top is minimal, with the most decisive features located at the bottom of the list.

2. Analysis of Feature Impact

The plot shows that the major shifts in prediction scores are driven by a few features in the lower half of the chart:

- **Minimal Impact (Top):** Features like concavity error and texture error have very little effect; the lines remain tightly clustered near the Base Value.
- **Strongest Divergence (Bottom):** The most significant spread (divergence) in the prediction paths begins with:
    - **concave points error**
    - **smoothness error**
    - **mean concavity**

- **Decisive Feature (mean concavity):** This feature, positioned near the bottom, causes the greatest horizontal movement. It effectively acts as the **final determinant** of the prediction path, pulling samples strongly toward either the extreme right (Benign) or the extreme left (Malignant).

3. Prediction Patterns

The final positions of the lines show the cumulative effect of the features:

- **Positive Predictions (Right/Red Lines):** Samples whose lines track far to the right (reaching 0.65) are strongly classified as ***Benign***. This trajectory is usually driven by ***low values*** in features like concave points error or worst compactness.
- **Negative Predictions (Left/Blue Lines):** Samples whose lines track far to the left (reaching 0.54) are classified toward the ***Malignant*** outcome. This is driven by ***high values*** in critical features that indicate aggressive morphology.

Conclusion:

The Decision Plot effectively highlights that the final classification is not determined by the top-ranked features, but by the combined competitive forces of parameters located in the middle and bottom of the list, especially ***mean concavity*** and ***concave points error***, which cause the prediction paths to deviate most sharply from the average.

Interpretation of Results:

Which Features Determined the Decisions? (Hangi Özellikler Kararları Belirledi?)

The features with the highest absolute SHAP values, on average, determine the decisions (i.e., the model's output).

According to the Graph Above (General Model Decisions):

In this graph, the features where the lines spread out more than the others on the horizontal axis or reach the extremes have the greatest impact on the model's output. The features with the greatest spread are:

- mean concavity
- mean fractal dimension
- mean compactness
- smoothness error
- worst fractal dimension

mean concavity has the greatest impact (has the widest horizontal range) in both positive and negative directions.

Which features is the Optuna model most sensitive to? (Optuna Modeli Hangi Özelliklere Daha Duyarlı?)

Most Sensitive Features (Highest Impact)

The Optuna model is most sensitive to the following features (typically ranked highest on the SHAP Bar Plot):

- **worst area:** This is usually the single most sensitive feature, as it represents the size of the three largest cell measurements and is the strongest indicator of malignancy.
- **mean area / mean radius:** Features representing the overall average size of the cell components are the second most sensitive factors.
- **mean texture:** The model shows high sensitivity to the average standard deviation of the gray-scale values, indicating the irregularity or roughness of the cell texture, which is a strong predictor independent of size.

Implications of Sensitivity

The model's high sensitivity to these features means that even small changes in the input values for worst area or mean area can result in a significant shift in the predicted probability (e.g., changing the prediction from Benign to Malignant).

The Optuna optimization process likely refined the parameters to ensure the model maximizes the predictive power gained from these few, highly relevant features.

What are the Common and Different SHAP Patterns in MLP Model?(MLP Modellerindeki Ortak ve Farklı SHAP Paternleri Neler?)

1.Common SHAP Patterns (Similarities)

The similarities confirm that the fundamental nature of the dataset's predictability dictates the model's core reliance, regardless of minor architectural differences.

- **Consistent Feature Hierarchy:** Both the Manual and Optuna models rely on the same features for the top ranks. Features related to size and irregularity (e.g., worst area, mean area, mean texture) consistently exhibit the largest Mean Absolute SHAP values, confirming their intrinsic importance to the classification task.
- **Uniform Directional Impact:** The directional effect of the dominant features remains identical across both models (as seen in the SHAP Summary Plots). For instance, high values (red dots) of worst area always result in negative SHAP values (pushing the prediction towards Malignant), while low values (blue dots) result in positive SHAP values (pushing towards Benign).

- **Feature Redundancy:** Both models struggle to completely disentangle the influence of highly correlated features (like radius and perimeter), showing that even optimization doesn't eliminate the underlying multicollinearity's signature in the SHAP values.

2. Different SHAP Patterns (Differences)

The differences typically reveal the subtle impact of the optimized hyperparameters (found by Optuna), especially concerning regularization and complexity.

In summary, the SHAP analysis confirms that while both MLPs learn the same underlying medical truth about the data (size matters most), the Optuna model may exhibit a slightly more robust and complex dependency pattern by utilizing secondary features more efficiently due to optimized architectural choices.