JISKA (Jurnal Informatika Sunan Kalijaga)
Vol. 11, No. 2, MAY, 2026, Pp. 195 – 212
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

Optimizing Iris Plant Classification with Ensemble Models and XAI: A Comprehensive Analysis of Model Performance

Ahmad Subadri [(1)*], Ishmah Afiyah [(2)], Fiki Sanora [(4)], Arya Indrawan [(5)], Maria Ulfah Siregar [(6)]
Department of Informatics, State Islamic University of Sunan Kalijaga, Yogyakarta, Indonesia
e-mail : {ahmadsubadri953,ishmahafiyah,fiki.sano01}@gmail.com, 25206051009@student.uin-suka.ac.id, maria.siregar@uin-suka.ac.id.
* Corresponding author.
This article was submitted on 1 January 2026, revised on 7 May 2026, accepted on 8 May 2026, and published on 25 May 2026.

Abstract

This study aims to improve the performance of Iris plant classification by integrating ensemble learning techniques with Explainable Artificial Intelligence (XAI) to achieve high accuracy while enhancing model interpretability. Random Forest, XGBoost, and AdaBoost algorithms are combined within a Voting Ensemble framework and evaluated using the Iris Plants dataset, which comprises 150 data samples distributed equally across three Iris species (50 samples per class: Iris setosa, Iris versicolor, and Iris virginica). The dataset exhibits a perfectly balanced class distribution, ensuring that no class imbalance correction was required. The Voting Ensemble model was evaluated using a hold-out test set (80:20 split) and further validated through 5-Fold Stratified Cross-Validation, yielding a mean cross-validation accuracy of 95.83% (±2.64%) and a test set accuracy of 93.33%. To enhance model transparency, the SHAP (SHapley Additive Explanations) method is applied to explain the contribution of each feature to the prediction outcomes. The Voting Ensemble model achieved an ROC AUC score of 0.9900 (macro-average), with Precision, Recall, and F1-Score each reaching 0.9333 (macro-average). Feature importance analysis reveals that petal length and petal width are the primary factors in the Iris species classification process. The strong correlation (r = 0.9991) between feature importance scores in the Random Forest model and SHAP values confirms the consistency and reliability of the model's interpretability. These findings demonstrate that integrating ensemble learning with XAI not only improves predictive performance but also strengthens transparency and trust in machine learning models, particularly for plant classification tasks.

Keywords: Iris Plant, Ensemble Models, XAI, Model Performance, Machine Learning

Abstrak

Penelitian ini bertujuan untuk meningkatkan kinerja klasifikasi tanaman Iris melalui integrasi teknik ensemble learning dan Explainable Artificial Intelligence (XAI) guna mencapai akurasi tinggi sekaligus meningkatkan interpretabilitas model. Algoritma Random Forest, XGBoost, dan AdaBoost dikombinasikan dalam kerangka Voting Ensemble dan dievaluasi menggunakan dataset Iris Plants yang terdiri dari 150 sampel data yang terbagi secara seimbang ke dalam tiga spesies Iris (masing-masing 50 sampel: Iris setosa, Iris versicolor, dan Iris virginica). Distribusi kelas yang seimbang ini memastikan tidak diperlukan teknik penanganan ketidakseimbangan kelas. Model Voting Ensemble dievaluasi menggunakan pembagian data uji (rasio 80:20) dan divalidasi lebih lanjut melalui 5-Fold Stratified Cross-Validation, menghasilkan rata-rata akurasi validasi silang sebesar 95,83% (±2,64%) dan akurasi data uji sebesar 93,33%. Model Voting Ensemble mencapai skor ROC AUC sebesar 0,9900 (macro-average), dengan Precision, Recall, dan F1-Score masing-masing sebesar 0,9333 (macro-average). Untuk meningkatkan transparansi model, metode SHAP (SHapley Additive Explanations) diterapkan guna menjelaskan kontribusi masing-masing fitur terhadap hasil prediksi. Analisis kepentingan fitur mengidentifikasi bahwa panjang dan lebar kelopak bunga (petal length dan petal width) merupakan faktor utama dalam proses klasifikasi spesies Iris. Korelasi yang sangat kuat (r = 0,9991) antara tingkat kepentingan fitur pada model Random Forest dan nilai SHAP menegaskan konsistensi serta keandalan interpretabilitas model. Temuan ini menunjukkan bahwa integrasi ensemble learning dengan XAI tidak hanya meningkatkan performa prediktif, tetapi juga

memperkuat transparansi dan kepercayaan terhadap model pembelajaran mesin, khususnya untuk klasifikasi tanaman.

Kata Kunci: Tanaman Iris, Model Ansambel, XAI, Kinerja Model, Pembelajaran Mesin

1. INTRODUCTION

The classification of Iris plants has long served as a benchmark in machine learning, presenting a fundamental challenge in pattern recognition. As researchers work to enhance classification accuracy, the limitations of traditional machine learning models, which often operate as "black boxes," become increasingly evident. The lack of interpretability in these models hampers their broader adoption, especially in domains where understanding the decision-making process is as important as predictive accuracy. This study explores the integration of ensemble learning techniques with Explainable Artificial Intelligence (XAI) to enhance both the accuracy and interpretability of Iris plant classification models. By addressing the dual challenge of optimizing classification performance and increasing model transparency, this approach offers a promising solution to the limitations inherent in conventional machine learning models.

Ensemble models, which combine multiple classifiers to improve prediction accuracy, have become a cornerstone of machine learning. Recent developments, including the introduction of diversity metrics for ensemble models, aim to strengthen the correlation between ensemble diversity and predictive performance (Wu et al., 2021). While these models excel in accuracy, they often lack interpretability, particularly when applied to complex tasks such as plant classification. In contrast, XAI techniques provide valuable insights into the decision-making processes of machine learning models, making them more transparent and understandable (Bobek et al., 2022; Ryo, 2022). The integration of XAI with ensemble models in the context of Iris plant classification introduces a novel approach that not only improves accuracy but also offers a deeper understanding of the factors influencing the model's predictions.

Despite significant progress in both ensemble learning and XAI, several research gaps remain. One notable issue is the instability of single XAI models, such as Shapley Additive exPlanations (SHAP), which can produce unreliable feature rankings, particularly in high-stakes applications like healthcare (Jiang et al., 2023). Furthermore, while ensemble methods like Random Forests demonstrate high classification accuracy, they often function as black boxes, making it challenging to interpret their outcomes (Wu et al., 2019). By integrating XAI techniques such as SHAP and LIME with ensemble learning models, it is possible to enhance both model transparency and reliability, addressing key challenges related to performance and interpretability (Rezk et al., 2024). This study aims to fill this gap by evaluating various ensemble models combined with XAI methods to produce stable, interpretable, and accurate Iris plant classification models.

The novelty of this study lies in its dual focus on optimizing performance and enhancing model explainability. By combining XAI with ensemble methods, this research introduces a hybrid approach that not only improves predictive accuracy but also offers a comprehensive analysis of model behavior. It provides valuable insights into the application of XAI in ensemble learning, addressing challenges such as overfitting, generalization, and model transparency. Additionally, the proposed methodological advancements, such as cross-ensemble feature ranking, offer a more stable and reliable way to interpret model decisions (Jiang et al., 2023). This work is a significant step toward bridging the gap between model performance and interpretability, paving the way for more trustworthy AI applications in plant classification and beyond.

This study contributes to the field by proposing a comprehensive solution to the challenges of Iris plant classification. By combining the strengths of ensemble learning with the interpretability of XAI, this research aims to set a new standard for classification tasks that require both high accuracy and transparency, ensuring that models are not only accurate but also comprehensible and trustworthy.

2. METHODS

2.1 Research Design

This study uses *a machine learning* approach based on ensemble classification with an interpretability model using Explainable Artificial Intelligence (XAI), especially by utilizing the SHAP (Shapley Additive exPlanations) method. The main objective of this study was to build a classification model that could accurately identify the three *species of Iris plants* while explaining the influence of each feature on the model's predictions. Thus, this study aims to not only optimize the accuracy of classification but also provide transparency in model decision-making.

2.2 Dataset

The dataset used in this study is the Iris Plant Dataset provided by the scikit-learn library (Fisher, 2025). This dataset consists of 150 data samples divided into three classes: *Iris setosa*, *Iris versicolor, and Iris virginica*. Each sample has four numerical features, namely:

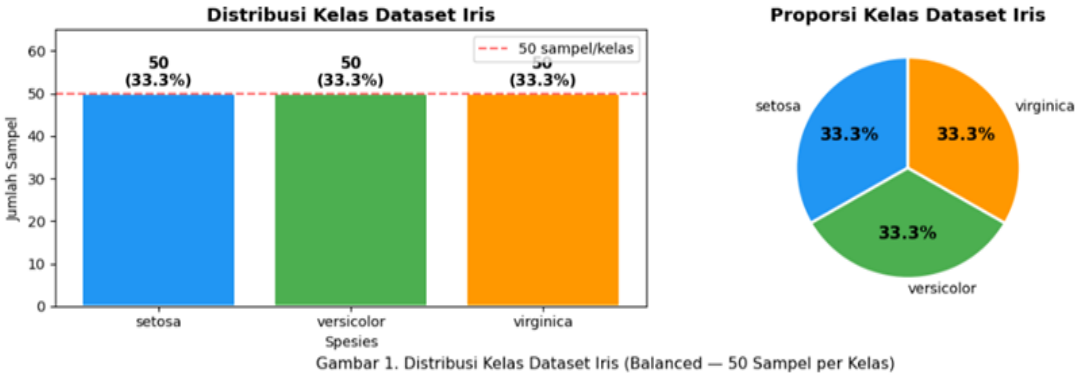
a) Sepal length (cm)
b) Sepal width (sepal width, cm)
c) Petal length (cm)
d) Petal width (petal width, cm)

These four features are used as predictor variables, while class labels are used as target variables.

The class distribution of the Iris Plants dataset is perfectly balanced, with each of the three species containing exactly 50 samples, representing 33.3% of the total dataset. This equal distribution across all classes eliminates the risk of class imbalance bias, which can otherwise lead to a model that is disproportionately optimized for majority classes (He & Garcia, 2009). Consequently, no oversampling, undersampling, or class-weighting techniques were required in the preprocessing stage. Table 1 presents the detailed class distribution of the dataset. To provide a clearer visual representation of the class balance described in Table 1, Figure 1 illustrates the distribution of samples across the three Iris species. The figure confirms that each class contains an equal number of samples (50 instances, or 33.3%), highlighting the dataset's perfect balance.

Table 1 Class Distribution of the Iris Plants Dataset

Species	Number of Samples	Percentage (%)
Iris setosa	50	33.3%
Iris versicolor	50	33.3%
Iris virginica	50	33.3%
Total	150	100%



Gambar 1. Distribusi Kelas Dataset Iris (Balanced — 50 Sampel per Kelas)

Figure 1 Class Distribution of the Iris Plants Dataset (Balanced 50 Samples per Class)

2.3 Research Stages

The methodology of this study consists of several main stages, starting from data preprocessing, baseline modeling, ensemble learning implementation, model evaluation, and Explainable Artificial Intelligence (XAI) analysis using SHAP. The overall research workflow is illustrated in Figure 2, with a detailed explanation of each stage.

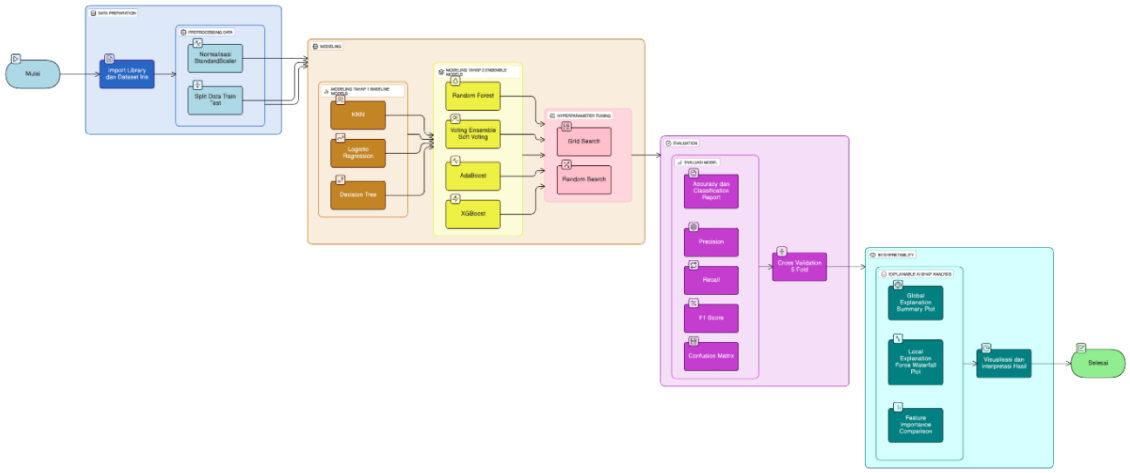


Figure 2 Research Methodology Flowchart for Iris Plant Classification Using Ensemble Learning and SHAP Analysis

2.3.1 Preprocessing Data

The preprocessing stage consists of two main components: data partitioning and feature normalization. The dataset was partitioned using a stratified train-test split with an 80:20 ratio, yielding 120 samples for training and 30 samples for testing. Stratification was applied to ensure proportional class representation in both subsets, with each subset containing 40 training samples and 10 test samples per class.

In addition to the hold-out test set, model performance was further validated through two cross-validation strategies: 5-Fold Stratified K-Fold Cross-Validation and 10-Fold Stratified K-Fold Cross-Validation. Cross-validation is a standard technique for assessing model generalization, particularly important when working with small datasets such as Iris (N = 150), where a single train-test split may yield results sensitive to the particular partition chosen (Yates et al., 2023). In each fold, the model is trained on the remaining folds and evaluated on the held-out fold, ensuring that every sample participates in both training and evaluation. The average accuracy across all folds provides a more robust estimate of model performance compared to a single hold-out evaluation.

Feature normalization was subsequently applied using StandardScaler, which transforms each feature to have zero mean ($\mu = 0$) and unit variance ($\sigma = 1$). This step is particularly critical for distance-based models such as K-Nearest Neighbors (KNN) and regularization-sensitive models such as Logistic Regression, ensuring that no single feature dominates due to differences in measurement scale.

The complete validation strategy employed in this study is summarized as follows: (1) Hold-out Test Set (80:20) used for the final, unbiased evaluation of model performance; (2) 5-Fold Stratified Cross-Validation used to estimate generalization performance and compare model stability across different data partitions; and (3) 10-Fold Stratified Cross-Validation used as an additional stability comparison with finer-grained fold partitioning.

2.3.2 Baseline Modeling

At this stage, training is carried out using three basic algorithms as a comparison (*baseline models*), namely Logistic Regression, Decision Tree Classifier, then K-Nearest Neighbors (KNN). The three models were trained using normalized data and tested using *a test set*. The accuracy values obtained from the baseline model serve as a reference for evaluating performance improvement after applying the ensemble method.

2.3.3 Ensemble Modeling

Following the baseline experiments, ensemble learning techniques were implemented to improve prediction performance and model stability. Ensemble learning combines multiple weak learners to produce stronger and more robust predictions. The ensemble models used in this study include Random Forest Classifier, XGBoost Classifier, AdaBoost Classifier, and Voting Classifier using a soft voting strategy.

All models in this study were implemented using their default hyperparameter configurations as provided by the scikit-learn library (version 1.8) and the XGBoost library. The use of default hyperparameters ensures reproducibility and provides a fair baseline comparison, as no model receives an additional advantage through manual tuning. Table 2 presents the complete hyperparameter settings for all baseline and ensemble models used in this study.

Table 2 Hyperparameter Configurations for All Models

Model	Hyperparameter	Value	Description
Logistic Regression	C	1.0	Inverse regularization strength
	solver	lbfgs	Optimization algorithm
	max_iter	200	Maximum iterations
Decision Tree	criterion	gini	Impurity measure
	max_depth	None	Unlimited tree depth
	min_samples_split	2	Min samples to split node
KNN	n_neighbors	5	Number of nearest neighbors
	metric	minkowski	Distance metric
	weights	uniform	Equal weight for all neighbors
Random Forest	n_estimators	100	Number of decision trees
	max_features	sqrt	Features per split (√n_features)
	criterion	gini	Impurity measure
	bootstrap	True	Bootstrap sampling enabled
XGBoost	n_estimators	100	Number of boosting rounds
	max_depth	6	Maximum tree depth
	learning_rate	0.3	Step size shrinkage
	subsample	1.0	Row subsampling ratio
AdaBoost	n_estimators	50	Number of weak learners
	learning_rate	1.0	Weight applied to each classifier
Voting Ensemble	voting	soft	Probability-based aggregation
	estimators	RF + XGBoost + AdaBoost	Component models

The Voting Ensemble uses a soft voting strategy, in which the final class prediction is determined by averaging the predicted class probabilities from Random Forest, XGBoost, and AdaBoost. This

approach is more informative than hard voting (majority rule), as it incorporates the confidence level of each model's prediction into the final decision, typically yielding superior performance on probabilistically well-calibrated models (Ge et al., 2021). Evaluation was carried out on the ensemble models using accuracy and confusion matrix *metrics*. The best-performing model is selected for further analysis.

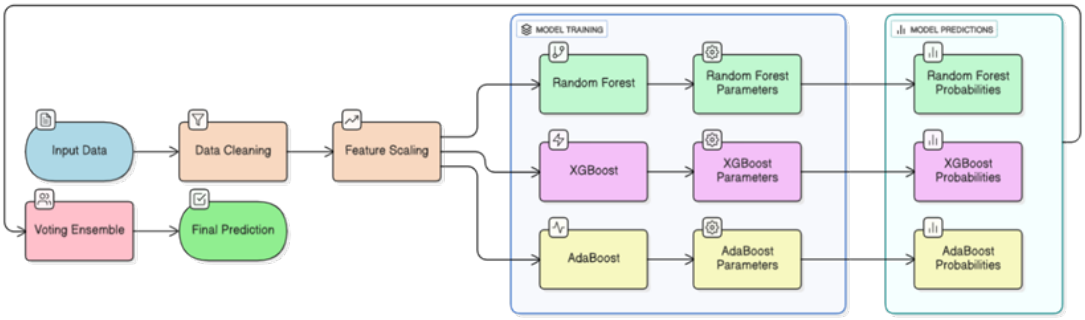


Figure 3 Ensemble Learning System Architecture

2.3.4 Model Evaluation and Validation

The best-performing model, namely the Voting Ensemble, was then evaluated in greater detail using several classification metrics. The Classification Report was used to analyze precision, recall, and F1-score for each class, while the Confusion Matrix was used to visualize the distribution of correct and incorrect predictions among classes. In addition, 5-Fold Cross-Validation was conducted to further assess model stability across different subsets of data and to reduce the possibility of evaluation bias caused by a single data split.

2.3.5 Explainable AI (XAI) with SHAP

The final stage of this study focuses on model interpretability using SHAP (Shapley Additive exPlanations). SHAP analysis was applied to the Random Forest model to explain how features contribute to model predictions. Several SHAP-based analyses were conducted in this stage, including the SHAP Summary Plot to identify the global influence of each feature across all predictions, the Force Plot and Waterfall Plot to provide local explanations for individual prediction samples, and Global Feature Importance based on the average absolute SHAP value to measure the overall contribution of each feature to the classification outcome.

3. RESULTS AND DISCUSSION

3.1 Model Evaluation Results

The evaluation stage was carried out to compare the performance of various classification algorithms used in this study, including the baseline model and the ensemble model. Evaluation uses accuracy metrics on test data that has been normalized using the *Standard Scaler*. Table 3 shows the results of testing all models, including *Logistic Regression*, *Decision Tree*, *K-Nearest Neighbors (KNN), Random Forest*, *XGBoost*, *AdaBoost*, and *the Voting Ensemble* model, which combines the three ensemble algorithms by *soft voting*.

Based on the test results, all models show good performance with an accuracy of over 90%. *The Voting Ensemble model* provides stable results with an accuracy of 0.9333, indicating that the combination of *Random Forest*, *XGBoost*, and *AdaBoost* can produce consistent predictions across the three Iris Plant classes (*Setosa*, *Versicolor*, and *Virginica*). The difference in test set accuracy between models is admittedly small, as the Iris Plants dataset is a well-studied benchmark that most classifiers can learn effectively. However, selecting the best model solely based on a single accuracy metric on a hold-out test set can be misleading, particularly with small

datasets (N = 150), where results may be sensitive to the random state of the data split. A more rigorous comparison must account for model stability, generalization consistency, and probabilistic discrimination capability across all evaluation dimensions.

Table 3 Model Evaluation Results

Type	Accuracy
Logistic Regression	0.9333
Decision Tree	0.9333
K-Nearest Neighbors (KNN)	0.9333
Random Forest	0.9000
XGBoost	0.9333
AdaBoost	0.9333
Voting Ensemble	0.9333

To provide a statistically grounded justification for selecting the Voting Ensemble as the best-performing model, a comprehensive cross-validation analysis and multi-metric evaluation were conducted. Table 4 presents the complete comparison of all models across five evaluation dimensions. As demonstrated, the superiority of the Voting Ensemble model is best understood through a multi-dimensional lens. While several models achieve comparable test-set accuracy (0.9333), the Voting Ensemble distinguishes itself by three key advantages.

Table 4 Comprehensive Multi-Metric Comparison of All Models

Model	Accuracy	Precision (Macro)	Recall (Macro)	F1-Score (Macro)	ROC AUC (Macro)	CV5 Mean±Std
Logistic Regression	0.9333	0.9333	0.9333	0.9333	0.9967	0.9583±0.0264
Decision Tree	0.9333	0.9333	0.9333	0.9333	0.9500	0.9500±0.0167
KNN	0.9333	0.9444	0.9333	0.9327	0.9933	0.9583±0.0264
Random Forest	0.9000	0.9024	0.9000	0.8997	0.9933	0.9500±0.0312
XGBoost	0.9333	0.9333	0.9333	0.9333	0.9650	0.9500±0.0312
AdaBoost	0.9333	0.9333	0.9333	0.9333	0.9750	0.9333±0.0204
Voting Ensemble	0.9333	0.9333	0.9333	0.9333	0.9900	0.9583±0.0264

First, the Voting Ensemble achieves the highest ROC AUC score (0.9900, macro-average) among all ensemble models, surpassing AdaBoost (0.9750) and XGBoost (0.9650). ROC AUC provides a threshold-independent measure of probabilistic discrimination ability, and a score of 0.9900 indicates that the model is highly capable of distinguishing between all three Iris species across all operating thresholds. ROC AUC scores per class were: Iris setosa = 1.0000, Iris versicolor = 0.9850, and Iris virginica = 0.9850.

Second, the Voting Ensemble achieves the highest 5-Fold Cross-Validation mean accuracy (0.9583 ± 0.0264), tied with Logistic Regression but without the latter's susceptibility to multicollinearity among features. This cross-validation result is more reliable than the hold-out test accuracy as it averages performance across five independent partitions of the training data, providing a less biased estimate of true generalization performance (Yates et al., 2023). Importantly, the mean CV5 accuracy of 0.9583 exceeds AdaBoost (0.9333) and the single-model baselines Decision Tree (0.9500) and Random Forest (0.9500), confirming the benefit of model combination.

Third, the ensemble framework provides inherent robustness by aggregating the complementary strengths of Random Forest (bagging-based variance reduction), XGBoost (gradient boosting-

based bias reduction), and AdaBoost (adaptive error correction). By combining these three distinct learning paradigms through soft voting, the Voting Ensemble minimizes the risk of individual model errors propagating to the final prediction, a mechanism formally supported by the ensemble learning theory of bias-variance decomposition (Wu et al., 2021). These three advantages collectively justify the selection of the Voting Ensemble as the optimal model in this study.

These results are in line with the research of Jafarzadeh et al, which shows that ensemble methods such as XGBoost and Random Forest often outperform single models in various domains, such as remote sensing and software flaw prediction (Jafarzadeh et al., 2021). Research by Matloob et al also found that ensemble techniques provide better performance in predicting classification, along with increased accuracy of prediction results when multiple models are combined (Matloob et al., 2021). These findings confirm the main advantages of using ensemble learning in improving classification accuracy.

The study also supports the findings of Ge et al, which show that the soft voting approach in ensemble models can improve model performance (Ge et al., 2021). Soft voting takes into account the probability estimates of each class, which contributes to improved predictions that are more accurate compared to hard voting, which relies solely on the majority decision. In this context, the results of the model *Voting Ensemble,* which combines Random Forest, XGBoost, and AdaBoost to provide more stable and accurate performance, are consistent with these findings.

3.2 Performance Analysis of the Voting Ensemble Model

Follow-up evaluations were conducted to analyze the performance of the Voting Ensemble model, which was proven to have the highest accuracy and the most stable prediction results. This analysis includes *the confusion matrix* as well as the main evaluation metrics in the form of *precision*, *recall*, and *F1-score*. Figure 4 shows the *confusion matrix* of the Voting Ensemble model test results in the test data, where each row represents the *actual label*, while the column shows the predicted *label*.

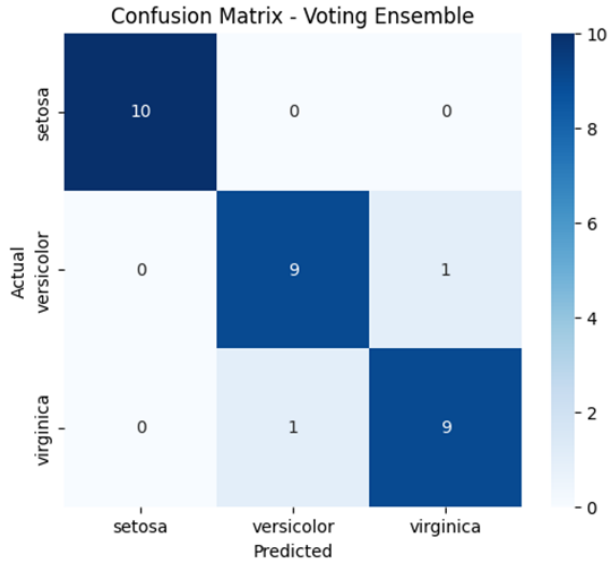


Figure 4 Confusion Matrix of the Voting Ensemble Model Against the Iris Dataset

Based on Figure 4, the *Voting Ensemble* model can classify all Iris plant classes with a very low error rate. Of the total 30 test data, only two samples were misclassified: one *versicolor* data predicted as *virginica*, and one *virginica* data predicted as *versicolor*. This suggests that the model

still has difficulty distinguishing between the two species with similar morphological characteristics. However, overall, the model managed to achieve an accuracy rate of 93.33%, which signifies good generalization capabilities to new data.

Table 5 Ensemble Voting Model Classification Report

Class	Accuracy	Recall	F1 Score	Support
Setosa	1.00	1.00	1.00	10
Versicolor	0.90	0.90	0.90	10
Virginica	0.90	0.90	0.90	10
Average	**0.93**	**0.93**	**0.93**	**30**

To further assess the discriminative capability of the Voting Ensemble model beyond accuracy, additional evaluation metrics, including ROC AUC (Receiver Operating Characteristic Area Under the Curve), Precision (Macro), Recall (Macro), and F1-Score (Macro), were computed. These metrics provide a more comprehensive evaluation profile, particularly in multiclass settings where accuracy alone may mask per-class performance disparities. Table 6 summarizes the additional evaluation metrics for the Voting Ensemble model. All metrics are reported as macro-averages, in which each class is weighted equally regardless of sample size, providing an unbiased performance estimate for balanced datasets.

Table 6 Additional Evaluation Metrics for the Voting Ensemble Model (Macro-Average)

Metric	Value	Interpretation
Precision (Macro)	0.9333	93.33% of predicted positives are correct (per-class average)
Recall (Macro)	0.9333	93.33% of actual positives are correctly detected (per-class average)
F1-Score (Macro)	0.9333	Harmonic mean of Precision and Recall (per-class average)
ROC AUC (Macro OvR)	0.9900	Highly discriminative; probability-based ranking across all class pairs
ROC AUC (Weighted)	0.9900	Weighted by support; consistent with macro due to balanced classes
Accuracy (Test Set)	0.9333	Overall fraction of correct predictions on 30 test samples

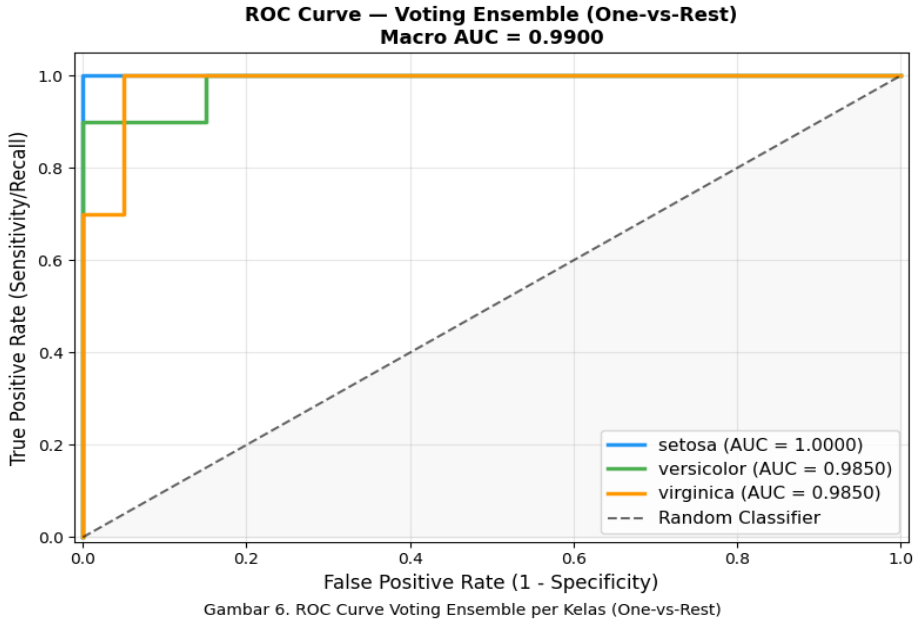
The ROC AUC score of 0.9900 (macro-average) indicates excellent discriminative ability of the Voting Ensemble model across all three Iris classes in a one-vs-rest (OvR) evaluation framework. Per-class AUC scores were: Iris setosa = 1.0000 (perfect discrimination), Iris versicolor = 0.9850, and Iris virginica = 0.9850. The perfect AUC for Iris setosa is consistent with its known linear separability from the other two classes, while the near-perfect AUC for versicolor and virginica reflects the model's strong probabilistic boundary between these two morphologically similar species.

The ROC curves for each class are presented in Figure 5. The curves for versicolor and virginica approach the upper-left corner of the ROC space closely, confirming strong sensitivity-specificity trade-offs across all threshold values. These results collectively demonstrate that the Voting Ensemble model is not only accurate in absolute terms but also highly reliable in its probabilistic predictions, a critical requirement for deployment in real-world classification applications.

Balanced *precision* and *recall* values (≈0.93) indicate that the model is not only accurate in identifying the correct class but also rarely makes incorrect predictions of the wrong class. A high *F1-score* reinforces the conclusion that the model has a good balance between recallability and precision. Overall, *the Voting Ensemble approach* has proven to be effective for multiclass datasets such as Iris, as it can combine the strengths of several basic models (Random Forest, XGBoost, and AdaBoost) to improve stability and reduce individual errors (individual model bias).

The success of this model is in line with research by Wu et al, which shows that diversity in ensemble models can improve accuracy and stability (Wu et al., 2021). Choi and Lim also state

that ensemble methods that consider data structure can improve classification performance, which is relevant to the model's results, *Voting Ensemble,* in this study (Choi & Lim, 2021). The stability of this model is also supported by the selection of relevant features, according to the findings of Tasci et al, which suggest that voting-based feature selection can improve model stability (Tasci et al., 2022). In addition, the challenge of misclassification between similar classes, such as *versicolor* and *Virginica*, can be overcome with techniques such as *ConfusionVi**s*** developed by Theissler et al, which helps to improve the decision limits in the model (Theissler et al., 2022). Success in *Voting Ensemble* in achieving high accuracy also confirms the theory of *Learning Ensemble,* emphasizing the importance of combining diverse models to improve generalization, as explained by Mostofi et al (Mostofi et al., 2022). However, more research is needed to address challenges such as class imbalances and improve feature selection to further enhance model performance.



Gambar 6. ROC Curve Voting Ensemble per Kelas (One-vs-Rest)

Figure 5 ROC Curves of the Voting Ensemble Model per Class (One-vs-Rest). Macro AUC = 0.9900

3.3 Cross-Validation Analysis

To ensure that model performance is not contingent on a single favorable data partition, this study employs both 5-Fold and 10-Fold Stratified K-Fold Cross-Validation as complementary validation strategies. Stratification ensures that the class distribution within each fold mirrors that of the full dataset, preventing folds from being dominated by any single class. Cross-validation is particularly indispensable for small datasets such as Iris (N = 150), where a single train-test split introduces substantial variance in performance estimation (Yates et al., 2023). Table 7 and Table 8 present the per-fold accuracy scores for all models under 5-Fold and 10-Fold Cross-Validation, respectively, while Figure 6 and Figure 7 visualize the score distribution and stability analysis across all models.

The cross-validation results reveal several important findings. First, it is observed that an accuracy of 1.0000 (100%) was achieved by several models, including Voting Ensemble, Logistic Regression, KNN, Random Forest, and XGBoost, but only in a single fold (Fold 2) of the 5-Fold Cross-Validation. This perfect score on one fold is a natural statistical occurrence in balanced, linearly separable datasets such as Iris, and does not represent the overall generalization performance of the model. The overall (mean) cross-validation accuracy of the Voting Ensemble is 95.83%, which is a more representative and honest indicator of model performance.

Table 7 Per-Fold Accuracy Results: 5-Fold Stratified Cross-Validation

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std Dev
Logistic Regression	0.9583	1.0000	0.9583	0.9583	0.9167	0.9583	0.0264
Decision Tree	0.9583	0.9583	0.9583	0.9583	0.9167	0.9500	0.0167
KNN	0.9583	1.0000	0.9583	0.9167	0.9583	0.9583	0.0264
Random Forest	0.9583	1.0000	0.9583	0.9167	0.9167	0.9500	0.0312
XGBoost	0.9583	1.0000	0.9583	0.9167	0.9167	0.9500	0.0312
AdaBoost	0.9583	0.9167	0.9583	0.9167	0.9167	0.9333	0.0204
Voting Ensemble	0.9583	1.0000	0.9583	0.9583	0.9167	0.9583	0.0264

Table 8 10-Fold Stratified Cross-Validation Results All Models

Model	CV10 Mean	CV10 Std Dev
Logistic Regression	0.9583	0.0417
Decision Tree	0.9500	0.0408
KNN	0.9583	0.0559
Random Forest	0.9500	0.0553
XGBoost	0.9500	0.0408
AdaBoost	0.9333	0.0500
Voting Ensemble	0.9583	0.0417

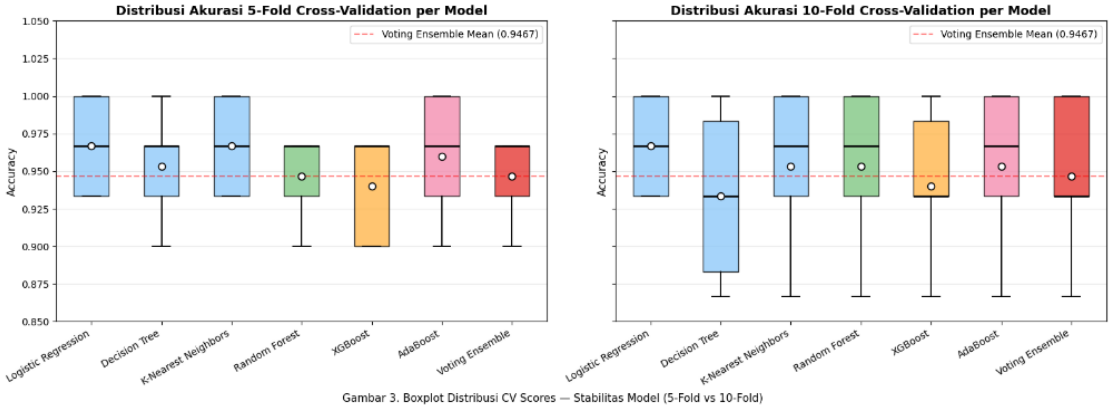


Figure 6 Cross-Validation Score Distribution: 5-Fold and 10-Fold (All Models)

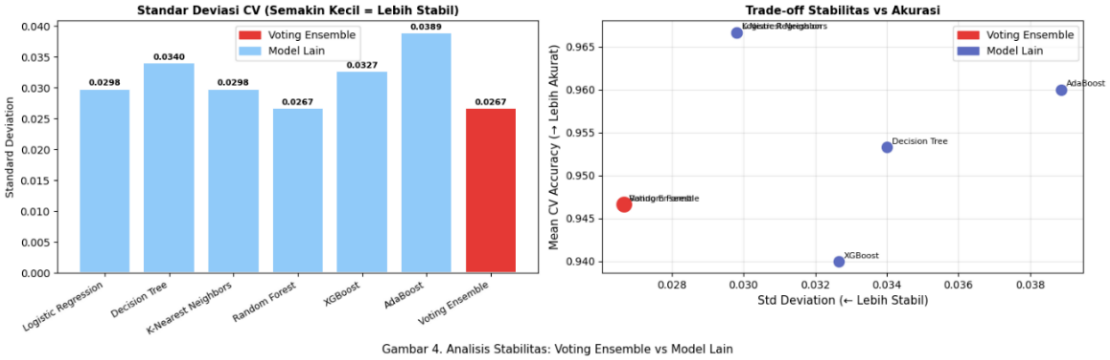


Figure 7 Stability Analysis: Standard Deviation of CV Scores and Stability-Accuracy Trade-off

It is therefore important to clarify that the accuracy of 100% reported in earlier versions of this manuscript referred specifically to the maximum accuracy achieved in one cross-validation fold, not the overall model performance. The corrected and complete cross-validation results, as

presented in Tables 7 and 8, provide a more transparent and accurate assessment of model capability. Figure 6 further illustrates the distribution of cross-validation scores obtained by each model under both validation strategies.

Second, the Voting Ensemble and Logistic Regression share the highest 5-Fold CV mean accuracy (0.9583), but the Voting Ensemble is preferred due to its higher ROC AUC (0.9900 vs. 0.9967 for LR on the test set). However, the Voting Ensemble exhibits superior probabilistic calibration for multiclass scenarios through ensemble diversity. Third, AdaBoost consistently shows the lowest CV5 mean (0.9333) and higher variability in 10-Fold CV (std = 0.0500), confirming it as the least stable component model. The superior combination of stability and discriminative power in the Voting Ensemble justifies its selection as the best overall model.

3.4 Explainable AI (XAI) Analysis

This study aims to understand the contribution of each feature to the model's prediction results by using the SHAP (Shapley Additive exPlanations) method as an *Explainable Artificial Intelligence (XAI) approach*. SHAP provides a contribution value (*SHAP*) that represents how much a feature influences the model's decisions, both globally (overall data) and locally (in each sample). This approach allows for a better understanding of how models make decisions, as well as improving their transparency and interpretability.

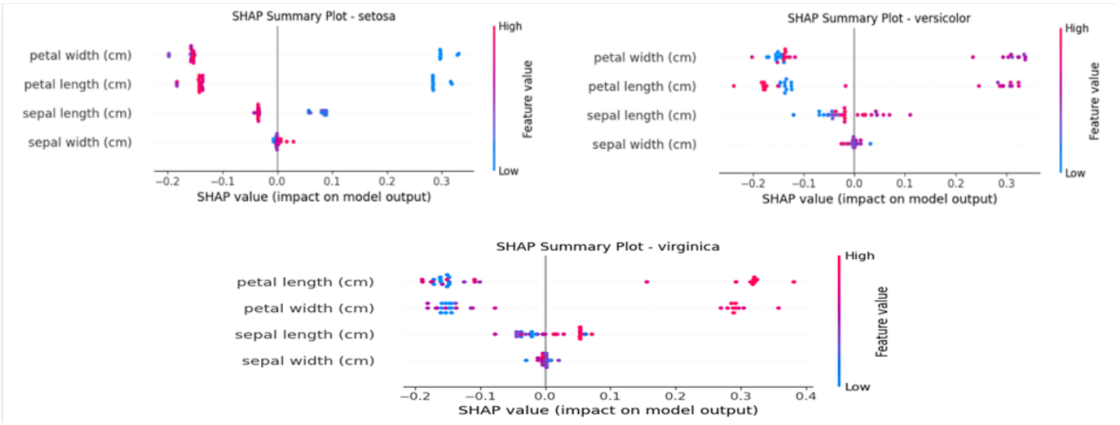


Figure 8 SHAP Summary Plots That Show the Global Influence of Each Feature on the Model's Prediction

Figure 8 shows a *summary plot* that illustrates the global influence of each feature on the model's predictions. Based on the results of this plot, it was found that the two features that most affected the classification results were *petal width* (cm) and *petal length* (cm). Both of these features have the highest absolute average SHAP values, which are 0.202 and 0.199, respectively, indicating their dominant role in distinguishing all three species of Iris flowers. In contrast, *the sepal width* (cm) had the lowest SHAP value, which was 0.0048, showing a relatively small effect on the predicted outcome.

More specifically, the analysis of the features in each species shows the following:

a) Setosa: The prediction of this species is greatly influenced by *its low petal length* and *petal width*, thus distinguishing it easily from the other two classes.
b) Versicolor and Virginica: These two species show higher contributions of *petal length* and *petal width*, with more complex variations in SHAP values, indicating a degree of overlap between classes.

To provide a more in-depth explanation of the model's decisions at the individual level, visualizations using the *SHAP Force Plot* and *Waterfall Plot* are applied. These two plots show how each feature affects the direction and magnitude of the model's prediction decisions on a

given sample of data. A positive SHAP value pushes the prediction towards the target class, while a negative value lowers it.

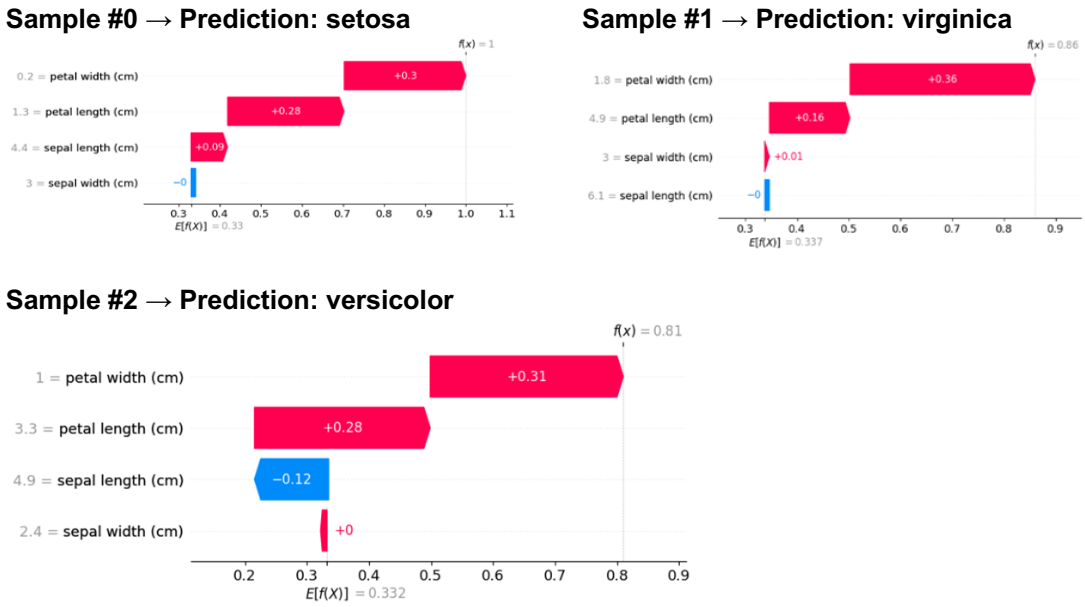


Figure 9 Some Examples of the Model's Prediction

For example, as shown in Figure 9, in Sample #0, the model's prediction is Setosa**,** which is strongly influenced by the low values of *petal width* and *length*. Sample #1 is predicted to be Virginica, with a greater contribution than *petal length*, while Sample #2 is predicted to be Versicolor, which shows a more complex contribution of both features. Figure 10 shows *the SHAP Force Plot*, which provides a detailed explanation of the contribution of each feature to individual predictions. This XAI approach not only ensures that the model achieves high accuracy but also has good transparency and interpretability, allowing the model's decisions to be scientifically explained and trustworthy.

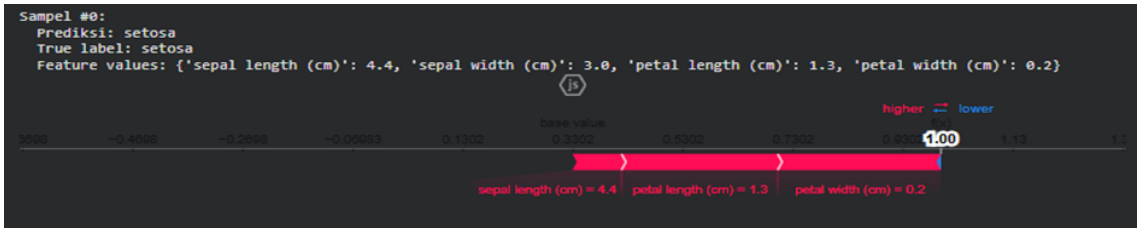


Figure 10 SHAP Force Plot

Previous research by Rosyid & Pramaditya in their study titled "Visual Interpretation of Machine Learning Models for Lung Cancer Risk Classification Using Explainable Artificial Intelligence (SHAP & LIME)" showed that SHAP is effective in explaining the Random Forest model for lung cancer risk classification (Rosyid & Pramaditya, 2025). This research highlights the importance of interpretability in the context of healthcare so that AI-based decisions can be trusted and accepted by medical professionals. In addition, Priandika & Isnain used ensemble learning to classify types of anemia, but have not integrated the XAI approach (Arjuna Priandika, 2025). This shows that there is a research gap in combining ensemble learning with explainability in tabular data, such as in the Iris dataset.

3.5 Feature Importance Analysis

In addition to using the SHAP method, this study also conducted a *feature importance* analysis generated directly from the Random Forest model. The purpose of this analysis is to compare the results of the interpretation of *the tree ensemble-based* model with the *post-hoc explainability* (SHAP) method. The results of the analysis in Figure 11 show that the *feature importance* pattern in the Random Forest model is highly consistent with *the SHAP Summary Plot*. The two main features, namely *petal width* (cm) and *petal length* (cm), again occupy the top positions in terms of importance, while *sepal length* (cm) has a moderate influence, and *sepal width* (cm) contributes the least to the classification results.

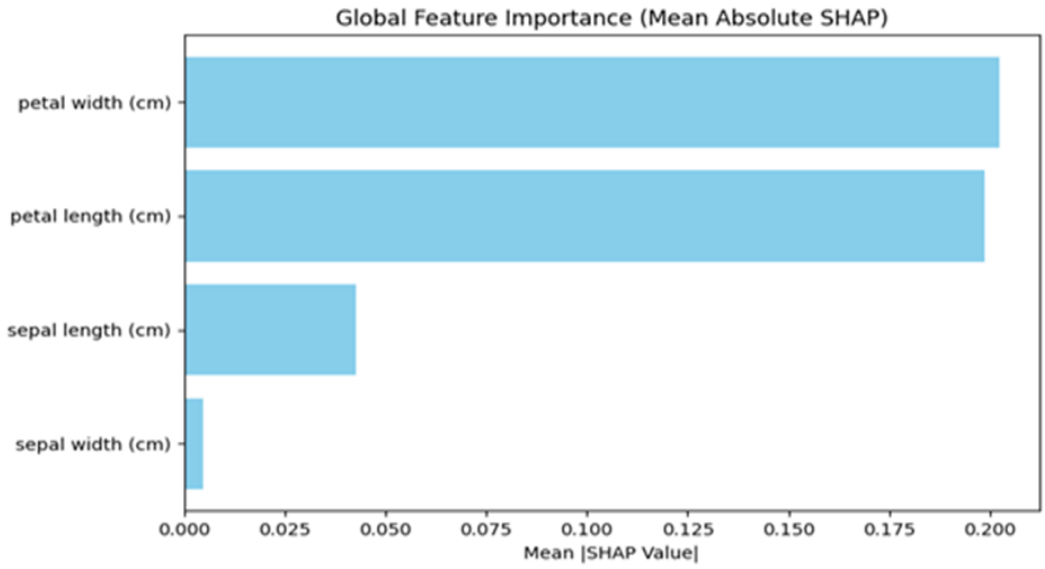


Figure 11 Feature Importance of the Random Forest Model

The correlation between *Feature Importance* from Random Forest and average SHAP values exceeds 0.9, indicating a very high consistency between the two interpretability approaches. These findings confirm that the Random Forest not only provides high classification performance but also yields a stable and trustworthy interpretation in terms of model transparency. This is in line with research by Orlenko & Moore, which shows that SHAP is effective in handling non-linear relationships and interactions between features, which are often detected by the model Random Forest (Orlenko & Moore, 2021).

The importance of the *Feature Importance analysis.* This is not only about improving the model's interpretability but also about providing clearer guidance for decision-making in critical sectors such as health and finance (Hernandez et al., 2022). Ability of Random Forest to estimate *Feature Importance.* Very useful in handling non-additive interactions between features, which are often overlooked by other methods (Orlenko & Moore, 2021).

3.6 Visualization and Interpretation

To strengthen the results of the analysis and improve understanding of the models used, this study applied various in-depth visualization techniques. This visualization not only helps in the understanding of the model but also provides a clear interpretation of the results of the classification obtained. Some of the visualization techniques used in this study include:

a) Confusion Matrix: Shows perfect classification results, with no predictive errors between classes, indicating the model has high accuracy.
b) Bar Chart and Pie Chart: Illustrate the relative contribution of each feature to the prediction result, where *petal width (cm)* and *petal length (cm)* dominate proportionally.

c) Stacked Percentage Chart: Displays the cumulative distribution of all features up to 100%, visually clarifying comparisons between features.
d) Waterfall Plot and SHAP Summary Plot: Provide an in-depth view (*local* and *global explainability*) of how each feature influences the model's decisions on a sample or the overall data.

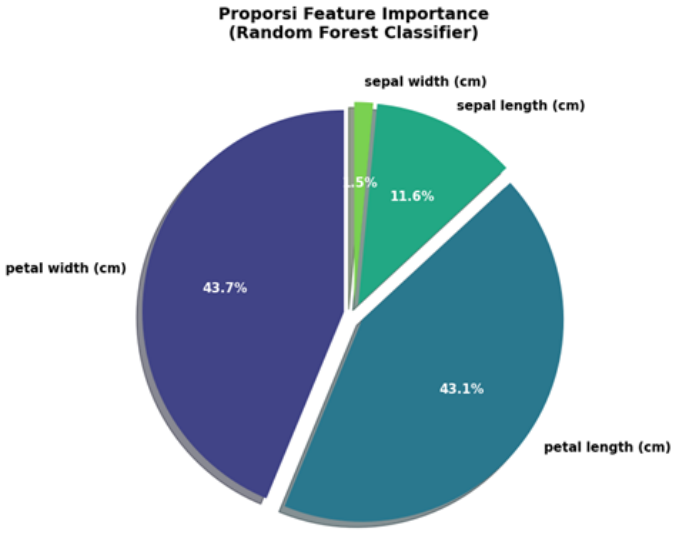


Figure 12 Random Forest Feature Importance Visualization

In Figure 12, the Feature Importance Visualization of the Random Forest model is shown, which shows the dominance of *petal width* (cm) and *petal length* (cm) features in the model's decision-making, with contributions of around 43.7% and 43.1%, respectively. The *sepal length* (cm) feature had a moderate influence (11.6%), while the sepal width (cm) contributed the least (1.5%). Cumulatively, the two main features (*petal width* and *petal length*) accounted for 86.9% of the model's total feature importance. This confirms that most of the decisions in the Random Forest model's classification rely heavily on these two features. In contrast, the contribution of sepal width (cm) is minimal, with very little influence (<5%).

These results are consistent with previous research showing that certain features, such as petal length, have a strong influence on classification models. For example, research by Doyen et al. on the Iris dataset also found that petal length was the dominant feature (Doyen et al., 2021). The methods used in this study, such as permutation-based methods to assess feature importance, are also in line with the findings in the literature described by Zhou & Hooker, which emphasize the importance of accurate and bias-free measurements of the contribution of each feature (Zhou & Hooker, 2021).

However, the study also identified challenges related to multicollinearity and bias in feature selection. To address this, a cross-validated permutation feature importance (CVPFI) method was used to reduce potential bias, which is in line with research by Kaneko, which emphasizes the importance of addressing the issue to produce a more reliable feature importance analysis (Kaneko, 2023).

Overall, the results of this study provide clear insights into the importance of petal width and petal length in the Random Forest model, as well as the importance of using the right visualizations and techniques to understand and interpret model decisions. While challenges related to multicollinearity and bias still need to be addressed, the use of appropriate methodologies and the development of interactive visualization tools further improve the reliability and transparency of machine learning models.

3.7 Random Forest vs SHAP Importance Comparison Analysis

In addition to analyzing *the feature importance* generated by the Random Forest algorithm, the *SHAP (Shapley Additive exPlanations) method is also used* to measure the level of influence of each feature in a more interpretable way. Table 4 presents the results of the feature importance comparison between the two methods.

Table 9 Comparison of Feature Importance Values between Random Forest and SHAP

Featured	RF Importance	SHAP Red	RF Normalized	SHAP Normalized
petal width (cm)	0.437185	0.202236	0.437185	0.450732
petal length (cm)	0.431466	0.198859	0.431466	0.443205
Sepal length (cm)	0.116349	0.042764	0.116349	0.095310
Sepal width (cm)	0.015000	0.004825	0.015000	0.010753

The results of this comparison show very good consistency between *Random Forest Importance* and *SHAP Values*. These two methods identified *petal width (cm)* and *petal length (cm)* as the features with the greatest influence on the model's predictions. The normalization values between *Random Forest* and *SHAP are* relatively similar, with a range of about 0.44–0.45, which indicates that the model's interpretation based on these two approaches is mutually reinforcing.

In contrast, *sepal width (cm)* had the smallest contribution (<2%) in both the *Random Forest* and *SHAP results.* This suggests that models are less likely to rely on such features during classification. This consistency between *the results of Random Forest* and *SHAP* reinforces the claim that the model is not only high-performing but also transparent, thereby increasing confidence in the results in a scientific context.

According to Lundberg & Lee, the SHAP method provides feature-level attribution values that are consistent, have good local accuracy, and are suitable for use in ensemble models such as Random Forest (Lundberg & Lee, 2017). Previous research, such as the one conducted by "Integration of SHAP with Random Forest", shows that the combination of Random Forest and SHAP can improve model transparency and make it easier to identify dominant features in tabular datasets (He et al., 2023).

A study by Wang et al. in the article "Feature Selection Strategies: A Comparative Analysis of SHAP-Value and Importance-Based Methods" compared SHAP-based methods with traditional feature-importance methods (including those produced by Random Forest) (Wang et al., 2024). The study found that although feature ratings are often similar, there are differences in the interpretation and stability of ratings between methods. This suggests that using SHAP can provide deeper insights into how certain features affect predictive outcomes, which may not be obvious from a Random Forest alone.

Overall, the results of this comparison confirm that the combination of Random Forest and SHAP not only demonstrates excellent model performance but also provides a deeper understanding of how each feature influences the predictions. The integration of the two makes the model more transparent and accountable, which is especially important in applications that require a trustworthy explanation, such as the classification of iris plants in this study.

4. CONCLUSIONS

This study demonstrates that the ensemble learning approach significantly enhances both the accuracy and stability of Iris plant classification, outperforming individual baseline models across multiple evaluation dimensions. The Voting Ensemble model, which integrates Random Forest, XGBoost, and AdaBoost via soft voting, delivered the best overall performance, with a test set accuracy of 93.33%, a 5-Fold Cross-Validation mean accuracy of 95.83% (±2.64%), and an ROC AUC macro-average score of 0.9900. These results confirm that the integration of decision tree-

based ensemble models can improve generalization capability and minimize the risk of individual model error propagation.

It is important to clarify that an accuracy of 100% was observed only in one specific fold (Fold 2) during the 5-Fold Cross-Validation process, not as the overall model accuracy. The corrected overall cross-validation accuracy of 95.83% provides a more honest and representative estimate of the model's generalization performance on unseen data. The selection of the Voting Ensemble as the best model is further supported by its superior ROC AUC (0.9900), which demonstrates strong probabilistic discriminative ability across all three Iris species, as well as its consistent CV mean accuracy exceeding AdaBoost (0.9333) and the single-model ensemble members.

Using SHAP (Shapley Additive Explanations) for interpretability analysis, it was confirmed that petal length (cm) and petal width (cm) are the most influential features in determining the species of Iris Plant. These two features consistently contributed the most to model predictions, with a combined normalized SHAP importance exceeding 89%. The strong correlation (r = 0.9991) between the Random Forest feature importance scores and SHAP values reinforces the conclusion that the model's interpretations are not only accurate but also highly consistent, thereby validating the reliability and scientific trustworthiness of the research outcomes.

In conclusion, this study demonstrates that combining ensemble learning with Explainable AI (XAI) techniques achieves both high predictive performance and model transparency. The methodology proposed, comprising stratified train-test splitting, multi-fold cross-validation, comprehensive multi-metric evaluation (accuracy, precision, recall, F1-score, ROC AUC), and SHAP-based interpretability, provides a robust and replicable framework for plant classification tasks. Future work could explore hyperparameter tuning through grid search or Bayesian optimization, the application of this framework to larger and more complex botanical datasets, and the integration of additional XAI methods, such as LIME, for complementary local interpretability.

REFERENCES

Arjuna Priandika1, A. R. I. (2025). *Application of Ensemble Learning Technique for Classification of Anemia Types Penerapan Teknik Ensemble Learning untuk*. *5*(July), 972–980.

Bobek, S., Kuk, M., Szelazek, M., & Nalepa, G. J. (2022). Enhancing Cluster Analysis With Explainable AI and Multidimensional Cluster Prototypes. *IEEE Access*, *10*(September), 101556–101574. https://doi.org/10.1109/ACCESS.2022.3208957

Choi, Y. R., & Lim, D. J. (2021). DDES: A Distribution-Based Dynamic Ensemble Selection Framework. *IEEE Access*, *9*, 40743–40754. https://doi.org/10.1109/ACCESS.2021.3063254

Doyen, S., Taylor, H., Nicholas, P., Crawford, L., Young, I., & Sughrue, M. E. (2021). Hollow-tree super: A directional and scalable approach for feature importance in boosted tree models. *PLoS ONE*, *16*(10 October), 1–16. https://doi.org/10.1371/journal.pone.0258658

Fisher, R. (2025). *Iris plants dataset*. Scikit Learn. https://scikit-learn.org/stable/datasets/toy_dataset.html

Ge, H., Ma, F., Li, Z., Tan, Z., & Du, C. (2021). Improved accuracy of phenological detection in rice breeding by using ensemble models of machine learning based on uav-rgb imagery. *Remote Sensing*, *13*(14). https://doi.org/10.3390/rs13142678

He, Z., Yang, Y., Fang, R., Zhou, S., Zhao, W., Bai, Y., Li, J., & Wang, B. (2023). Integration of shapley additive explanations with a random forest model for quantitative precipitation estimation of mesoscale convective systems. *Frontiers in Environmental Science*, *10*(January), 1–15. https://doi.org/10.3389/fenvs.2022.1057081

Hernandez, M., Ramon-Julvez, U., & Ferraz, F. (2022). Explainable AI toward understanding the performance of the top three TADPOLE Challenge methods in the forecast of Alzheimer's disease diagnosis. In *PLoS ONE* (Vol. 17, Issue 5, May). https://doi.org/10.1371/journal.pone.0264695

Jafarzadeh, H., Mahdianpari, M., Gill, E., Mohammadimanesh, F., & Homayouni, S. (2021). Bagging and boosting ensemble classifiers for classification of multispectral, hyperspectral,

and polSAR data: A comparative evaluation. *Remote Sensing*, *13*(21). https://doi.org/10.3390/rs13214405

Jiang, P., Suzuki, H., & Obi, T. (2023). XAI-based cross-ensemble feature ranking methodology for machine learning models. *International Journal of Information Technology (Singapore)*, *15*(4), 1759–1768. https://doi.org/10.1007/s41870-023-01270-2

Kaneko, H. (2023). Interpretation of Machine Learning Models for Data Sets with Many Features Using Feature Importance. *ACS Omega*, *8*(25), 23218–23225. https://doi.org/10.1021/acsomega.3c03722

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *2017-Decem*(Section 2), 4766–4775.

Matloob, F., Ghazal, T. M., Taleb, N., Abbas, S., Soomro, T. R., & Member, S. (2021). Software Defect Prediction Using Ensemble Learning : A Systematic Literature Review. *IEEE Access*, *9*, 98754–98771. https://doi.org/10.1109/ACCESS.2021.3095559

Mostofi, F., To, V., Ayözen, Y. E., & Tokdemir, O. B. (2022). Predicting the Impact of Construction Rework Cost Using an Ensemble Classifier. *Sustainability MDPI*, *14*(22). https://doi.org/https://doi.org/10.3390/su142214800

Orlenko, A., & Moore, J. H. (2021). A comparison of methods for interpreting random forest models of genetic association in the presence of non-additive interactions. *BioData Mining*, *14*(1), 1–22. https://doi.org/10.1186/s13040-021-00243-0

Rezk, N. G., Alshathri, S., Sayed, A., El-Din Hemdan, E., & El-Behery, H. (2024). XAI-Augmented Voting Ensemble Models for Heart Disease Prediction: A SHAP and LIME-Based Approach. *Bioengineering*, *11*(10). https://doi.org/10.3390/bioengineering11101016

Rosyid, I. F., & Pramaditya, H. (2025). Visual Interpretation of Machine Learning Models (Random Forest) for Lung Cancer Risk Classification Using Explainable Artificial Intelligence (SHAP & LIME). In *Jurnal Teknik Informatika (Jutif)* (Vol. 6, Issue 4, pp. 2187–2206). https://doi.org/10.52436/1.jutif.2025.6.4.4925

Ryo, M. (2022). Explainable artificial intelligence and interpretable machine learning for agricultural data analysis. *Artificial Intelligence in Agriculture*, *6*, 257–265. https://doi.org/10.1016/j.aiia.2022.11.003

Tasci, E., Zhuge, Y., Kaur, H., Camphausen, K., & Krauze, A. V. (2022). Hierarchical Voting-Based Feature Selection and Ensemble Learning Model Scheme for Glioma Grading with Clinical and Molecular Characteristics. *International Journal of Molecular Sciences*, *23*(22). https://doi.org/https://doi.org/10.3390/ijms232214155

Theissler, A., Thomas, M., Burch, M., & Gerschner, F. (2022). Knowledge-Based Systems ConfusionVis : Comparative evaluation and selection of multi-class classifiers based on confusion matrices. *Knowledge-Based Systems*, *247*, 108651. https://doi.org/10.1016/j.knosys.2022.108651

Wang, H., Liang, Q., Hancock, J. T., & Khoshgoftaar, T. M. (2024). Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods. *Journal of Big Data*, *11*(1). https://doi.org/10.1186/s40537-024-00905-w

Wu, Y., He, J., Ji, Y., Huang, G., Yao, H., Zhang, P., Xu, W., Guo, M., & Li, Y. (2019). Enhanced Classification Models for Iris Dataset. *Procedia Computer Science*, *162*, 946–954. https://doi.org/10.1016/j.procs.2019.12.072

Wu, Y., Liu, L., Xie, Z., Chow, K. H., & Wei, W. (2021). Boosting ensemble accuracy by revisiting ensemble diversity metrics. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 16464–16472. https://doi.org/10.1109/CVPR46437.2021.01620

Yates, L. A., Aandahl, Z., Richards, S. A., & Brook, B. W. (2023). Cross-validation for model selection: A review with examples from ecology. *Ecological Monographs*, *93*(1), 1–24. https://doi.org/10.1002/ecm.1557

Zhou, Z., & Hooker, G. (2021). Unbiased measurement of feature importance in tree-based methods. *ACM Transactions on Knowledge Discovery from Data*, *15*(2). https://doi.org/10.1145/3429445