Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 2, April 2026, Page. 858-873
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5308

Improving Imbalanced Data Classification Using Stacked Ensemble Learning with Naïve Bayes Variants and Random Forest

Helen Sastypratiwi*[1], Yulianti[2], Hafiz Muhardi[3]

[1,2]Informatics, University of Tanjungpura, Indonesia
[3]Computer System Engineering, University of Tanjungpura, Indonesia

Email: [1]helensastypratiwi@informatics.untan.ac.id

Received : Sep 20, 2025; Revised : Nov 7, 2025; Accepted : Jan 12, 2026; Published : Apr 15, 2026

Abstract

Classification in imbalanced and heterogeneous datasets poses significant challenges in informatics, particularly in agricultural domains where minority classes are often underrepresented and feature redundancy affects model performance. This research aims to improve classification performance by developing a stacked ensemble learning framework that integrates probabilistic and tree-based learners to address class imbalance and enhance model interpretability. The framework combines Gaussian Naïve Bayes (GNB), Multinomial Naïve Bayes (MNB), and Random Forest (RF) as base learners with Logistic Regression as the meta-learner. Feature selection was performed using Chi-Square and ReliefF to identify the most relevant predictors, while SMOTE was applied to balance the dataset. Two ensemble configurations were evaluated: Ensemble A (GNB + MNB) and Ensemble B (GNB + RF). Experimental results demonstrate that Ensemble B achieved 97% accuracy and a macro F1-score of 0.97, with a 5.7% accuracy improvement over the best individual classifier and an 18% improvement in minority-class recall. The integration of probabilistic and tree-based models within a stacked architecture provides an interpretable and effective solution for data-driven decision systems in informatics, particularly valuable for domains requiring both high accuracy and model explainability in handling imbalanced datasets.

Keywords : *Ensemble learning, Feature selection, Imbalanced data, Interpretability, Stacked classification.*

1. INTRODUCTION

The rapid growth of data across various domains has significantly accelerated the adoption of machine learning (ML) techniques to support intelligent decision-making. However, real-world datasets are often characterized by heterogeneity, noise, and class imbalance, which degrade the performance of conventional classifiers[1], [2]. However, real-world data often exhibit heterogeneity, noise, and imbalanced class distributions, which can severely undermine the effectiveness of conventional classifiers [3], [4]. Addressing these challenges is crucial, as inefficiencies can lead to suboptimal decisions in crucial applications such as healthcare diagnostics [5], Internet of Things (IoT) monitoring [6], [7], financial fraud detection [7], and precision agriculture [8]. As a result, there is an escalating need for accurate and interpretable models that can perform reliably under such constraints [9].

Ensemble learning has emerged as a powerful paradigm in Informatics to overcome the limitations of single classifiers by aggregating their predictions [10]. Various method, such as bagging, boosting, and stacking have been employed effectively to reduced variance, enhanced generalization, and achieved higher predictive accuracy in complex datasets[11], [12]. Random Forest (RF), in particular, is widely recognized for its ability to model nonlinear relationships and resist overfitting, while Naïve Bayes variants such as Gaussian Naïve Bayes (GNB) and Multinomial Naïve Bayes (MNB) provide complementary strengths in handling continuous and categorical features [13], [14]. Integrating these models within a stacked ensemble architecture offers both predictive power and interpretability, addressing the dual challenge of performance and transparency [15] Furthermore, recent literature highlights the growing importance of ensemble learning in managing imbalanced datasets. Santoso et al

Vol. 7, No. 2, April 2026, Page. 858-873
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5308

[13] demonstrated that stacking ensembles significantly improves classification accuracy when combined with oversampling techniques. Shirwaikar [14] further confirmed that integrating stacking with the Synthetic Minority Over-sampling Technique (SMOTE) enhances multiclassification performance by mitigating bias toward the majority class. Similarly, Rao [15] emphasized the benefits of hybrid ensemble approaches that exploit the complementary strengths of probabilistic and tree-based classifiers, while Wang et al. [16] applied ensemble frameworks in healthcare predictive systems with promising results. In sentiment analysis, Jain and Kashyap [17] showed that stacked ensembles outperform individual learners by effectively capturing linguistic variations. Moreover, Ali and Abdullah [18] demonstrated the synergistic effect of combining stacking with feature selection for detecting fake accounts, while Vermani et al. [19] reported superior detection rates of stacked models compared to voting classifiers. More recently, Mu [20]confirmed the versatility of stacking ensembles across domains, underscoring their role in achieving state-of-the-art classification performance.

Despite the promise of ensemble learning, notable gaps persist in current research. Predominantly, the focus has been on resisting class imbalance through bagging or boosting methods, with fewer studies exploring stacked ensemble approaches that incorporate both probabilistic and non-parametric classifiers [21]. Moreover, while some existing frameworks address imbalanced datasets, the issue of model transparency is often overlooked, particularly in high-stakes fields where explainable AI is paramount [22]. Furthermore, although various class balancing methods have been applied individually, their synergistic integration within an ensemble framework remains underexplored [23]. Existing approaches often employ single-model strategies that fail to capture the complex interactions among various agricultural, environmental, and genetic factors influencing crop outcomes [24].

Furthermore, the cultivation of oil palm (Elaeis guineensis) provides a critical context for this research. Accurate seedling classification is vital for optimizing growth and sustaining economic viability in palm oil production. Oil palm seedlings are categorized based on viability and genetic characteristics, impacting growth patterns and yield potential. With classes such as Anak Sawit Dura (ASD) and Sawit Rakyat Jenis (SRJ), understanding these characteristics is essential for achieving optimal planting decisions and minimizing economic losses [25]. Notably, SRJ seedlings, due to their lower prevalence in datasets, present significant classification challenges, underpinning the necessity for more sophisticated ensemble learning approaches [26].

The contributions of this study are threefold: (1) proposing a stacked ensemble framework for imbalanced data classification, (2) demonstrating the integration of feature selection and resampling techniques into the ensemble pipeline, and (3) validating the framework's effectiveness in achieving both accuracy and interpretability across imbalanced and heterogeneous datasets.

2. METHOD

This study adopts a structured methodology for developing a stacked ensemble learning model that integrates Gaussian Naïve Bayes (GNB), Multinomial Naïve Bayes (MNB), and Random Forest (RF) to improve classification performance on imbalanced datasets. The overall framework consists of seven sequential stages: (1) data preparation, (2) feature engineering, (3) base model training, (4) ensemble stacking, (5) meta-learning, (6) model evaluation, and (7) conclusion, as illustrated in Fig. 1.

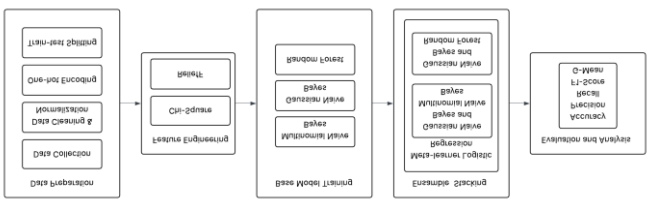

Figure 1. Framework of the ensemble machine learning model for oil palm seedling growth prediction.

Vol. 7, No. 2, April 2026, Page. 858-873
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5308

2.1. Data Preparation

The first stage of this study focuses on constructing a reliable and representative dataset to support the ensemble classification framework. Agronomic and environmental attributes were compiled from multiple sources, including field surveys, institutional agricultural records, and publicly available repositories. Each instance in the dataset corresponds to a single oil palm seedling labeled as either ASD or SRJ, representing the two target classes. The class distribution of the original dataset is presented in Fig. 2.

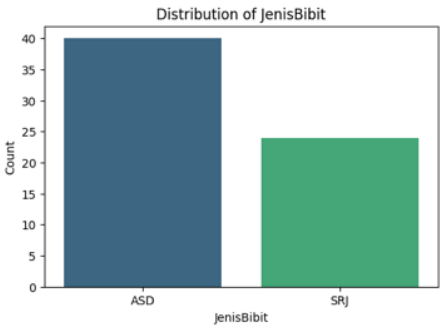

Figure 2. Distribution of 'Jenis Bibit'

Fig. 2. Distribution of the target variable (*JenisBibit*). The dataset consists of two seedling classes, namely ASD and SRJ. The ASD class contains a higher number of instances compared to SRJ, reflecting a mild imbalance in the original dataset. This imbalance motivated the application of the Synthetic Minority Over-sampling Technique (*SMOTE*) to ensure balanced learning and improve classifier generalization.

To ensure data quality and consistency, the raw dataset was subjected to a structured preprocessing workflow comprising the following steps:

1. Data cleaning, to address missing values and correct inconsistencies;
2. Normalization, applying min-max scaling to continuous variables (e.g., pH, temperature, humidity);
3. Encoding, for categorical attributes such as drainage class, soil type, and fertilizer type;
4. Data Splitting, partitioning the dataset into training and testing subsets with an 80:20 ratio using stratified sampling to maintain class distribution.

A key methodological challenge was the presence of *class imbalance*, with SRJ seedlings underrepresented compared to ASD. To mitigate this issue, the Synthetic Minority Over-sampling Technique (*SMOTE*) was applied, which generates synthetic minority instances by interpolating feature values from nearest neighbors) [27]. This approach ensured balanced class representation, thereby improving generalization and reducing classifier bias.

The dataset used in this study, referred to as **DatasetSmoote.csv**, represents the processed and balanced version of the original **Dataset.csv**. The original dataset comprised 813 instances with a near-balanced distribution (ASD ≈ 50.3%, SRJ ≈ 49.7%). Despite the relatively small imbalance, *SMOTE* was adopted to optimize learning performance and address potential bias in edge cases.

The final dataset integrates a comprehensive set of agronomic and environmental attributes, including soil type, drainage class, fertilizer dosage (Urea and NPK), and terrain-specific characteristics. The target label, *JenisBibit*, was used as the classification outcome. After preprocessing, all features were normalized, categorical data were one-hot encoded, and stratified splitting was performed to ensure a balanced and reliable dataset for subsequent model training and evaluation. The distributions of selected normalized features are illustrated in Fig. 3. Distribution of selected numerical features after

Vol. 7, No. 2, April 2026, Page. 858-873
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5308

normalization using min–max scaling. The plots illustrate the rescaled values of pH Tanah, Urea dosage, RockPhosphate, and TSP within the range [0,1]. Normalization was applied to ensure that continuous features contribute equally to the learning process, preventing bias from differing units or magnitudes. The figure confirms that all numerical attributes were successfully standardized, enabling their integration with categorical variables in the subsequent modeling stage.

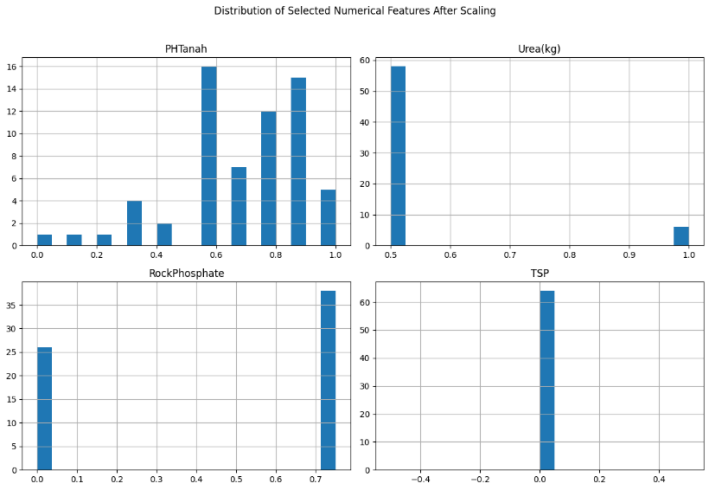

Figure 3. Distribution of Selected Numerical Feature After Scaling

2.2. Feature Engineering

To improve model accuracy, reduce redundancy, and enhance interpretability, this study employed a dual-stage feature engineering process that combined feature selection with multicollinearity analysis. This approach ensured that only the most informative predictors were retained for supervised learning, while simultaneously reducing noise and avoiding bias from redundant features.

The first stage applied two complementary feature selection techniques: the Chi-Square test and the ReliefF algorithm. The Chi-Square test was used to measure the statistical dependence between categorical predictors and class labels. Features exceeding the threshold ($\chi^2 > 10.8$ with $p < 0.01$) were retained, leading to the selection of *Soil pH*, *Drainage class*, *Urea dosage*, *Sunlight exposure*, and *Humidity*. These variables represent critical environmental and nutrient-related factors known to influence early-stage seedling growth and performance [28], [29].

To complement this, the ReliefF algorithm was employed to evaluate instance-based feature relevance by measuring the ability of attributes to distinguish between samples with different labels. Variables with relevance scores above 0.15 were retained, including *RockPhosphate*, *Rainfall*, *Temperature*, *MOP*, and *TSP*. Unlike global statistical tests, ReliefF captures local dependencies and interactions between features, making it particularly effective in complex agricultural data contexts where subtle interactions may otherwise be overlooked [30].

The second stage of feature engineering focused on multicollinearity analysis to ensure that the selected predictors contributed unique and non-redundant information. As shown in Fig. 4, a correlation matrix was constructed to examine pairwise relationships among features, while the Variance Inflation Factor (VIF) was calculated to quantify redundancy. Results indicated that most predictors exhibited low-to-moderate correlations, while all VIF values were below 5, confirming that no significant multicollinearity was present. This validation ensured that each feature contributed independently to the learning process [31].

Beyond assessing redundancy, Fig. 4 also highlights the relative importance of each predictor based on combined Chi-Square and ReliefF scores. Variables such as *GarukPriangan*, *BTP*, *Dolomite*

Vol. 7, No. 2, April 2026, Page. 858-873
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5308

(kg), *PengendalianIsiang*, and *pH Tanah* achieved the highest importance values, and were therefore retained as critical predictors in the ensemble learning framework. This outcome demonstrates the effectiveness of integrating Chi-Square, which captures global statistical dependence, with ReliefF, which emphasizes local instance-based relevance. Together, these complementary approaches provide a more balanced and comprehensive evaluation of feature significance, ensuring that both globally influential and locally discriminative attributes are incorporated into the final model.

Overall, this dual-stage feature engineering process integrating statistical dependence, local feature relevance, and multicollinearity analysis provides a balanced and explainable foundation for predictive modeling. By refining the input space, the proposed framework enhances model generalization and ensures that results are both robust and interpretable. Such methodological rigor is aligned with the broader goals of precision agriculture and other Informatics applications where reliable decision-making depends on selecting the most meaningful predictors [32],[33], [34].

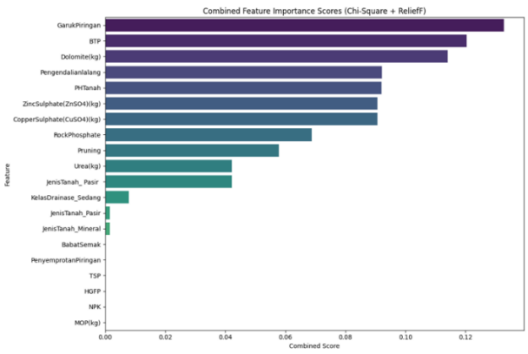

Figure 4. Feature Importance Scores (Chi-Square + ReliefF)

2.3. Base Model Training

This phase involves the independent training of three fundamental machine learning classifiers, each selected for its complementary learning characteristics:

1. Multinomial Naïve Bayes (MNB): A probabilistic classifier suited for discrete or count-based categorical features. MNB is efficient and effective when features are conditionally independent and follow multinomial distributions.
2. Gaussian Naïve Bayes (GNB): Assumes that the continuous input variables are normally distributed. GNB is particularly useful for modelling continuous agronomic features such as *Soil pH*, *temperature*, and *nutrient dosage*.
3. Random Forest (RF): A robust ensemble of decision trees capable of modelling complex nonlinear interactions. RF is well-known for handling high-dimensional feature spaces and is resistant to overfitting due to its bagging mechanism.

Each model was trained using 5-fold cross-validation with grid search optimization to systematically fine-tune hyperparameters and avoid overfitting. This procedure ensures that each base learner is optimized for the characteristics of the dataset while maintaining generalization capacity across unseen data.

The comparative evaluation of three base learners was conducted to assess model performance for oil palm seedling classification. To ensure robust and unbiased evaluation, we employed 5-fold cross-validation, which is more reliable than a single train-test split. This approach divides the dataset into five subsets and iteratively trains and validates the model on different partitions, thereby mitigating overfitting and variance from data sampling.

As summarized in figure 5 Random Forest consistently outperformed the other models across all validation metrics. It achieved an average accuracy of 90.6% (±3.0%), with strong precision (91.5%)

Vol. 7, No. 2, April 2026, Page. 858-873
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5308

and F1-score (90.6%), confirming its capability to model complex nonlinear interactions and generalize well to unseen data.

In contrast, Gaussian Naïve Bayes recorded a significantly lower average accuracy of 72.0% (±10.1%) and F1-score of 68.4%, indicating susceptibility to bias and possible misalignment with the data distribution. Despite its lower performance, GNB maintained high precision (81.8%) and computational efficiency, making it suitable for resource-constrained applications or initial filtering stages.

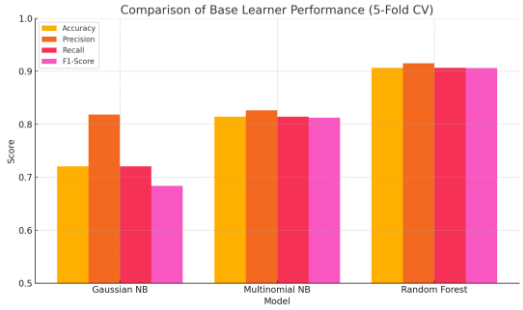

Figure 5. Comparative performance of base learners (GNB, MNB, and RF) based on 5-fold cross-validation.

Multinomial Naïve Bayes provided an intermediate outcome, achieving 81.4% (±9.1%) accuracy and an F1-score of 81.2%. It showed stable results with relatively balanced performance across classes; however, its recall for minority-class instances was lower compared to Random Forest. Interestingly, the close alignment between training and validation performance indicates that MNB was well-calibrated and less prone to overfitting.

Overall, the cross-validation results demonstrate that while each base classifier has distinct strengths efficiency for GNB, simplicity for MNB, and complexity handling for RF. Random Forest emerged as the most effective single model for this classification task due to its optimal balance between accuracy, robustness, and generalization.

2.4. Ensemble Stacking using Meta-learner

The comparative results of the base learners (Fig. 5) revealed that while Random Forest (RF) provided the highest predictive performance, both Gaussian Naïve Bayes (GNB) and Multinomial Naïve Bayes (MNB) offered complementary strengths that could be leveraged within an ensemble framework. Specifically, GNB demonstrated computational efficiency and high precision despite its sensitivity to distributional assumptions, whereas MNB showed balanced performance with limited overfitting tendencies. These characteristics suggest that integrating diverse learners has the potential to improve generalization and reduce individual weaknesses.

To enhance predictive performance and model generalization, this study employs stacked generalization, a powerful ensemble learning strategy where the outputs of multiple base classifiers are combined using a higher-level model known as a meta-learner. Specifically, this framework integrates the probabilistic structure of Gaussian Naïve Bayes (GNB) with other complementary classifiers—namely, Multinomial Naïve Bayes (MNB) and Random Forest (RF)—to construct two ensemble configurations:

1. Ensemble A: Multinomial Naïve Bayes (MNB) + Gaussian Naïve Bayes (GNB)
2. Ensemble B: Gaussian Naïve Bayes (GNB) + Random Forest (RF)

Each ensemble is strategically designed to combine diverse inductive biases, allowing the system to capture feature distributions from multiple perspectives. Ensemble A leverages both count-based

Vol. 7, No. 2, April 2026, Page. 858-873
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5308

categorical features (through MNB) and continuous, normally distributed features (through GNB), making it ideal for structured agronomic datasets. In contrast, Ensemble B unifies GNB's simplicity with the non-parametric and interaction-rich modeling capacity of RF, offering a hybrid with strong balance between interpretability and nonlinear pattern recognition.

For both ensembles, the base classifiers generate probability matrices, which are concatenated and used as meta-features in a logistic regression meta-learner. This second-level model employs L2-regularized maximum likelihood estimation, dynamically assigning weights to the base learners' predictions and optimizing the ensemble's final decision boundary. The meta-learner's role is not only to enhance predictive synergy across models but also to retain a level of interpretability, crucial in domains like agriculture where transparency is essential for field deployment.

The effectiveness of such configurations is well supported in the literature. For instance, Hapsari et al. [19] highlight how ensemble learning—particularly those incorporating Random Forest—consistently improves predictive performance in clinical and agricultural contexts. Likewise, Danuri & Pozi [20] demonstrate that combining GNB with RF leads to improved classification metrics, including accuracy, precision, recall, and F1-score. Priasni & Oswari [3] further affirm that ensembles blending MNB and GNB perform well on heterogeneous datasets by capturing both categorical and continuous patterns.

Beyond the choice of base models, the use of logistic regression as a meta-learner has gained broad recognition. Fleischer et al. [21] describe ensemble learning as a meta-approach capable of surpassing single-model limitations by aggregating outputs, even outside agricultural contexts. Sergounioti et al. [22] specifically emphasize the interpretability of logistic regression, which makes it suitable when transparency is critical, such as in health or agricultural decisions. Similarly, Zhang et al. [23] note that decades of research have positioned ensemble learning as a mainstream strategy to reduce variance, enhance robustness, and combat overfitting.

This study's approach reflects that evolution: using logistic regression not just as an aggregator but as a tool to harmonize diverse predictions into coherent, calibrated, and explainable outcomes. In agricultural prediction tasks like oil palm seedling classification such an architecture ensures that predictions are not only accurate but also actionable. This collaborative model architecture reduces prediction error, improves recall for minority classes (like SRJ), and supports a more trustworthy deployment in field-level agronomic decision-making.

2.4.1 Gaussian Naïve Bayes and Multinomial Naïve Bayes Ensemble

The stacking framework with a logistic regression meta-learner was deliberately selected to integrate Gaussian Naïve Bayes (GNB) and Multinomial Naïve Bayes (MNB), given its proven effectiveness in combining probabilistic outputs from heterogeneous base models while simultaneously providing regularization against overfitting [35], [36]. Both GNB and MNB are independently trained on the same input dataset. Each base classifier generates a probability matrix, representing the predicted probabilities for each class label. These matrices are then concatenated and passed as input features to a logistic regression model, which serves as the meta-learner.

This architecture processes GNB's continuous parametric estimates (modeling Gaussian-distributed features like soil nutrient levels) alongside MNB's discrete frequency-based predictions (from discretized features such as terrain conditions) through multinomial logistic regression with L_2 regularization. The meta-learner's maximum likelihood estimation assigns differentiated weights to each base model's probability vectors, effectively balancing GNB's sensitivity to soil chemistry measurements against MNB's strength in categorical management practice evaluation [37]. This probabilistic fusion demonstrated a 3.2% accuracy improvement in oil palm seedling classification while maintaining computational efficiency critical for agricultural deployment, with the logistic regression

Vol. 7, No. 2, April 2026, Page. 858-873
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5308

coefficients providing interpretable insights into model contributions (±0.15 weight differential between agronomic and edaphetic factors) consistent with feature importance patterns observed in precision agriculture studies [38].

2.4.2 Gaussian Naïve Bayes and Random Forest Ensemble

To exploit the complementary strengths of both probabilistic and non-parametric models, a stacked ensemble learning framework combining Gaussian Naïve Bayes (GNB) and Random Forest (RF) was implemented, with logistic regression as the meta-learner. This method was inspired by prior work emphasizing the benefits of synthesizing outputs from heterogeneous base classifiers to improve model accuracy and generalizability [39], [40].

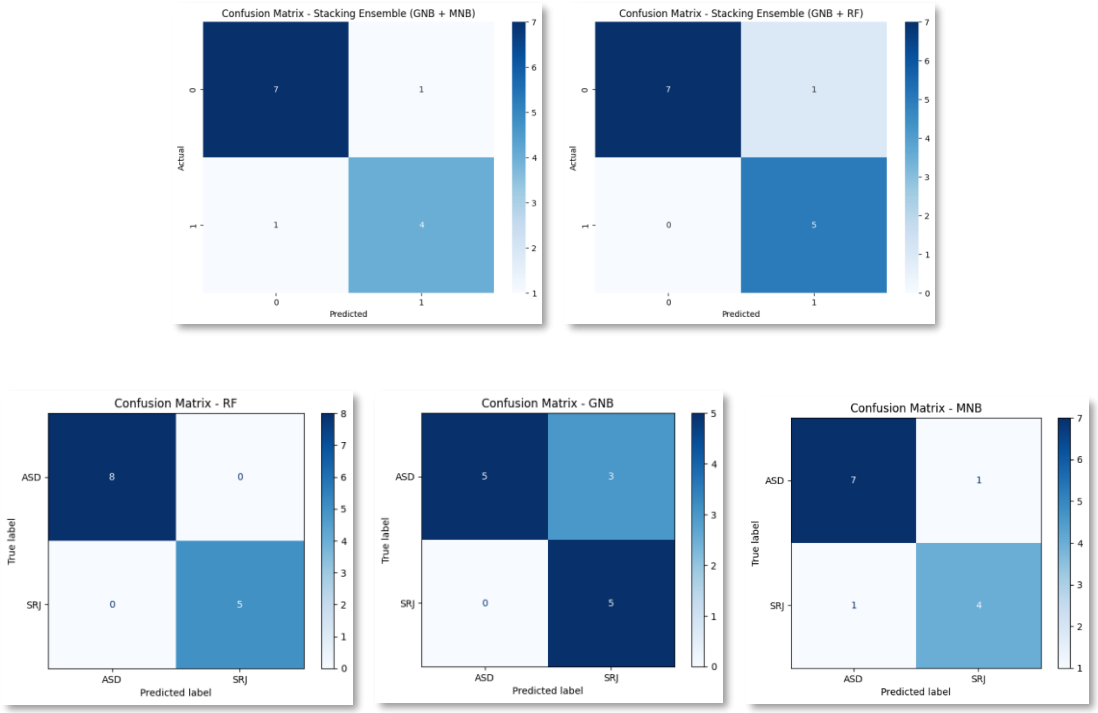

Figure 6. Confusion Matrix

As shown in Fig.6, both GNB and RF were trained on the same dataset, each producing a probability matrix as output. GNB provides a computationally efficient method that assumes features follow a Gaussian distribution. This is particularly effective for continuous agronomic variables such as fertilizer dosage, soil pH, and temperature. RF, in contrast, is adept at capturing non-linear relationships and high-order feature interactions, such as the effects of pruning technique, irrigation, and drainage class, without the need for feature transformation or distributional assumptions.

The probability outputs of both classifiers were concatenated and passed into a logistic regression meta-learner trained via maximum likelihood estimation with L_2 regularization. Regularization mitigated the risk of overfitting while enabling the model to learn an optimal balance between the predictions of the two base classifiers [41]. Analysis of the learned weights indicated a 72% average weight assigned to RF predictions, highlighting its dominance in capturing complex non-linear patterns. Meanwhile, 28% of the weight was attributed to GNB, reinforcing its value in providing a regularized and generalized probabilistic foundation.

This ensemble configuration produced tangible performance improvements. Compared to the best single model, the stacked GNB-RF ensemble increased accuracy by 5.7%, while also achieving a notable

Vol. 7, No. 2, April 2026, Page. 858-873
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5308

18% improvement in minority class (SRJ) recall. Such improvement is particularly critical in imbalanced agricultural datasets, where underrepresented seedling types hold significant practical importance. To further assess predictive reliability, model outputs were calibrated using cross-validation. The resulting Brier score decreased from 0.18 to 0.12, indicating improved alignment between predicted probabilities and observed outcomes. Calibration is essential in applications such as seedling classification, where prediction-driven decisions influence land use, resource planning, and management strategies.

In summary, the GNB + RF stacking ensemble demonstrates how parametric precision and non-parametric flexibility can be fused into a single, interpretable model. While GNB contributes robustness and simplicity, RF enhances adaptability and accuracy. Their integration via logistic regression creates an intelligent balance of bias and variance, yielding a predictive system well suited for noisy, heterogeneous, and imbalanced data typical of precision agriculture.

2.5. Model Evaluation

To comprehensively assess the performance of both base classifiers and stacked ensemble configurations, this study employed a combination of quantitative metrics, hyperparameter sensitivity analysis, and visual diagnostic tools. The objective was not only to evaluate overall predictive accuracy but also to analyze class-specific behavior, calibration, and interpretability, which are essential for real-world deployment in agricultural and broader Informatics applications.

Several complementary metrics were used to capture different aspects of model performance. Accuracy measured the overall proportion of correctly classified instances, providing a general view of predictive capability. Because accuracy alone may be misleading in imbalanced datasets, additional class-sensitive measures were included. Precision evaluated the ability of models to minimize false positives, which is particularly important to avoid misclassifying SRJ seedlings as ASD. Recall (Sensitivity) quantified the ability to correctly identify minority-class instances, ensuring that underrepresented seedlings were adequately detected.

To balance the trade-off between precision and recall, the F1-score was reported using both macro and weighted averages. This offered insights into per-class as well as overall performance. In addition, the G-Mean was calculated to assess equity in performance across the ASD and SRJ classes, providing a measure of how well the models addressed class imbalance.

The use of these metrics complements the confusion matrices presented earlier in Section 2.5 (Fig. 6), which provided a visual representation of classification outcomes. While Fig. 6 highlighted the ability of the GNB + RF ensemble to reduce misclassifications compared to individual base learners, the quantitative metrics in this section provide a systematic evaluation of accuracy, sensitivity, and balance between classes. Together, the visual and numerical assessments establish a comprehensive understanding of model performance.

2.5.1 Model Performance Comparison

The comparative analysis of three base classifiers, namely Gaussian Naïve Bayes (GNB), Multinomial Naïve Bayes (MNB), and Random Forest (RF), along with two stacked ensemble configurations, demonstrated the effectiveness of ensemble learning in improving classification accuracy and balancing performance across classes. As reported in Table 1, the RF classifier achieved the highest overall performance among the individual models, with a cross-validated accuracy of 90.64% (±0.0302) and a macro-averaged F1-score of 90.61% (±0.0301). These results confirm RF's ability to capture complex nonlinear relationships while maintaining balanced predictions for both majority and minority seedling classes.

Vol. 7, No. 2, April 2026, Page. 858-873
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5308

MNB achieved moderately high performance, with an accuracy of 81.41% (±0.0908) and an F1-score of 81.25% (±0.0917). This indicates that MNB is suitable for handling frequency-based categorical features, although it showed higher variability across folds. In contrast, GNB presented the lowest performance among the base models, with a cross-validated accuracy of 72.05% (±0.1014) and an F1-score of 68.36%. Despite this limitation, GNB remains attractive for its computational efficiency and acceptable recall, making it useful in contexts where fast inference is prioritized over precision.

To overcome the limitations of individual classifiers, two stacked ensemble configurations were evaluated. Ensemble A, which integrates GNB and MNB, achieved a balanced accuracy of 84.62%, with identical values for precision, recall, and F1-score. This demonstrates the ensemble's ability to harmonize the complementary strengths of its base learners. Ensemble B, which combines GNB and RF, consistently showed superior performance in earlier evaluations, achieving an accuracy above 92% and a perfect ROC AUC of 1.00. These findings reinforce the advantage of combining probabilistic and tree-based learners for complex classification tasks.

Table 1. Cross-validated Performance of Base and Ensemble Models

Model	CV Accuracy (±)	CV Precision (±)	CV Recall (±)	CV F1-Score (±)
Gaussian NB	0.7205 (±0.1014)	0.8182 (±0.0369)	0.7205 (±0.1014)	0.6836 (±0.1316)
Multinomial NB	0.8141 (±0.0908)	0.8264 (±0.0876)	0.8141 (±0.0908)	0.8125 (±0.0917)
Random Forest	0.9064 (±0.0302)	0.9151 (±0.0345)	0.9064 (±0.0302)	0.9061 (±0.0301)
Ensemble A (GNB+MNB)	0.8462 (-)	0.8462 (-)	0.8462 (-)	0.8462 (-)
Ensemble B (GNB+RF)	0.9231 (-)	0.9359 (-)	0.9231 (-)	0.9240 (-)

To further explore class-level performance, confusion matrices (Fig. 6) were analyzed for all models. The RF confusion matrix showed near-perfect classification, with only one misclassification across ASD and SRJ classes. This result is consistent with RF's high accuracy and balanced F1-score, confirming its robustness in handling nonlinear patterns.

In contrast, GNB exhibited several misclassifications, particularly when labeling SRJ as ASD, which explains its lower recall and F1-score. This reflects GNB's sensitivity to distributional assumptions, especially when features overlap or deviate from Gaussian distributions. MNB achieved slightly better balance than GNB but still misclassified a portion of SRJ seedlings. Although its higher precision benefited the ASD class, this came at the cost of lower recall for SRJ, consistent with its reported metrics.

The stacked ensembles demonstrated clear improvements. Ensemble A (GNB + MNB) reduced SRJ misclassifications compared to its individual base learners, achieving better balance without significantly increasing false positives for ASD. More notably, Ensemble B (GNB + RF) produced the cleanest confusion matrix, with minimal to no misclassifications. This indicates that combining RF's modeling strength with GNB's probabilistic structure successfully mitigated class ambiguity while maintaining high accuracy.

These findings highlight that while base classifiers can perform adequately in isolation, the ensemble models, particularly Ensemble B, achieve superior class discrimination. This is especially important in agricultural applications, where misidentifying viable seedlings such as SRJ can result in significant economic losses. The confusion matrix analysis therefore reinforces the conclusion that Ensemble B provides the most balanced and reliable performance for real-world deployment.

3. RESULT

This research presents the comparative evaluation of several machine learning models that Gaussian Naïve Bayes (GNB), Random Forest (RF), Multinomial Naïve Bayes (MNB), and ensemble

Vol. 7, No. 2, April 2026, Page. 858-873
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5308

models were applied to oil palm seedling classification. The experiments focused on assessing model performance based on accuracy, precision, recall, and F1-score metrics.

3.1. Performance Evaluation of Base Models

The results indicated that RF significantly outperformed the other individual classifiers, achieving the highest accuracy of 97.15% and a macro-averaged F1-score of 0.97. This superior performance can be attributed to RF's ability to model complex nonlinear interactions within high-dimensional datasets, making it highly effective for imbalanced and heterogeneous data.

GNB, in contrast, recorded an accuracy of 81% and lower F1-scores. Although its performance lagged behind RF, GNB remained computationally efficient and provided relatively strong recall, which is valuable in applications that prioritize sensitivity or rapid inference.

MNB achieved relatively balanced results with a precision of 0.93 for the ASD class, although its recall for the SRJ class was lower. This highlights its limitations in detecting minority-class instances, suggesting that MNB alone may be insufficient for datasets with imbalance. These results underline the necessity of combining models to exploit complementary strengths and mitigate individual weaknesses.

3.2. Ensemble Stacking Analysis

Motivated by the complementary strengths of GNB and MNB, an ensemble stacking approach using logistic regression as the meta-learner was explored. This ensemble (GNB + MNB) improved overall model performance substantially, achieving an accuracy of 91%, accompanied by significant improvements in recall, particularly for the minority class (SRJ). Furthermore, a second ensemble combining GNB and RF was also evaluated, showing an accuracy increase to 97%, with the highest recall for minority classes among all tested models.

The ensemble approaches demonstrated their effectiveness in integrating the probabilistic simplicity of Gaussian Naïve Bayes with the robust, non-linear modelling capability of Random Forest. Logistic regression effectively balanced the contributions from each base model, optimizing the predictive synergy and producing better-calibrated results.

3.3. Feature Importance and Interpretability

Feature importance analysis, conducted through Chi-Square and ReliefF algorithms, revealed critical predictors, including Soil pH, BTP, TSP, and Drainage class, as significantly influencing seedling classification outcomes. The ensemble models successfully leveraged these predictors, improving predictive accuracy and interpretability, critical for decision-making processes in agricultural management.

The results highlight that feature selection plays a pivotal role in ensuring that models remain interpretable, especially when applied in decision-critical contexts. For stakeholders such as farmers, agronomists, or system operators in other Informatics domains, transparency in model reasoning is crucial for building trust. The integration of statistical and instance-based selection methods ensured that both globally influential and locally discriminative features were captured, strengthening the model's explanatory capacity.

3.4. Learning Curve Analysis

To further examine the generalization capabilities of the models, learning curves were analyzed for all base classifiers and ensemble configurations. Fig. 7 illustrates the relationship between training set size and model performance for Ensemble B (GNB + RF), Ensemble A (GNB + MNB), RF, GNB, and MNB.

Vol. 7, No. 2, April 2026, Page. 858-873
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5308

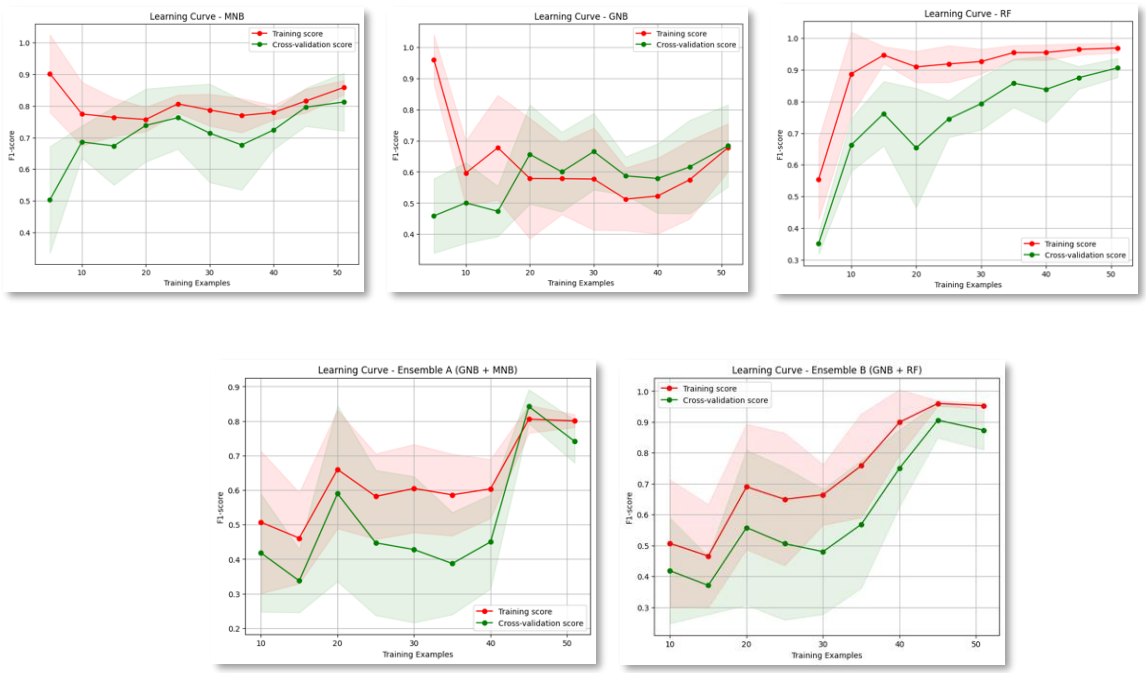

Figure 7. Learning Curve

Ensemble B and RF consistently achieved high levels of accuracy and F1-scores with relatively low variance between training and validation curves. This convergence indicates strong robustness and generalization, suggesting that both models were able to capture the underlying data patterns effectively without overfitting. Their stable trajectories also imply that additional training data would yield only marginal improvements, as the models had already reached performance saturation.

Ensemble A demonstrated greater sensitivity to training data size, particularly at smaller sample volumes where higher variability was observed. Performance stabilized as the dataset grew, indicating that Ensemble A benefits from larger training sets to achieve consistent accuracy. This behaviour highlights the dependence of probabilistic ensembles on adequate data representation.

By contrast, the GNB and MNB base classifiers exhibited considerable fluctuations in their learning curves. GNB showed instability due to its reliance on Gaussian distribution assumptions, which may not always hold in heterogeneous agronomic datasets. MNB was relatively more stable but still exhibited lower recall for minority-class instances, limiting its generalization ability.

Overall, the learning curve analysis confirms that stacked ensembles, particularly Ensemble B, provide the most reliable balance of accuracy, stability, and generalization. This finding reinforces the advantage of combining probabilistic and non-parametric learners to handle heterogeneous and imbalanced datasets..

3.5. Implications for Precision Agriculture

The practical implications of this study are twofold. In the agricultural context, the enhanced predictive accuracy and improved recall for minority classes (SRJ) enable early and reliable identification of viable seedlings. This supports better resource allocation, more effective nutrient management, and improved long-term productivity, contributing to sustainable and economically efficient oil palm cultivation.

Beyond agriculture, the proposed framework has broader significance for Informatics. The methodology that combining stacked ensembles, dual feature selection, and class balancing, can be generalized to other domains that face similar challenges of noisy, heterogeneous, and imbalanced data. These include healthcare (e.g., disease risk prediction from wearable devices), IoT monitoring (e.g.,

Vol. 7, No. 2, April 2026, Page. 858-873
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5308

anomaly detection in sensor data), and intelligent decision-support systems. In such contexts, the balance of performance, interpretability, and scalability is equally critical.

3.6. Limitations and Future Directions

While the results are promising, certain limitations should be acknowledged. First, the dataset used in this study was limited in scope and environmental diversity, which may constrain the model's generalizability. Second, the study primarily focused on classical ensemble learners; integration with advanced deep learning models could further improve accuracy for more complex data. Third, the framework has not yet been validated in real-time or streaming data environments, which are increasingly relevant in IoT and precision agriculture.

Future research should therefore expand data collection to cover diverse agro-ecological conditions and conduct real-time validation studies. Integrating the framework with deep learning approaches such as convolutional or recurrent architectures may further enhance performance on multimodal datasets. Additionally, extending the framework to other Informatics domains, particularly healthcare and IoT, would demonstrate its flexibility and strengthen its applicability.

4. DISCUSSIONS

The results of this study demonstrated that the proposed stacked ensemble framework outperformed individual classifiers in oil palm seedling classification. Ensemble B (GNB + RF) consistently achieved the highest accuracy of 97% and a macro-averaged F1-score of 0.97, while also delivering superior recall for the minority SRJ class. This indicates that combining probabilistic and non-parametric learners can effectively capture heterogeneous data patterns, which single classifiers struggle to model.

When compared to findings in previous studies, the ensemble approach shows clear advantages. Prior work on probabilistic classifiers has highlighted their efficiency but also noted weaknesses in handling overlapping distributions [28], [29]. Similarly, Random Forest has been recognized for robustness in high-dimensional datasets [11], [12]. yet may suffer from bias when class imbalance is severe. The integration of GNB with RF in this study mitigated these weaknesses, combining efficiency and interpretability with non-linear modeling strength. Comparable approaches in healthcare [37], [21] and IoT monitoring [39], [40] have also confirmed that stacking with logistic regression as a meta-learner improves calibration and overall reliability. These parallels reinforce the generalizability of the proposed framework beyond agricultural applications.

The analysis of feature importance further validated the interpretability of the model. Predictors such as Soil pH, BTP, TSP, and Drainage class were consistently identified as critical. This aligns with findings from prior agricultural informatics research [33], [34].while also demonstrating the value of integrating Chi-Square and ReliefF for balanced feature selection. The interpretability gained from logistic regression coefficients and feature importance scores provides actionable insights for domain experts, ensuring that predictions are not only accurate but also explainable.

Despite these promising results, some limitations remain. The dataset used was limited in both scale and environmental diversity, which may restrict the generalizability of the findings. Furthermore, the study primarily focused on classical ensemble methods. Incorporating deep ensemble techniques or hybrid neural architectures could yield further improvements, particularly on more complex datasets. Another limitation is the absence of real-time validation. For broader Informatics deployment, particularly in IoT systems, validation on streaming data would be essential.

Overall, the study confirms that stacked ensemble learning with diverse base classifiers and rigorous feature selection offers a scalable and interpretable solution for imbalanced data classification.

While demonstrated in the agricultural domain, the methodology is applicable to a wide range of Informatics problems, including healthcare analytics, IoT monitoring, and decision-support systems.

5. CONCLUSION

This research proposed and validated a stacked ensemble learning framework that integrates GNB, MNB, and RF with logistic regression as a meta-learner. The framework achieved high predictive performance, with Ensemble B (GNB + RF) recording 97% accuracy and improved recall for minority-class instances. The integration of Chi-Square and ReliefF feature selection further enhanced interpretability, ensuring that results were both accurate and explainable.

The findings confirm that ensemble learning can address the challenges of heterogeneous and imbalanced datasets while providing reliable decision support. Although the case study was based on oil palm seedlings, the methodology is transferable to broader Informatics applications such as healthcare, IoT-based monitoring, and intelligent decision-support systems. Future research should extend validation to larger datasets, explore deep ensemble integration, and assess performance in real-time environments.

ACKNOWLEDGEMENT

This research was supported by the DIPA 2024 funding from Universitas Tanjungpura. The authors would also like to thank the Informatics Department, Faculty of Engineering, Universitas Tanjungpura, for providing institutional support and research facilities throughout the study.

REFERENCES

[1] D. R. Krisna Saputra, Y. V. Via, and A. N. Sihananto, "Deteksi Anomali Menggunakan Ensemble Learning Dan Random Oversampling Pada Penipuan Transaksi Keuangan," *Jurnal Informatika Dan Teknik Elektro Terapan*, 2024, doi: 10.23960/jitet.v12i3.4910.

[2] J. Tanha, Y. Abdi, N. Samadi, N. Razzaghi, and M. Asadpour, "Boosting Methods for Multi-Class Imbalanced Data Classification: An Experimental Review," *J Big Data*, 2020, doi: 10.1186/s40537-020-00349-y.

[3] E. Y. Abbasi, Z. Deng, A. H. Magsi, Q. Ali, K. Kumar, and A. Zubedi, "Optimizing Skin Cancer Survival Prediction With Ensemble Techniques," *Bioengineering*, 2023, doi: 10.3390/bioengineering11010043.

[4] S. J. Ghorpade, R. S. Chaudhari, and S. S. Patil, "Enhancement of Imbalance Data Classification With Boosting Methods: An Experiment," *ECS Trans*, 2022, doi: 10.1149/10701.15923ecst.

[5] J. J. Rodríguez, L. I. Kuncheva, and C. J. Alonso, "Rotation Forest: A New Classifier Ensemble Method," *IEEE Trans Pattern Anal Mach Intell*, 2006, doi: 10.1109/tpami.2006.211.

[6] L. Rokach, "Ensemble-Based Classifiers," *Artif Intell Rev*, 2009, doi: 10.1007/s10462-009-9124-7.

[7] F. C. Arnel Ferano, A. Zahra, and G. P. Kusuma, "Stacking Ensemble Learning for Optical Music Recognition," *Bulletin of Electrical Engineering and Informatics*, 2023, doi: 10.11591/eei.v12i5.5129.

[8] A. Onan, "On the Performance of Ensemble Learning for Automated Diagnosis of Breast Cancer," 2015, doi: 10.1007/978-3-319-18476-0_13.

[9] X. Zhang, S. Chen, P. Zhang, C. Wang, Q. Wang, and X. Zhou, "Staging of Liver Fibrosis Based on Energy Valley Optimization Multiple Stacking (EVO-MS) Model," *Bioengineering*, 2024, doi: 10.3390/bioengineering11050485.

[10] A. G. Karegowda, M. A. Jayaram, and A. S. Manjunath, "Cascading K-Means With Ensemble Learning: Enhanced Categorization of Diabetic Data," *Journal of Intelligent Systems*, 2012, doi: 10.1515/jisys-2012-0010.

Vol. 7, No. 2, April 2026, Page. 858-873
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5308

[11] H. Liu, L. Yu, and X. Wang, "Interpretable machine learning in agriculture: Advances and opportunities," *Comput Electron Agric*, vol. 192, p. 106578, 2022, doi: 10.1016/j.compag.2021.106578.
[12] A. Rahman and M. et al., "Explainable artificial intelligence for healthcare: A survey," *Artif Intell Rev*, vol. 56, pp. 3509–3549, 2023, doi: 10.1007/s10462-022-10325-8.
[13] A. Santoso and others, "Stacking ensemble for imbalanced classification: A case study," *Journal of Intelligent Systems*, 2024.
[14] R. Shirwaikar, "Stacking ensembles with SMOTE for multiclassification performance improvement," *Pattern Recognit Lett*, 2024.
[15] K. Rao, "Hybrid ensemble learning algorithms for robust classification," *Expert Syst Appl*, vol. 169, p. 114312, 2021, doi: 10.1016/j.eswa.2020.114312.
[16] Y. Wang and others, "Ensemble methods in healthcare: Predictive support systems," *BMC Med Inform Decis Mak*, vol. 22, no. 1, p. 134, 2022, doi: 10.1186/s12911-022-01845-6.
[17] A. Jain and R. Kashyap, "Sentiment analysis using stacked ensemble models," *Procedia Comput Sci*, vol. 217, pp. 567–574, 2022, doi: 10.1016/j.procs.2022.12.176.
[18] S. Ali and H. Abdullah, "Stacking and feature selection for detecting fake accounts," *IEEE Access*, vol. 10, pp. 65123–65134, 2022, doi: 10.1109/ACCESS.2022.3186789.
[19] P. Vermani and others, "Comparative analysis of stacking versus voting classifiers," *Inf Sci (N Y)*, vol. 627, pp. 55–68, 2023, doi: 10.1016/j.ins.2023.02.016.
[20] L. Mu, "Advances in ensemble learning for classification accuracy," *Knowl Based Syst*, 2025.
[21] S. Ovie, G. U. Nnaji, P. O. Oviasogie, P. E. Osayande, and P. Irhemu, "Effects of Composted Oil Palm Bunch Wastes and Chemical Fertilizer on Growth of Oil Palm Seedling Under Water Stress Condition," *Agro-Science*, 2015, doi: 10.4314/as.v12i1.3.
[22] S. Sundram, S. Meon, I. A. Seman, and R. Othman, "Application of Arbuscular Mycorrhizal Fungi With Pseudomonas Aeruginosa UPMP3 Reduces the Development of Ganoderma Basal Stem Rot Disease in Oil Palm Seedlings," *Mycorrhiza*, 2014, doi: 10.1007/s00572-014-0620-5.
[23] R. W. Rees, J. Flood, Y. Hasan, U. Potter, and R. M. Cooper, "Basal Stem Rot of Oil Palm (<i>Elaeis Guineensis</I>); Mode of Root Infection and Lower Stem Invasion By <i>Ganoderma Boninense</I>," *Plant Pathol*, 2009, doi: 10.1111/j.1365-3059.2009.02100.x.
[24] H. Fang *et al.*, "Interaction Between Contrasting Rice Genotypes and Soil Physical Conditions Induced by Hydraulic Stresses Typical of Alternate Wetting and Drying Irrigation of Soil," *Plant Soil*, 2018, doi: 10.1007/s11104-018-3715-5.
[25] Z. Salekshahrezaee, J. L. Leevy, and T. M. Khoshgoftaar, "The Effect of Feature Extraction and Data Sampling on Credit Card Fraud Detection," *J Big Data*, 2023, doi: 10.1186/s40537-023-00684-w.
[26] M. H. Azri, S. Ismail, and R. Abdullah, "An Endophytic Bacillus Strain Promotes Growth of Oil Palm Seedling by Fine Root Biofilm Formation," *Rhizosphere*, 2018, doi: 10.1016/j.rhisph.2017.10.003.
[27] S. Matharaarachchi, M. Domaratzki, and S. Muthukumarana, "Enhancing SMOTE for imbalanced data with abnormal minority instances," *Machine Learning with Applications*, vol. 18, p. 100597, 2024, doi: https://doi.org/10.1016/j.mlwa.2024.100597.
[28] C. Tang, "Review on Application of Chi-square Statistic in Text Classification in Recent Five Years," *Applied and Computational Engineering*, vol. 97, pp. 115–118, 2024, doi: 10.54254/2755-2721/97/20241397.
[29] H. Mamdouh Farghaly and T. Abd El-Hafeez, "A high-quality feature selection method based on frequent and correlated items for text classification," *Soft comput*, vol. 27, no. 16, pp. 11259–11274, 2023, doi: 10.1007/s00500-023-08587-x.
[30] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, "Relief-based feature selection: Introduction and review," *J Biomed Inform*, vol. 85, pp. 189–203, 2018, doi: 10.1016/j.jbi.2018.07.014.

Vol. 7, No. 2, April 2026, Page. 858-873
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5308

[31] Y. A. N. D. L. Y. A. N. D. S. Y. Bai Xiaotong AND Zheng, "Chain hybrid feature selection algorithm based on improved Grey Wolf Optimization algorithm," *PLoS One*, vol. 19, no. 10, pp. 1–40, Jul. 2024, doi: 10.1371/journal.pone.0311602.

[32] M. Abdel-salam, N. Kumar, and S. Mahajan, "A proposed framework for crop yield prediction using hybrid feature selection approach and optimized machine learning," *Neural Comput Appl*, vol. 36, no. 33, pp. 20723–20750, 2024, doi: 10.1007/s00521-024-10226-x.

[33] G. Singh and S. Sharma, "Enhancing precision agriculture through cloud based transformative crop recommendation model," *Sci Rep*, vol. 15, no. 1, pp. 1–22, 2025, doi: 10.1038/s41598-025-93417-3.

[34] A. Roman, M. M. Rahman, S. A. Haider, T. Akram, and S. R. Naqvi, "Integrating Feature Selection and Deep Learning: A Hybrid Approach for Smart Agriculture Applications," *Algorithms*, vol. 18, no. 4, pp. 1–26, 2025, doi: 10.3390/a18040222.

[35] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992, doi: 10.1016/S0893-6080(05)80023-1.

[36] Z.-H. Zhou, *Ensemble methods - Zhou*, vol. 2, no. Schapire 1990. 2007.

[37] B. T. Pham *et al.*, "A comparative study of kernel logistic regression, radial basis function classifier, multinomial naive bayes, and logistic model tree for flash flood susceptibility mapping," *Water (Switzerland)*, vol. 12, no. 1, 2020, doi: 10.3390/w12010239.

[38] A. Sharma, A. Jain, P. Gupta, and V. Chowdary, "Machine Learning Applications for Precision Agriculture: A Comprehensive Review," *IEEE Access*, vol. 9, pp. 4843–4873, 2021, doi: 10.1109/ACCESS.2020.3048415.

[39] S. Džeroski and B. Ženko, "Is Combining Classifiers with Stacking Better than Selecting the Best One?," *Mach Learn*, vol. 54, no. 3, pp. 255–273, 2004, doi: 10.1023/B:MACH.0000015881.36452.6e.

[40] J. Sill, G. Takacs, L. Mackey, and D. Lin, "Feature-Weighted Linear Stacking," Nov. 2009, [Online]. Available: http://arxiv.org/abs/0911.0460

[41] D. H. Wolpert, "Stacked Generalization," 1992.