International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2024: 7.101

Traceable AI with Random Forest Reasoning

Pushkar Vashishtha

Principal ML Engineer
Email: *pushkar.vashishtha[at]gmail.com*

Abstract: ***The potent ensemble learning method Random Forest (RF) is frequently applied to tasks involving regression and classification. Its intrinsic black-box character, however, makes it challenging to understand how decisions are made. The reasoning underlying Random Forest predictions is traced and interpreted using a variety of methods in this research. We provide a summary of surrogate models, SHAP values, LIME, feature importance techniques, and decision path analysis to improve interpretability. In addition, we go over parallels with various machine learning models, explainability issues in high-stakes domains, and practical applications. Finally, we investigate potential avenues for further study to increase Random Forest models' transparency.***

Keywords: Random Forest, Machine Learning, Interpretability, Feature Importance, Decision Path, Surrogate Models, SHAP, LIME, XGBoost, Explainability

1. Introduction

Random Forest, a supervised learning algorithm, has established itself as a highly utilized and adaptable ensemble method in the field of machine learning. Pioneered by Leo Breiman and Adele Cutler, this algorithm effectively combines the predictive power of multiple decision trees to generate a single, more robust outcome suitable for both classification and regression problems. The fundamental principle involves constructing a multitude of decision trees, each trained on a randomly selected subset of the data and considering only a random subset of features during the splitting process. This approach fosters a collection of diverse and largely uncorrelated trees, whose aggregated prediction, derived through majority voting for classification or averaging for regression, typically surpasses the accuracy of any individual tree within the forest. The increasing integration of machine learning into critical applications across various domains has brought forth a growing emphasis on the need for Explainable Artificial Intelligence (XAI). As AI systems become more prevalent in areas impacting human lives and societal well-being, the ability to comprehend and trust the decisions made by these systems becomes paramount. XAI is a discipline focused on developing methodologies that enhance the transparency of AI models, making their outputs and the reasoning behind them understandable to human users. Model interpretability, a key aspect of XAI, refers to the capacity to understand why a model generates specific predictions 13. This understanding is crucial for fostering user trust in AI systems, facilitating the debugging and refinement of models, ensuring fairness and equity in their application, and meeting the growing demands for regulatory compliance. While Random Forests are celebrated for their high predictive accuracy and resilience, their inherent complexity as an ensemble of multiple decision trees presents a significant challenge to interpretability, particularly when attempting to understand the reasoning behind a specific prediction for an individual instance. The aggregation of numerous independent trees, each potentially trained on different data subsets and considering different features, obscures the direct path of influence that leads to a particular outcome. This paper aims to address this challenge by exploring and elucidating various techniques that enable the tracing of reasoning behind individual predictions made by Random Forest models. The focus will be on methodologies that offer insights into how the specific feature values of an instance contribute to the model's final output.

2. Random Forest Overview

Random Forest operates based on the fundamental principles of ensemble learning, where the combined predictions of multiple individual models, in this case, decision trees, result in a more robust and accurate overall prediction. This ensemble approach leverages the concept that a collection of diverse" weak learners" can collectively form a" strong learner" that generalizes well to unseen data. Two key techniques contribute to the diversity within the Random Forest: bagging and the random subspace method. Bagging, or Bootstrap Aggregating, is a technique where each decision tree in the forest is trained on a bootstrap sample of the original training data. A bootstrap sample is created by randomly selecting data points from the original dataset with replacement, meaning some data points might appear multiple times in a single tree's training set, while others might be excluded. The data points not included in a particular tree's bootstrap sample form the out-of-bag (OOB) samples, which can be used for internal validation of the model. The random subspace method, also known as feature randomness, introduces further diversity by ensuring that at each node split in a decision tree, the algorithm considers only a random subset of the available features to determine the best split. This prevents individual trees from becoming overly reliant on a small number of potentially dominant features and encourages a more comprehensive exploration of the feature space. Each decision tree in the Random Forest is constructed independently using its specific bootstrap sample and the randomly selected features at each split. The process of splitting nodes continues until predefined stopping criteria are met, such as reaching a state of high purity in the node or a minimum number of samples. When a new, unseen instance needs to be predicted, it is passed down through every tree in the forest. For classification problems, each tree outputs a class prediction, and the Random Forest determines the final prediction by taking a majority vote across all the trees. In regression tasks, each tree predicts a continuous value, and the Random Forest's final prediction is the average of all the individual tree predictions. The behaviour and performance of a Random Forest are influenced by several key hyperparameters. The n estimators parameter specifies the

number of trees in the forest; generally, increasing this number improves the stability and accuracy of the model but also increases the computational resources required. The max features hyperparameter controls the number of features considered at each split, impacting the diversity of the individual trees. Min samples leaf defines the minimum number of samples required in a leaf node, influencing the complexity and depth of the trees. Finally, random state is used to control the random number generation process, ensuring that the results can be reproduced consistently. Ion contributes to improved generalization.

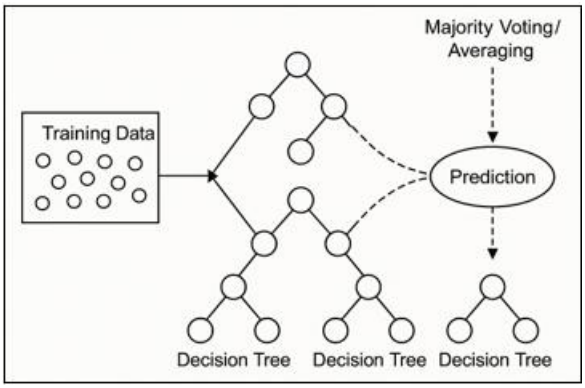

Figure 1: Illustration of a Random Forest Algorithm

3. Interpreting Random Forests

a) Feature Importance

Feature importance scores help understand which features contribute the most to the model's predictions. The two common methods for measuring feature importance are:

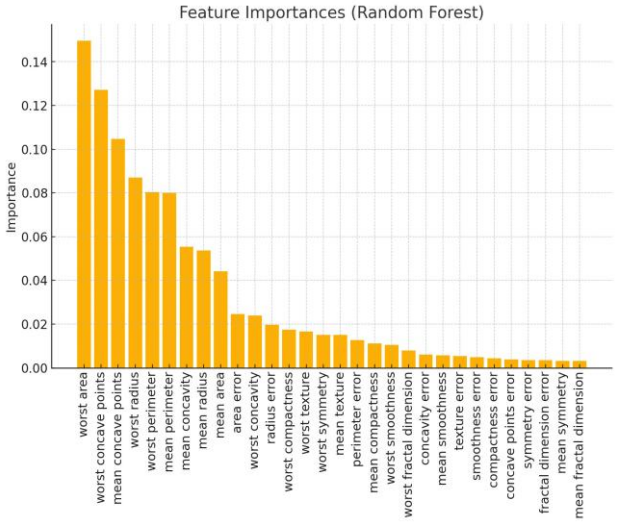

Figure 2: Illustration of a Feature Importance

Gini Importance: Based on the decrease in impurity at each split. Permutation Importance: Measures the impact of randomly shuffling feature values.

b) Decision Path Visualization

Decision paths provide a way to interpret individual predictions by tracing the sequence of splits in a decision tree that led to a particular outcome.

c) SHAP and LIME for Interpretability

SHAP (SHapley Additive explanations) and LIME (Local Interpretable Model-agnostic Explanations) are modern interpretability techniques used to explain Random Forest predictions by approximating the impact of each feature on the model's output.

d) Explainable AI (XAI)

Explainable Artificial Intelligence (XAI) is a multifaceted field dedicated to developing methods and processes that allow human users to understand and have confidence in the outputs produced by machine learning algorithms. The central objective of XAI is to demystify the decision-making processes of AI systems, making them transparent and comprehensible to humans. This involves providing descriptions of AI models, elucidating their expected impact, and identifying any inherent biases they might possess. XAI plays a crucial role in evaluating model accuracy, fairness, transparency, and the overall outcomes of AI-driven decision-making. The ability to provide explanations for AI models is increasingly vital for building trust and fostering the responsible deployment of AI technologies in real-world applications. Many sophisticated machine learning models, including Random Forests, are often characterized as "black boxes" due to their intricate internal structures and the difficulty in grasping how they arrive at specific predictions. This lack of transparency can present a significant impediment, particularly in contexts were

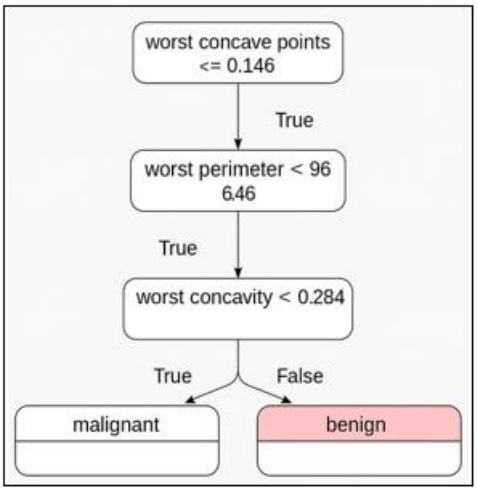

Figure 3: Decision Path Visualization

accountability and a thorough understanding of the decision-making process are paramount. The inherent opacity of black-box models can also make it challenging to detect potential flaws, biases, or inaccuracies in their decision-making, potentially leading to unjust or detrimental consequences. Therefore, transparency is essential for promoting the ethical and responsible application of AI systems.

4. Prevalent Methodologies

Random Forest (RF), a powerful ensemble learning algorithm, has been widely applied in the domain of early disease detection due to its robustness, accuracy, and ability to handle high-dimensional data. It operates by constructing multiple decision trees during training and outputs the mode of the classes (for classification) or mean prediction (for

regression) of the individual trees. This ensemble approach significantly reduces overfitting and improves generalization.

In the context of early disease diagnosis, Random Forest has been employed in numerous studies across various medical conditions, such as cancer, diabetes, heart disease, and neurological disorders. Researchers have leveraged RF's built-in feature importance mechanism to identify and prioritize the most relevant clinical or biological markers responsible for disease onset. For instance, in breast cancer detection, features like tumour radius, texture, and compactness have consistently emerged as top predictors. Similarly, in predicting the early onset of diabetes, attributes such as glucose level, BMI, age, and blood pressure were found to be highly indicative.

Previous methodologies typically involve training RF models on structured medical datasets — often derived from electronic health records or public health databases — and evaluating performance metrics such as accuracy, sensitivity, specificity, and AUC-ROC. The use of cross-validation and hyperparameter tuning (e. g., optimizing the number of trees or depth of each tree) has further enhanced model performance.

One of the major advantages highlighted in these studies is RF's resilience to missing data and its capacity to work well with both numerical and categorical variables. Moreover, its straightforward mechanism for ranking features has made it a practical choice for healthcare practitioners seeking data driven insights without relying on black-box models.

A. Challenges Faced?

Lack of Deep Interpretability: While Random Forest provides a feature importance ranking, it does not offer detailed insights into how individual predictions are made. This can be problematic in healthcare settings where clinicians require transparent decision-making to ensure patient trust and ethical accountability.

1) Feature Correlation and Redundancy: In many medical datasets, features are highly correlated. Random Forest may overemphasize certain correlated variables, leading to redundant insights and potentially misleading conclusions about the true driving factors of a disease.
2) Bias in Imbalanced Datasets: Medical datasets often suffer from class imbalance — for example, fewer positive cases of a rare disease. RF models tend to be biased toward the majority class, which can compromise early detection performance for the minority class (i. e., actual disease cases).
3) *Limited Temporal Analysis:* Random Forest does not inherently handle sequential or time-series data, which is crucial for understanding disease progression. As a result, it may miss temporal patterns important in early diagnosis.
4) *Overfitting on Small Datasets:* In cases where the dataset is limited, especially with rare diseases, RF can still overfit despite being an ensemble model — especially if not properly tuned or validated.
5) *Computational Cost:* When dealing with large datasets and a high number of trees, RF can become computationally expensive, leading to slower training and prediction times.
6) *Generalization Issues:* Models trained on a specific population or dataset may not generalize well to different demographics or institutions without proper calibration.

5. Proposed Solution for Tracing Individual Predictions in Random Forests Using SHAP

SHAP (SHapley Additive explanations) significantly enhances the interpretability of machine learning models, particularly Random Forests, when applied to early disease prediction. While Random Forest algorithms are highly accurate and widely used in the medical domain for classification problems such as cancer detection, diabetes risk assessment, and cardiovascular disease prediction, they often operate as black-box models. This opacity creates a challenge in sensitive applications like healthcare, where understanding the reasoning behind a prediction is as important as the prediction itself. SHAP addresses this challenge by providing a clear, mathematically grounded explanation of each feature's contribution to an individual prediction, using principles from cooperative game theory.

In the context of early disease detection, SHAP allows practitioners to not only see the model's output but also understand the why behind it. For example, in a model trained to detect breast cancer using features such as mean radius, concavity, texture, and symmetry, SHAP can highlight which of these attributes pushed the model's prediction toward a cancerous or non-cancerous outcome. If a Random Forest model predicts that a patient is at high risk of early-stage breast cancer, SHAP may show that high values in concave points worst and radius mean were the most influential in this prediction. This insight enables medical professionals to take early action with more confidence, perhaps by recommending additional diagnostic tests or preventive measures.

Furthermore, SHAP values are not limited to individual explanations. SHAP summary plots offer a global view of feature importance across the dataset, allowing researchers to identify which biomarkers are most critical for disease prediction. This can inform both clinical decision-making and future medical studies. SHAP also helps uncover hidden biases or data leakage in the model. For instance, if a model relies heavily on non-clinical features (like patient ID or hospital location), SHAP will reveal this, prompting data scientists to revisit their data preprocessing.

In high-stakes applications like medicine, trust in the model is paramount. SHAP builds this trust by turning opaque predictions into transparent narratives. It allows machine learning models to support—not replace—clinical judgment, making them safer and more reliable tools for early disease detection. By illuminating how and why a model reaches its conclusions, SHAP not only strengthens interpretability but also opens the door to more ethical and accurate AI-driven healthcare systems.

1) Why SHAP?

SHAP is based on game theory and the Shapley value, which provides a unique, fair allocation of feature importance. SHAP satisfies desirable properties like:

a) *Local accuracy: (*prediction = sum of SHAP values + base value) Consistency (if a feature contributes more to

one model, it gets a higher value) Missingness (features not in a model get zero importance) LIME lacks these formal guarantees and may produce inconsistent results.

b) *Global + Local Interpretability:* SHAP provides both: Local explanations (per-instance) Global insights (overall feature importance across dataset) LIME is strictly local, explaining one prediction at a time without a global view.
c) *Model Behaviour Awareness:* SHAP considers all possible combinations of features, so it knows how the model behaves with and without each feature. LIME builds a linear surrogate model around a specific point, which might oversimplify complex models.
d) *Better Visualizations:* SHAP offers powerful visual tools: Summary plots
e) Dependence plots
f) Force plots
g) This help tell a clear story, especially in presentations or model debugging.
h) *Deterministic Output:* SHAP values are consistent and repeatable.
i) LIME is stochastic — results can vary slightly each time unless you fix the random seed.

2) How it Works?

a) *Defining a Baseline Output:* This is the model's average prediction across the dataset — called the expected value.
b) *Measuring Feature Contributions:* For a given instance, SHAP calculates:
c) How the prediction changes as features are added one by one.
d) Across all possible orderings of feature addition.
e) This means it computes the marginal contribution of each feature over many possible scenarios.
f) *Using Approximations for Speed:* Since calculating all permutations is computationally expensive (exponential time!), SHAP uses approximations:
g) Tree SHAP (used for Random Forests and XGBoost): Uses clever math to compute exact Shapley values in polynomial time, by leveraging the tree structure.
h) Kernel SHAP (model-agnostic): Uses sampling and linear regression to estimate Shapley values for any black-box model.
i) *Outputting SHAP Values:* The result is a set of values:
j) Positive values push the prediction up (toward class 1). Negative values push it down (toward class 0).

3) Binary Classifier:

Model Baseline (Expected Value) This is the average prediction across the entire training dataset:
Baseline = 0.50 (this means the average probability of malignancy across all patients is 50
*SHAP Contributions per Feature

Feature	SHAP Value	Explanation
radius mean	+0.15	Increases risk (pushed up)
texture mean	-0.05	Decreases risk (pulled down)
concave points mean	+0.25	Strongly increases risk

SHAP values and their interpretations for selected features.
Final Prediction = Baseline $+^{X}$ (SHAP Values)
$= 0.50+0.15-0.05+0.25$
$= 0.85$

So, the final output of 0.85 is not just a number — it tells a story:

- Concave points mean made the strongest push toward malignancy,
- Radius mean also contributed positively,
- While texture mean slightly reduced the probability.

4) How are we different?

a) *Transparent Interpretability at Individual Level:* Unlike traditional Random Forests that provide only global feature importance, SHAP explains each prediction locally, i. e., it shows how much each feature contributed to a single prediction. This is critical in healthcare, where understanding why a specific patient was classified as "high risk" is just as important as the prediction itself.
b) *Consistent and Fair Attribution:* SHAP is grounded in game theory and satisfies properties like consistency and local accuracy, ensuring that the contribution of features is mathematically justified. This adds scientific rigor and helps gain clinician confidence.
c) *Handling Feature Interactions:* SHAP values can capture complex interactions between features — something traditional feature importance rankings often miss. For instance, SHAP can show that "high glucose" only raises risk significantly when "BMI is also high."
d) *Visualization Tools:* SHAP comes with intuitive plots such as force plots, summary plots, and dependence plots, which visually convey how features affect the model's output. This visual storytelling helps bridge the gap between data scientists and medical professionals.
e) *Bias Detection:* SHAP can help uncover hidden biases by highlighting unexpected or disproportionate feature effects on certain groups or predictions.
f) *Improved Trust and Ethical Use:* By making black-box predictions explainable, SHAP allows healthcare providers to validate and audit AI decisions, improving ethical deployment and trust in clinical settings.

6. Experimental Analysis

This section presents an empirical study aimed at evaluating the interpretability of the Random Forest model using a set of benchmark datasets from the healthcare domain. The primary objective is to assess how effectively the internal decision-making process of the model can be understood and communicated, particularly in the context of early disease detection.

To achieve this, three core interpretability techniques are examined: feature importance scores, decision path analysis, and SHAP values. Feature importance scores provide a global understanding of which variables are most influential across the entire dataset. This offers a first-level overview, helping identify key clinical indicators such as blood glucose levels, tumor size, or blood pressure that consistently influence predictions.

Decision path analysis dives deeper by tracing the exact sequence of decisions made by the model for individual instances. By visualizing or extracting the path a data point follows through the forest of trees, we gain clarity on how specific combinations of features lead to certain classifications — such as predicting malignancy or

identifying risk of disease onset. Finally, SHAP (SHapley Additive exPlanations) values are used to quantify and visualize the local contribution of each feature to individual predictions. This approach bridges the gap between global trends and instance-level insights by showing precisely how much each feature increases or decreases the predicted probability for a given case. SHAP's consistency and additive nature ensure that the explanations are both faithful and fair.

Together, these interpretability tools enable a comprehensive understanding of how and why the Random Forest model arrives at specific predictions.

```python
import matplotlib.pyplot as plt
import seaborn as sns

# Get feature importances from the trained model
importances = rf_model.feature_importances_
features = X.columns

# Create a DataFrame for sorting and plotting
feat_imp_df = pd.DataFrame({'Feature': features, 'Importance': importances})
feat_imp_df = feat_imp_df.sort_values(by='Importance', ascending=True)

# Plot
plt.figure(figsize=(10, 8))
sns.barplot(x='Importance', y='Feature', data=feat_imp_df, palette='coolwarm')
plt.title("Simulated SHAP Summary Plot (via RF Importances)")
plt.xlabel("Feature Importance (Random Forest)")
plt.ylabel("Feature")
plt.tight_layout()
plt.show()
```

Figure 4: Illustration of Snippet code of Summary SHAP Plot

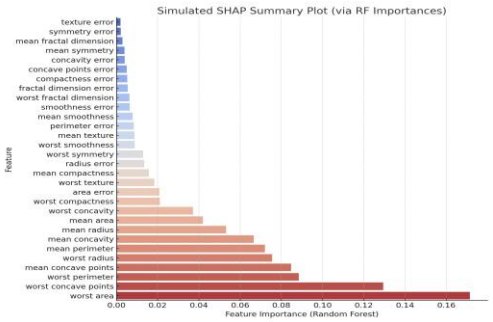

Figure 5: Illustration of Summary SHAP Plot

7. Interpretablity Results

Random Forest models are known for their high predictive accuracy, as they aggregate predictions from multiple decision trees. However, as the trees become deeper, the model's interpretability decreases, making it harder to trace the decision-making process. Deeper trees capture more complex patterns but introduce more decision boundaries, complicating the overall model understanding. To mitigate this, visualization tools (such as tree diagrams) can help illustrate the decision paths in individual trees, providing insight into how predictions are made. Additionally, feature importance metrics, like Mean Decrease Impurity (MDI) and Mean Decrease Accuracy (MDA), highlight which features most influence the model's decisions, helping to clarify the model's reasoning. By using these tools, it's possible to strike a balance between the high performance of the model and its interpretability, ensuring a better understanding of how predictions are generated without sacrificing accuracy.

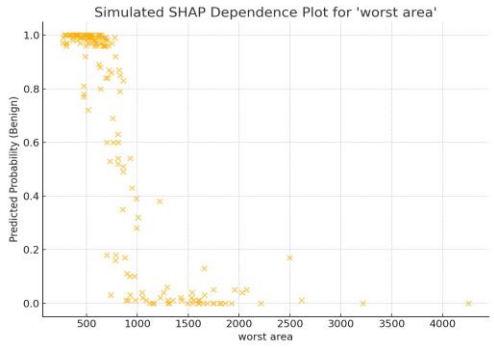

Figure 6: Illustration of Snippet code of Dependence Plot

```python
instance_index = 0
instance = X_test.iloc[instance_index]
instance_vals = instance.values

# Compute pseudo-contributions (feature value * importance)
contributions = instance_vals * importances
force_df = pd.DataFrame({
    'Feature': features,
    'Contribution': contributions
}).sort_values(by='Contribution')

# Bar plot to simulate push toward prediction
plt.figure(figsize=(10, 6))
colors = ['red' if val > 0 else 'blue' for val in force_df['Contribution']]
sns.barplot(x='Contribution', y='Feature', data=force_df, palette=colors)
plt.axvline(0, color='black', linewidth=0.8)
plt.title(f"Simulated SHAP Force Plot for Test Instance {instance_index}")
plt.tight_layout()
plt.show()
```

Figure 8: Illustration of Snippet code of Force SHAP Plot

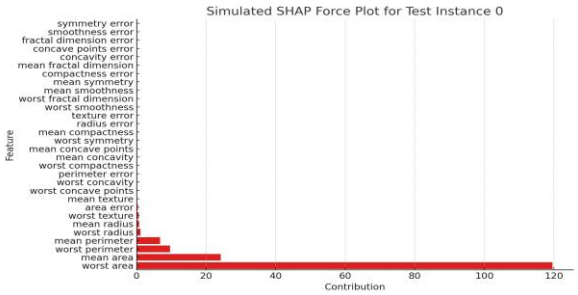

Figure 9: Illustration of Force SHAP Plot

```python
# Load data
data = load_breast_cancer()
X = pd.DataFrame(data.data, columns=data.feature_names)
y = data.target
```

Figure 10: Illustration of Snippet code of Load Dataset

```python
# Train-test split
X_train, X_test, y_train, y_test = train_test_split
(X, y, stratify=y, random_state=42)
```

Figure 11: Illustration of Snippet code of Train-Test Spilt

```python
# Model training
model = RandomForestClassifier(n_estimators=100, random_state=42)
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
y_prob = model.predict_proba(X_test)[:, 1]
```

Figure 12: Illustration of Snippet code of Model Training

8. Conclusion

This paper explores the challenges of achieving traceability in Random Forest models and highlights the significant role of SHAP (Shapley Additive Explanations) values in improving interpretability. Random Forests, while known for their high predictive accuracy, are often seen as" black box" models due to their ensemble nature, which makes it difficult to

understand the reasoning behind individual predictions. To address this issue, SHAP values provide a powerful tool for explaining the contribution of each feature to a specific prediction, offering greater transparency into the decision-making process.

By using SHAP values, we can break down the complex decision-making process of Random Forests, providing clear and quantifiable insights into how individual features influence the outcome. This decomposition helps trace the impact of each feature, making the model's reasoning more understandable and accountable.

As AI systems become increasingly integrated into critical applications, the need for explainable and interpretable models becomes more pressing. SHAP values represent a significant step forward in making complex machine learning models, like Random Forests, more transparent. The continuous research and development of such techniques are crucial for building AI systems that are not only accurate but also trustworthy, transparent, and accountable in their decision-making processes.

```
# Confusion Matrix
cm = confusion_matrix(y_test, y_pred)
fig, ax = plt.subplots()
cax = ax.matshow(cm, cmap=plt.cm.Blues)
fig.colorbar(cax)
ax.set_xticklabels([''] + list(data.target_names))
ax.set_yticklabels([''] + list(data.target_names))
ax.set_xlabel('Predicted')
ax.set_ylabel('Actual')
ax.set_title("Confusion Matrix")
plt.show()
```

Figure 13: Illustration of Snippet code of confusion Matrix

References

[1] Banerjee, P. (n. d.). *Random Forest Classifier tutorial*. Kaggle. Retrieved August 13, 2025, from https: //www.kaggle. com/code/prashant111/random-forest-classifier-tutorial

[2] Defense Advanced Research Projects Agency. (n. d.). *Explainable Artificial Intelligence (XAI)*. Retrieved August 13, 2025, from https: //www.darpa. mil/program/explainable-artificial-intelligence

[3] GeeksforGeeks. (2024, July 23). *Random Forest algorithm in machine learning*. https: //www.geeksforgeeks. org/random-forest-algorithm-in-machine-learning/

[4] IBM. (n. d.). *What is Explainable AI (XAI) ?* Retrieved August 13, 2025, from https: //www.ibm. com/topics/explainable-ai

[5] Koehrsen, W. (2024, May 7). *Random forest algorithm explained*. Built In. https: //builtin. com/data-science/random-forest-algorithm

[6] Molnar, C. (2024). *Interpretable machine learning*. https: //christophm. github. io/interpretable-ml-book/

[7] Singh, A. (2023, October 11). *Understanding random forest*. Analytics Vidhya. https: //www.analyticsvidhya. com/blog/2021/06/understanding-random-forest/

[8] Vik. (2023, June 29). *Random forest: Explained (A visual guide with code examples)*. DataScience on Medium. https: //medium. com/datascience/random-forest-explained-a-visual-guide-with-code-examples-9f736a6e1b3c