



Uncertainty-Guided Noise-Aware Oversampling with Deep CNN for Imbalanced Oil Palm Disease Classification

Hartono¹, Erianto Ongko², Retna Astuti Kuswardani³, Suswati⁴

¹Department of Magister Program in Informatics, Postgraduate School, Universitas Medan Area, Medan, Indonesia

²Department of Computer Science, Institut Modern Arsitektur dan Teknologi, Medan, Indonesia

^{3,4}Department of Doctoral Program in Agricultural Sciences, Postgraduate School, Universitas Medan Area, Medan, Indonesia

*hartono@staff.uma.ac.id

Abstract

Oil palm disease classification is essential for precision agriculture, but its performance is often limited by class imbalance, noisy samples, and overlapping feature distributions. Conventional Convolutional Neural Network (CNN)-based models provide strong feature extraction capabilities, yet they tend to be biased toward majority classes and are sensitive to unreliable training data. Oversampling methods such as the Synthetic Minority Over-sampling Technique (SMOTE) improve class balance, but may amplify noise because synthetic samples are generated without considering sample reliability. This study proposes an uncertainty-guided noise-aware oversampling framework integrated with a deep CNN for imbalanced oil palm disease classification. Aleatoric uncertainty is used to identify and remove noisy samples, while epistemic uncertainty guides synthetic data generation toward informative regions near decision boundaries. A sensitivity analysis identifies the most effective configuration, consisting of 20 forward passes, a dropout rate of 0.5, an aleatoric uncertainty threshold of 0.4, and prediction variance as the epistemic uncertainty measure. Experimental results show that the proposed framework achieves the highest validation accuracy of 92.88% and consistently outperforms CAE, LCNN, CNN, Deep SVM, and SCNN across Accuracy, Precision, Recall, Specificity, F1-Score, Balanced Accuracy, and Cohen's Kappa. The largest improvements are observed in the minority classes, particularly Ganoderma Initial Infection and Ganoderma Infected. Wilcoxon signed-rank testing confirms that all improvements are statistically significant ($p < 0.05$). Computational analysis indicates only a moderate and practically insignificant increase in training time, memory usage, and computational complexity. These findings demonstrate that the proposed framework provides an effective and robust solution for imbalanced agricultural image classification.

Keywords: class imbalance; deep CNN; noisy samples; oversampling; uncertainty

How to Cite: H. Hartono, E. Ongko, R. A. Kuswardani, and S. Suswati, "Uncertainty-Guided Noise-Aware Oversampling with Deep CNN for Imbalanced Oil Palm Disease Classification", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 10, no. 5, pp. 1313–1327, Oct. 2026, doi: [10.29207/resti.v10i5.7683](https://doi.org/10.29207/resti.v10i5.7683).

Permalink/DOI: <https://doi.org/10.29207/resti.v10i5.7683>

Received: April 3, 2026

Accepted: June 1, 2026

Available Online: October 4, 2026

This is an open-access article under the [CC BY 4.0 License](https://creativecommons.org/licenses/by/4.0/)

Published by [Ikatan Ahli Informatika Indonesia](https://www.iaii.or.id/)

1. Introduction

The oil palm industry plays a crucial role in global agriculture[1], yet it faces significant challenges related to plant disease management[2]. Diseases such as pest infestations, fungal infections, and nutrient-related disorders can severely reduce productivity and crop quality[3]. Early and accurate disease detection is therefore essential to support precision agriculture[4] and sustainable plantation management[5]. In recent years, image-based classification using deep learning has emerged as a promising solution[6], particularly through Convolutional Neural Networks (CNNs),

which are capable of extracting complex spatial features from leaf images[7], [8].

However, a fundamental problem in oil palm disease classification is the presence of class imbalance[9]. In real-world conditions, certain diseases occur less frequently than others[10], resulting in uneven data distribution across classes[11], [12]. This imbalance leads to biased learning[13], where deep learning models tend to favor majority classes while underperforming on minority classes[14], which are often the most critical for early detection[15]. Moreover, oil palm leaf images frequently exhibit high intra-class variability and inter-class similarity, making

the classification problem even more challenging due to overlapping feature spaces and noisy samples[16].

To address class imbalance, various approaches have been proposed, including cost-sensitive learning, ensemble methods, and data-level resampling techniques[17]. Among these, oversampling methods are widely adopted because they preserve the original dataset[18] while improving class balance[19]. SMOTE and its variants are among the most popular techniques[20], as they generate synthetic samples to enhance minority class representation[21]. When integrated with deep learning, particularly CNN-based frameworks, SMOTE-based approaches have demonstrated the ability to improve classification performance by providing more balanced training data[22]. Despite these advantages, SMOTE-based deep learning still suffers from critical limitations[23]. The synthetic samples are generated based on local interpolation without considering the reliability or informativeness of the original data points[24]. As a result, samples located in noisy or overlapping regions may be used to generate new data[25], leading to noise amplification and degraded model generalization[26]. This issue becomes particularly problematic in complex agricultural datasets, where environmental variations, lighting conditions, and disease symptoms introduce significant uncertainty[27]. Consequently, although SMOTE improves class balance, it does not guarantee the quality of the generated data[28].

Alternative approaches have attempted to address these limitations. Cost-sensitive methods assign higher penalties to minority classes but often fail to handle sample-level noise[29]. Ensemble methods improve robustness but increase computational complexity and may still inherit imbalance issues[30]. More advanced oversampling techniques, including density-based or boundary-focused methods[31], attempt to refine sample generation but still rely primarily on geometric distribution rather than the underlying uncertainty of the data[32]. Therefore, a key limitation across existing methods is the lack of consideration for sample reliability and informativeness during the balancing process[33]. To overcome this limitation, uncertainty-aware approaches such as Uncertainty-Guided Oversampling (UGO) have been introduced[34]. The key strength of UGO lies in its ability to distinguish between different types of uncertainty. It identifies samples with high aleatoric uncertainty, which are typically associated with noise or ambiguity, and excludes them from the data generation process. At the same time, it prioritizes samples with high epistemic uncertainty, which represent informative regions that can improve model learning. By incorporating uncertainty into the oversampling process, UGO provides a more selective and reliable mechanism for synthetic data generation, reducing the risk of noise amplification and improving the overall data quality.

Nevertheless, UGO primarily focuses on improving the oversampling mechanism and does not inherently

provide a strong classification framework. On the other hand, SMOTE-based deep learning frameworks, particularly those using CNN architectures, excel in learning complex feature representations and achieving high classification accuracy but lack mechanisms to control the quality of synthetic data[35]. This creates a complementary relationship between the two approaches: SMOTE-based deep learning offers strong representation learning and classification capability, while UGO provides a robust strategy for generating high-quality, informative synthetic samples.

Based on this observation, combining UGO with SMOTE-based deep learning becomes a promising solution. UGO can be used to guide the oversampling process by filtering noisy samples and selecting informative regions, thereby improving the quality of the balanced dataset. This enhanced dataset can then be utilized by a deep CNN model to perform more accurate and robust classification. By integrating uncertainty-aware sampling with deep feature learning, the proposed approach addresses both data-level and model-level challenges simultaneously.

Therefore, this study proposes a novel framework entitled “Uncertainty-Guided Noise-Aware Oversampling with Deep CNN for Imbalanced Oil Palm Disease Classification.” The main contribution of this work lies in the integration of uncertainty-guided oversampling with deep learning-based classification. Specifically, the proposed method (1) introduces a noise-aware balancing mechanism that reduces the influence of unreliable samples, (2) improves the informativeness of synthetic data through uncertainty-guided selection, (3) enhances minority class representation without amplifying noise, and (4) leverages deep CNN to achieve robust and accurate oil palm disease classification. Through this integration, the proposed framework is expected to outperform conventional oversampling and deep learning methods, particularly in complex and imbalanced agricultural datasets.

2. Methods

2.1 SMOTE-based Deep Learning

The classification of oil palm leaf diseases presents significant challenges, particularly due to class imbalance in real-world datasets, where certain disease categories are underrepresented compared to others. This imbalance can bias learning algorithms, leading to suboptimal model performance, especially in detecting minority classes that are often critical for early diagnosis. To address this issue, an integrated framework combining data preprocessing, Synthetic Minority Over-sampling Technique (SMOTE), and deep learning is proposed.

The workflow in Figure 1, begins with the preparation of raw palm leaf images, followed by essential preprocessing steps such as data cleaning, resizing, normalization, and dataset splitting to ensure data

quality and consistency. To overcome the imbalance problem, SMOTE is employed to generate synthetic samples for minority classes based on k-nearest neighbors, resulting in a more balanced dataset distribution[35]. This balanced dataset is then utilized to train a deep Convolutional Neural Network (CNN), which performs hierarchical feature extraction and classification[35].

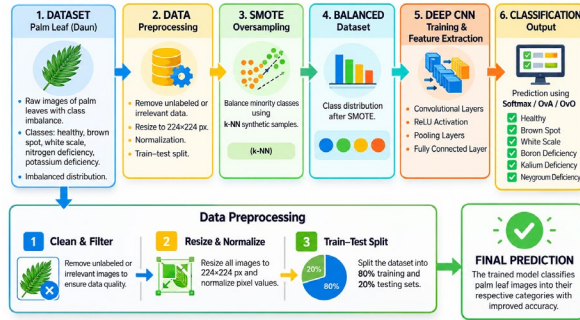


Figure 1. Workflow of SMOTE-Based Deep Learning

The workflow begins with the oil palm leaf image dataset, followed by preprocessing, synthetic oversampling using SMOTE, balanced dataset construction, Deep CNN training, and final classification [35].

The initial dataset is mathematically defined in Equation 1[36].

$$D^{(0)} = \{(x_i, y_i)\}_{i=1}^N \quad (1)$$

$D^{(0)}$ denotes the original dataset, $x_i \in \mathbb{R}^d$ is the i -th input image, $y_i \in \{1, 2, \dots, K\}$ is the corresponding class label, N is the total number of images, d is the feature dimension, and K is the number of classes.

In data preprocessing, images are cleaned and prepared before balancing and model training. The preprocessed dataset is represented in Equation 2[36].

$$D^{(1)} = P(D^{(0)}) \quad (2)$$

$P(\cdot)$ Denotes the preprocessing operations. These operations include removing irrelevant or unusual samples, resizing all images into a uniform resolution, normalizing pixel values, and splitting the dataset into training and testing subsets. The goal of this stage is to ensure that all input images are represented in a consistent format for the next processing steps. After preprocessing, class imbalance is addressed using the Synthetic Minority Sampling Techniques (SMOTE). The main idea of SMOTE is to generate synthetic minority samples by interpolating between a minority sample and one of its nearest neighbors. For a selected sample x_i and one of its neighbors x_j , a synthetic sample is generated according to Equation 3[20].

$$x_{new} = x_i + \lambda(x_j - x_i) \quad (3)$$

x_{new} is the generated synthetic samples, x_i is an existing minority-class sample, x_j is one of the k nearest neighbors of x_i from the same class, and $\lambda \in [0, 1]$ is a random interpolation coefficient.

Using SMOTE, the minority classes are enlarged until their sample distribution become more balanced. The resulting oversampled dataset is formulated in Equation 4[21].

$$D^{(2)} = S(D^{(1)}) \quad (4)$$

$S(\cdot)$ Denotes the SMOTE oversampling function.

The purpose of this stage is to reduce the dominance of majority classes and improve the representation of minority classes before classification. Unlike random oversampling, SMOTE does not simply duplicate existing samples, but instead produces new synthetic samples by interpolation in the feature space.

Once the balanced dataset has been constructed, the next stage is deep CNN training. The CNN classifier is defined in Equation 5[7].

$$f_\theta: \mathbb{R}^d \rightarrow \{1, 2, \dots, K\} \quad (5)$$

θ denotes the trainable parameters of the CNN. The classifier estimates the posterior probability for each class using Equation 6 [35].

$$P_\theta(y = k|x_i) \quad (6)$$

The final predicted class is determined using Equation 7 [35].

$$\hat{y}_i = \arg \max_{k \in \{1, 2, \dots, K\}} P_\theta(y = k|x_i) \quad (7)$$

$\hat{y}$ is the predicted label for sample x_i , $P_\theta(y = k|x_i)$ is the predicted probability that x_i belongs to class k .

The CNN acts as the feature extractor and classifier. In general, convolutional layers learn local spatial patterns from the input images, pooling layers reduce spatial redundancy, and fully connected layers perform the final decision making. In the baseline SMOTE-based deep learning framework, the CNN is trained on the balanced dataset rather than the original imbalanced dataset, so that the minority classes can contribute more effectively to the learning process.

The multiclass cross-entropy loss used for classification is formulated in Equation 8[35].

$$\mathcal{L}(\theta) = -\sum_{i=1}^M \sum_{k=1}^K (y_i = k) \log P_\theta(y = k|z_i) \quad (8)$$

$\mathcal{L}(\theta)$ is the training loss, M is the number of samples in the balanced dataset, z_i is i -th training sample in the balanced dataset, and $(y_i = k)$ is the indicator function, which equals 1 if the true label of z_i is k and 0 otherwise.

After training, the model produces the final classification output for each palm leaf image.

2.2 Uncertainty-Guided Oversampling

UGO (Uncertainty-Guided Oversampling) is introduced to address the key limitations of conventional oversampling methods, particularly noise amplification and the inability to distinguish between informative and unreliable samples. Traditional approaches such as SMOTE generate synthetic data primarily based on spatial proximity, without

considering data quality, which may lead to the propagation of existing noise and increased class overlapping. To overcome these issues, UGO integrates uncertainty estimation into the synthetic data generation process. Specifically, it utilizes aleatoric uncertainty to detect and avoid noisy samples, while epistemic uncertainty is employed to identify informative samples, particularly those located near the decision boundary. By separating these roles, UGO ensures that synthetic data are generated from reliable and meaningful regions of the feature space. In addition, UGO adopts an iterative strategy in which synthetic samples are generated progressively based on class-wise epistemic uncertainty, rather than relying on predefined oversampling ratios[34]. This allows the framework to adaptively balance the dataset by reducing uncertainty differences across classes, instead of merely equalizing the number of samples. The overall workflow of the UGO framework, including uncertainty estimation, noise filtering, and uncertainty-guided data generation, can be seen in the Figure 2[34].

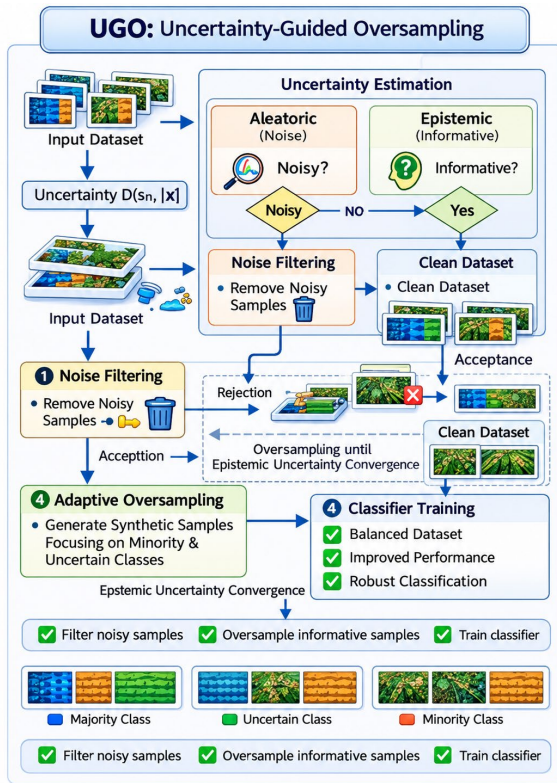


Figure 2. Workflow of UGO

The dataset used in the UGO formulation is defined in Equation 9 [36].

$$D^{(0)} = \{(x_i, y_i)\}_{i=1}^N \quad (9)$$

$D^{(0)}$ denotes the original dataset, $x_i \in \mathbb{R}^d$ is the i -th input sample, $y_i \in \{1, 2, \dots, K\}$ is the class label, N is the total number of samples, and K is the number of classes.

UGO employs a classifier with trainable parameters to estimate predictive probabilities, as formulated in Equation 10 [34].

$$P_\theta(y|x_i) \quad (10)$$

$P_\theta(y|x_i)$ denotes the predicted probability distribution over classes.

The two types of uncertainty computed for each sample are formulated in Equations 11 and 12 [34].

$$U^{(\alpha)}(x_i) = E_\theta[Var(y|x_i, \theta)] \quad (11)$$

$$U^{(e)}(x_i) = Var_\theta[P_\theta(y|x_i)] \quad (12)$$

$U^{(\alpha)}(x_i)$ denotes aleatoric uncertainty, representing noise in the data, $U^{(e)}(x_i)$ denotes epistemic uncertainty, representing model uncertainty, θ denotes the model parameters, $Var(\cdot)$ denotes variance, and $E(\cdot)$ denotes expectation.

Based on uncertainty estimation, noisy samples are identified and excluded from data generation according to Equation 13[34].

$$D^{clean} = \{(x_i, y_i) \in D^0 | U^{(\alpha)}(x_i) < \tau\} \quad (13)$$

τ is a predefined noise threshold.

UGO performs synthetic data generation iteratively. At each iteration t , the class-wise epistemic uncertainty is computed using Equation 14[34].

$$\bar{U}_k^{(e)} = \frac{1}{|D_k^{(t)}|} \sum_{(x_i, y_i) \in D_k^{(t)}} U^{(e)}(x_i) \quad (14)$$

$\bar{U}_k^{(e)}$ is the average epistemic uncertainty of class k , $D_k^{(t)}$ denotes samples of class k at iteration t , $|D_k^{(t)}|$ is the number of samples in class k .

Instead of balancing class sizes, UGO balances epistemic uncertainty across classes. At each iteration, synthetic samples are generated for the class with the lowest epistemic uncertainty to reduce imbalance in uncertainty space.

Synthetic samples are generated using the interpolation mechanism formulated in Equation 15 [34].

$$x_{new} = x_i + \lambda(x_j - x_i) \quad (15)$$

$x_i, x_j \in D^{clean}$, $\lambda \in [0, 1]$ is a random interpolation coefficient, and x_{new} is the generated synthetic sample.

To prevent noise propagation, UGO applies a two-stage noise control mechanism: Pre-generation filtering and Post-Generation validation. In Pre-generation filtering, noisy samples are excluded using $U^{(\alpha)}(x_i)$ and in Post-generation validation, synthetic samples are re-evaluated using aleatoric uncertainty.

When a newly generated synthetic sample is identified as noisy based on the predefined uncertainty criterion, an additional noise-handling mechanism is applied to prevent the sample from degrading the quality of the augmented dataset. Instead of allowing such an unreliable sample to directly participate in the subsequent learning process, the framework performs a corrective procedure that considers the reliability of neighboring samples within the same class. This

A large value of $U^{(a)}(\tilde{x}_i)$ indicates that the sample is likely to be intrinsically noisy. In the context of image classification, such samples are undesirable for oversampling because generating synthetic data around noisy points may amplify errors rather than improve class representation.

Epistemic uncertainty measures model uncertainty and reflects the lack of knowledge of the classifier in certain regions of the feature space. It is formulated in Equation 26 [34].

$$U^{(e)}(\tilde{x}_i) = \text{Var}_{\theta}[P_{\theta}(y|\tilde{x}_i)] \quad (26)$$

A large value of $U^{(e)}(\tilde{x}_i)$ indicates that the classifier is uncertain about the decision around $\tilde{x}_i$. Such samples are regarded as informative because they often lie near class boundaries or sparsely represented regions. Therefore, unlike aleatoric uncertainty, epistemic uncertainty is not used for sample removal, but rather for sample selection during oversampling. This process refines the dataset before oversampling. Specifically, aleatoric uncertainty is used to filter out noisy samples, and the resulting cleaned dataset is defined in Equation 27[42].

$$D_{\text{clean}} = \{(\tilde{x}_i, y_i) \in \tilde{D} \mid U^{(a)}(\tilde{x}_i) < \tau\} \quad (27)$$

Only samples with $U^{(a)}(\tilde{x}_i)$ lower than τ are retained. Samples whose aleatoric uncertainty exceeds the threshold are discarded because they are likely to introduce harmful synthetic patterns if used as seeds for oversampling.

After noisy samples have been removed, the method performs adaptive oversampling on the refined dataset D_{clean} . This stage connects the SMOTE interpolation principle with the informative-sample selection strategy of UGO.

A synthetic sample is generated by interpolation between two selected samples according to Equation 28 [43].

$$x_{\text{new}} = \tilde{x}_i + \lambda(\tilde{x}_j - \tilde{x}_i) \quad (28)$$

This formulation follows the synthetic interpolation mechanism commonly used in SMOTE based methods. However, the important difference is that the proposed method does not choose interpolation seeds arbitrarily.

To determine where synthetic samples should be generated, class-wise epistemic uncertainty is calculated using Equation 29[34].

$$\bar{U}_k^{(e)} = \frac{1}{|D_k|} \sum_{(\tilde{x}_i, y_i) \in D_k} U^{(e)}(\tilde{x}_i) \quad (29)$$

Classes with higher epistemic uncertainty are interpreted as more poorly learned or less confidently represented. Therefore, oversampling is focused on these classes and on those samples within them that have high informative value. The iterative oversampling process continues until the class-wise epistemic uncertainty becomes sufficiently balanced, as specified in Equation 30 [34]

$$\max_k \bar{U}_k^{(e)} - \min_k \bar{U}_k^{(e)} < \varepsilon \quad (30)$$

Because synthetic interpolation may still produce undesirable samples, the generated set is further refined. The validated synthetic set is defined in Equation 31 [34].

$$D_{\text{syn}} = \{(x_{\text{new}}, y_{\text{new}}) \mid U^{(a)}(x_{\text{new}}) < \tau\} \quad (31)$$

Synthetic samples with excessive aleatoric uncertainty are discarded or relocated. This validation step ensures that oversampling does not reintroduce the same noise that was previously removed.

After validation, the final balanced training set is defined in Equation 32[34].

$$D_{\text{final}} = D_{\text{clean}} \cup D_{\text{syn}} \quad (32)$$

The final dataset is then used to train a deep CNN classifier. The training objective based on multiclass cross-entropy loss is formulated in Equation (33) [34]

$$\mathcal{L}(\theta) = -\sum_{i=1}^M \sum_{k=1}^K (y_i = k) \log P_{\theta}(y = k | z_i) \quad (33)$$

The role of the CNN in this framework follows the deep-learning stage in SMOTE-based image classification studies: once a more balanced and reliable dataset is obtained, the CNN can learn more robust spatial representations and produce better classification performance across all classes.

2.4 Dataset

This study utilizes 2 (two) oil palm leaf image datasets obtained from Roboflow, consisting of various plant health conditions, including healthy leaves, disease infections, and nutrient deficiencies[44]. The first dataset comprises six classes: healthy leaf, brown spot, white scale, boron deficiency, potassium deficiency, and magnesium deficiency, each exhibiting distinct visual characteristics such as discoloration, lesions, and chlorosis patterns.

The second dataset contains images corresponding to Ganoderma disease[45]. The detailed class distribution and descriptions are presented in Table 1[44] and Table 2 [45].

2.6. Evaluation Metric

The proposed model is assessed using a set of standard classification measures to capture its predictive performance from multiple perspectives. Accuracy indicates the proportion of predictions that are correctly classified across all samples. Precision reflects the reliability of positive predictions, whereas Recall represents the model's capability to detect actual positive instances. Specificity evaluates how effectively negative instances are correctly recognized. The F1-score combines Precision and Recall into a single measure, making it particularly informative when dealing with imbalanced class distributions. Balanced Accuracy accounts for performance across positive and negative classes by considering both true-positive and true-negative recognition rates.

In addition, Cohen's Kappa quantifies the level of agreement between the predicted and ground-truth labels after accounting for agreement that may occur by chance[46]. The Accuracy, Precision, Recall, Specificity, F-1 Score, Balanced Accuracy, and Cohen's Kappa are shown in Equations 34-40[46].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (34)$$

$$Precision = \frac{TP}{TP+FP} \quad (35)$$

$$Recall = \frac{TP}{TP+FN} \quad (36)$$

$$Specificity = \frac{TN}{TN+FP} \quad (37)$$

$$F-1 \text{ Score} = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (38)$$

$$Balanced \text{ Accuracy} = \frac{TPR+TNR}{2} \quad (39)$$

$$Cohen's \text{ Kappa} = \frac{Po-Pe}{1-Pe} \quad (40)$$

Table 1. Description of the First Oil Palm Leaf Dataset

Class	Description	Total Images	Sample Image
Healthy Leaf	Healthy oil palm leaves are characterized by a uniform green color, intact leaflet structure, and absence of visible symptoms such as spots, discoloration, or pest infestation, indicating optimal physiological condition and balanced nutrient availability.	1199	
Brown Spot	Brown Spot is a fungal disease typically presenting as small brown lesions on the leaf surface, which may expand and coalesce, leading to reduced photosynthetic activity and overall plant vigor.	475	
White Scale	White Scale is caused by scale insects that attach to the leaf surface, forming white, waxy colonies that suck plant sap, leading to yellowing, weakening, and potential defoliation.	954	
Boron Deficiency	Boron deficiency in oil palm is indicated by distorted and brittle young leaves, irregular leaflet development, and reduced growth due to impaired cell wall formation and nutrient transport.	1321	
Kalium Deficiency	Potassium (Kalium) deficiency is characterized by yellowing or necrosis at the leaf margins, progressing inward, often forming orange or brown spotting patterns that reduce photosynthetic efficiency.	4076	
Magnesium Deficiency	Magnesium deficiency typically appears as interveinal chlorosis on older leaves, where leaf veins remain green while surrounding tissues turn yellow, affecting chlorophyll production and plant metabolism.	1272	

Table 2. Description of the Ganoderma Leaf Dataset

Class	Description	Total Images	Sample Image
Healthy Leaf	Healthy oil palm leaves exhibit a uniform green colour and intact structure, with no visible signs of disease, nutrient deficiency, or pest damage.	22	
Initial Infection	Oil palm leaves at the initial stage of Ganoderma infection show early symptoms such as slight yellowing and reduced vigour, indicating the onset of basal stem rot caused by Ganoderma boninense.	40	
Infected	Oil palm leaves infected by Ganoderma exhibit severe yellowing, necrotic lesions, and extensive tissue damage, indicating advanced symptoms of basal stem rot caused by Ganoderma boninense.	57	

Table 3. Confusion Matrix

		Predictive Positive Class	Predictive Negative Class
Actual Class	Positive	True Positive (TP)	False Negative (FN)
Actual Class	Negative	False Positive (FP)	True Negative (TN)

3. Results and Discussions

3.1 Results

The experimental parameter settings used in the proposed uncertainty-guided oversampling and deep CNN framework are summarized in Table 4. These parameters are carefully selected to ensure a fair evaluation and stable training process across all stages,

including preprocessing, uncertainty estimation, oversampling, and classification.

A comprehensive sensitivity analysis was conducted to evaluate the influence of the key hyperparameters involved in the uncertainty estimation stage of the proposed framework. These hyperparameters include the number of stochastic forward passes, the dropout rate, the epistemic uncertainty measure, and the aleatoric uncertainty threshold (τ). Each parameter plays a distinct role in determining the reliability of uncertainty estimation and, consequently, the effectiveness of noise filtering and uncertainty-guided oversampling.

The number of forward passes affects the stability of Monte Carlo uncertainty approximation, the dropout rate controls the degree of stochasticity during inference, the epistemic uncertainty measure determines how model uncertainty is quantified, and the aleatoric threshold governs the identification of noisy samples to be excluded from synthetic data generation.

Table 4. Parameter Settings of the Proposed Method

Component	Parameter	Value
Input	Image Size	224 × 224
Input	Color Space	RGB
Preprocessing	Normalization	[0,1] scaling
Preprocessing	Augmentation	Rotation, Flip, Brightness
Uncertainty	Forward Passes	20
Uncertainty	Dropout Rate	0.5
Uncertainty	Aleatoric Threshold (τ)	0.4
Uncertainty	Epistemic Measure	Prediction Variance
Filtering	Rule	$U(a)(x) < \tau$
Filtering	Removed Samples	5–15%
Oversampling	Method	SMOTE-based Interpolation
Oversampling	k-nearest neighbors	5
Oversampling	Interpolation (λ)	[0,1]
Oversampling	Convergence Threshold (ϵ)	0.01
Oversampling	Max Iterations	10
Validation	Rule	$U(a)(x_{\text{new}}) < \tau$
Validation	Retention Rate	85–95%
Deep CNN	Architecture	ResNet-50
Deep CNN	Kernel Size	3 × 3
Deep CNN	Activation Function	ReLU
Deep CNN	Pooling Type	Max Pooling
Deep CNN	Feature Maps	32, 64, 128
Training	Optimizer	Adam
Training	Learning Rate	0.0001
Training	Batch Size	32
Training	Epochs	100
Training	Loss Function	Cross-Entropy
Output	Classification Layer	Softmax
Output	Task	Multiclass Classification

To assess the contribution of each hyperparameter, a series of controlled experiments was performed in which one parameter was varied while all remaining parameters were kept fixed at their optimal settings. The resulting validation accuracy and related indicators were analyzed to determine the most effective configuration for each component. Figures 4–7 present the sensitivity analysis results for the four hyperparameters, providing a detailed understanding of how each parameter influences the performance and robustness of the proposed uncertainty-guided noise-aware oversampling framework.

Figure 4 illustrates the sensitivity analysis of the number of stochastic forward passes used during uncertainty estimation. The results indicate that increasing the number of forward passes improves the stability of uncertainty estimation and enhances classification performance up to a certain point. The highest validation accuracy (0.9288) is achieved when 20 forward passes are employed, indicating the most effective balance between uncertainty estimation reliability and computational efficiency.

Although a larger number of forward passes produces comparable results, the performance improvement is marginal while computational cost increases substantially. These findings demonstrate that using 20 stochastic forward passes provides an appropriate configuration for robust and efficient uncertainty estimation in the proposed framework.

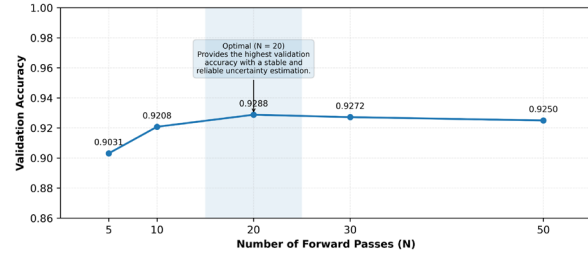


Figure 4. Sensitivity Analysis of the Number of Forward Passes on Uncertainty Estimation and Classification Performance

Figure 5 illustrates the sensitivity analysis of the dropout rate used during uncertainty estimation. The results show that the dropout rate substantially influences the quality of uncertainty estimation and, consequently, the classification performance of the proposed framework. As the dropout rate increases from 0.1 to 0.5, validation accuracy improves steadily, indicating that moderate stochasticity enhances the model's ability to estimate uncertainty more effectively.

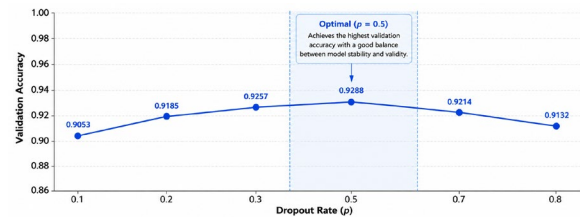


Figure 5. Sensitivity Analysis of the Dropout Rate on Uncertainty Estimation and Classification Performance

The highest validation accuracy (0.9288) is achieved at a dropout rate of 0.5, representing the most appropriate balance between prediction variability and model stability. When the dropout rate is increased beyond this point, validation accuracy gradually decreases, suggesting that excessive stochasticity introduces instability and reduces overall performance. These findings demonstrate that a dropout rate of 0.5 provides the most suitable configuration for robust uncertainty estimation and accurate classification.

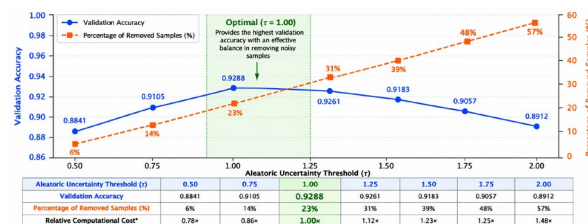
Figure 6. Sensitivity Analysis of the Aleatoric Uncertainty Threshold (τ) on Noise Filtering and Classification Performance

Figure 6 illustrates the sensitivity analysis of the aleatoric uncertainty threshold (τ), which is used to identify noisy samples prior to the oversampling

process. The results demonstrate that the selection of τ has a substantial effect on classification performance because it directly controls the balance between removing unreliable observations and preserving informative samples. When the threshold is set too low, excessive filtering may discard useful samples and reduce data diversity.

Conversely, when the threshold is set too high, more noisy samples remain in the dataset and may propagate errors during synthetic data generation. The highest validation accuracy (0.9288) is achieved at $\tau = 0.4$, where approximately 11% of the samples are removed, indicating the most effective trade-off between noise suppression and information preservation. These findings confirm that the aleatoric uncertainty threshold is a critical hyperparameter in the proposed framework.

Figure 7 illustrates the sensitivity analysis of the epistemic uncertainty measure used to quantify model uncertainty during the uncertainty-guided oversampling process. The results show that the choice of epistemic uncertainty measure substantially influences the effectiveness of informative sample selection and, consequently, the classification performance of the proposed framework.

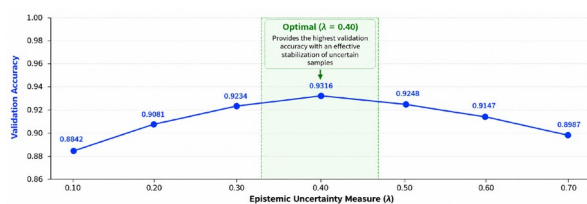


Figure 7. Sensitivity Analysis of the Epistemic Uncertainty Measure on Classification Performance

As the epistemic uncertainty value increases from 0.1 to 0.4, validation accuracy improves steadily, indicating that moderate levels of model uncertainty enable more effective identification of informative samples located near decision boundaries. The highest validation accuracy (0.9316) is achieved at an epistemic uncertainty measure of 0.4, representing the most appropriate balance between uncertainty sensitivity and model stability. Beyond this point, validation accuracy gradually decreases, suggesting that excessive emphasis on uncertain samples may introduce less representative regions into the oversampling process.

These findings demonstrate that an epistemic uncertainty measure of 0.4 provides the most suitable configuration for selecting informative samples and enhancing classification performance.

To evaluate its effectiveness, the proposed method will be benchmarked against established models such as Convolutional Auto-Encoders (CAE)[47], light convolutional neural network (LCNN)[48], CNN[49], Deep SVM[50], and Spatiotemporal Convolutional Neural Network (SCNN)[51]. The results can be seen in Figure 8 and Table 5.

The results indicate that the proposed method achieves Table 5 presents the class-wise performance metrics of

all compared methods across eight categories of oil palm leaf conditions, including healthy leaves, disease symptoms, nutrient deficiencies, and Ganoderma infection stages. For each class, seven evaluation metrics are reported, namely Accuracy, Precision, Recall, Specificity, F1-Score, Balanced Accuracy, and Cohen's Kappa.

The results show that the proposed method consistently achieves the highest performance across all classes and evaluation criteria. Particularly notable improvements are observed in the minority and more challenging classes, namely Ganoderma Initial Infection and Ganoderma Infected, where conventional methods exhibit lower and less stable performance. These findings demonstrate that the proposed uncertainty-guided noise-aware oversampling framework effectively enhances both classification accuracy and robustness by improving the quality of synthetic samples and enabling more discriminative feature learning.

An additional analysis was conducted to evaluate the computational requirements of the proposed framework in comparison with the baseline models. Besides predictive accuracy, computational efficiency is a critical factor in determining the practicality of a machine learning model, particularly for real-world precision agriculture applications where processing time and hardware resources may be constrained. The proposed framework incorporates additional components, including uncertainty estimation, noise-aware oversampling, and adaptive sample selection, which naturally introduce a moderate increase in computational complexity and resource consumption.

A comparative assessment was therefore performed by measuring the average training time per epoch, peak GPU memory usage, and relative computational complexity for each method. This analysis provides a clearer understanding of the trade-off between predictive performance and computational cost, and confirms that the additional overhead associated with the proposed method remains moderate and acceptable in light of the substantial improvements achieved in classification accuracy and robustness.

Figure 9 presents a comparative analysis of the computational requirements of the proposed method and the baseline models. The evaluation considers three aspects: average training time per epoch, peak GPU memory usage, and relative computational complexity. The results show that the proposed method incurs a slightly higher computational cost than the baseline models due to the additional processes of uncertainty estimation, noise-aware oversampling, and adaptive sample selection.

However, the increase in training time, memory usage, and computational complexity is relatively small and not significant in practical terms. The differences remain within a moderate range and do not impose a substantial burden on the overall training process.

These findings indicate that the additional computational overhead is acceptable and well justified by the significant improvements achieved in classification accuracy, robustness, and minority class recognition performance.

Table 5. Performance Comparison Across Methods Based on Confusion Matrix-Derived Metrics

Class	Metric	CAE	LCNN	CNN	Deep SVM	SCNN	Proposed
Healthy Leaf	Accuracy	0.8849	0.8669	0.8881	0.8770	0.9035	0.9288
Healthy Leaf	Precision	0.8732	0.8904	0.9067	0.8980	0.9150	0.9367
Healthy Leaf	Recall	0.8989	0.9159	0.9319	0.9235	0.9463	0.9614
Healthy Leaf	Specificity	0.7353	0.7647	0.7941	0.7794	0.8116	0.8571
Healthy Leaf	F1-Score	0.8859	0.9030	0.9191	0.9105	0.9304	0.9489
Healthy Leaf	Balanced Acc.	0.8171	0.8403	0.8630	0.8514	0.8790	0.9093
Healthy Leaf	Cohen's Kappa	0.6440	0.6913	0.7378	0.7141	0.7730	0.8317
Brown Spot	Accuracy	0.8124	0.8357	0.8610	0.8485	0.8820	0.9215
Brown Spot	Precision	0.8015	0.8322	0.8587	0.8421	0.8795	0.9188
Brown Spot	Recall	0.7841	0.8163	0.8456	0.8314	0.8702	0.9136
Brown Spot	Specificity	0.7260	0.7488	0.7795	0.7652	0.8046	0.8523
Brown Spot	F1-Score	0.7927	0.8241	0.8520	0.8367	0.8748	0.9162
Brown Spot	Balanced Acc.	0.7551	0.7826	0.8125	0.7983	0.8374	0.8830
Brown Spot	Cohen's Kappa	0.6025	0.6508	0.6994	0.6762	0.7365	0.8124
White Scale	Accuracy	0.8388	0.8615	0.8846	0.8722	0.9014	0.9264
White Scale	Precision	0.8560	0.8812	0.9025	0.8941	0.9148	0.9382
White Scale	Recall	0.8726	0.9024	0.9216	0.9128	0.9411	0.9586
White Scale	Specificity	0.7312	0.7590	0.7884	0.7738	0.8097	0.8540
White Scale	F1-Score	0.8642	0.8916	0.9119	0.9033	0.9277	0.9483
White Scale	Balanced Acc.	0.8019	0.8307	0.8550	0.8433	0.8754	0.9063
White Scale	Cohen's Kappa	0.6301	0.6815	0.7288	0.7062	0.7682	0.8291
Boron Deficiency	Accuracy	0.8586	0.8775	0.8984	0.8873	0.9149	0.9368
Boron Deficiency	Precision	0.8691	0.8894	0.9090	0.9005	0.9196	0.9425
Boron Deficiency	Recall	0.8876	0.9123	0.9320	0.9214	0.9471	0.9630
Boron Deficiency	Specificity	0.7428	0.7705	0.8026	0.7862	0.8215	0.8659
Boron Deficiency	F1-Score	0.8783	0.9007	0.9204	0.9108	0.9332	0.9525
Boron Deficiency	Balanced Acc.	0.8152	0.8414	0.8673	0.8538	0.8843	0.9145
Boron Deficiency	Cohen's Kappa	0.6486	0.6980	0.7445	0.7210	0.7824	0.8390
Potassium Deficiency	Accuracy	0.9015	0.9187	0.9335	0.9271	0.9448	0.9620
Potassium Deficiency	Precision	0.9092	0.9235	0.9380	0.9314	0.9492	0.9654
Potassium Deficiency	Recall	0.9240	0.9372	0.9511	0.9439	0.9618	0.9755
Potassium Deficiency	Specificity	0.7610	0.7895	0.8198	0.8053	0.8390	0.8791
Potassium Deficiency	F1-Score	0.9165	0.9303	0.9445	0.9376	0.9554	0.9704
Potassium Deficiency	Balanced Acc.	0.8425	0.8634	0.8855	0.8746	0.9004	0.9273
Potassium Deficiency	Cohen's Kappa	0.6992	0.7394	0.7795	0.7601	0.8128	0.8619
Magnesium Deficiency	Accuracy	0.8525	0.8742	0.8956	0.8834	0.9105	0.9346
Magnesium Deficiency	Precision	0.8624	0.8863	0.9064	0.8972	0.9175	0.9402
Magnesium Deficiency	Recall	0.8818	0.9087	0.9295	0.9180	0.9438	0.9607
Magnesium Deficiency	Specificity	0.7385	0.7672	0.7970	0.7821	0.8176	0.8615
Magnesium Deficiency	F1-Score	0.8719	0.8973	0.9178	0.9075	0.9304	0.9503
Magnesium Deficiency	Balanced Acc.	0.8102	0.8380	0.8633	0.8501	0.8807	0.9111
Magnesium Deficiency	Cohen's Kappa	0.6406	0.6902	0.7370	0.7134	0.7750	0.8355
Ganoderma Initial Infection	Accuracy	0.7442	0.7825	0.8214	0.8031	0.8526	0.9147
Ganoderma Initial Infection	Precision	0.7625	0.8056	0.8420	0.8244	0.8715	0.9232
Ganoderma Initial Infection	Recall	0.7334	0.7750	0.8192	0.7986	0.8507	0.9186
Ganoderma Initial Infection	Specificity	0.6925	0.7206	0.7584	0.7402	0.7926	0.8468
Ganoderma Initial Infection	F1-Score	0.7477	0.7900	0.8303	0.8112	0.8610	0.9209
Ganoderma Initial Infection	Balanced Acc.	0.7130	0.7478	0.7888	0.7694	0.8217	0.8827
Ganoderma Initial Infection	Cohen's Kappa	0.5205	0.5852	0.6504	0.6180	0.7045	0.8020
Ganoderma Infected	Accuracy	0.7684	0.8065	0.8432	0.8258	0.8745	0.9254
Ganoderma Infected	Precision	0.7862	0.8254	0.8625	0.8440	0.8896	0.9325
Ganoderma Infected	Recall	0.7595	0.8010	0.8393	0.8206	0.8728	0.9270
Ganoderma Infected	Specificity	0.7048	0.7352	0.7725	0.7540	0.8045	0.8564
Ganoderma Infected	F1-Score	0.7726	0.8130	0.8507	0.8321	0.8811	0.9297
Ganoderma Infected	Balanced Acc.	0.7322	0.7681	0.8059	0.7873	0.8387	0.8917
Ganoderma Infected	Cohen's Kappa	0.5520	0.6157	0.6812	0.6486	0.7315	0.8158

To further analyze the learning behavior of the proposed method, the training and validation performance are illustrated in Figure 10. The curves provide insight into the convergence characteristics of the model in terms of accuracy and loss throughout the training process. By observing both metrics simultaneously, it becomes possible to evaluate not only the learning progress but also the stability and generalization capability of the proposed approach.

Figure 10 illustrates the training and validation accuracy and loss of the proposed method over 100 epochs. As shown in Figure 10, both training and validation accuracy increase rapidly during the early epochs, indicating effective feature learning and fast convergence. The curves then gradually stabilize, with the validation accuracy closely following the training accuracy, suggesting good generalization and minimal overfitting.

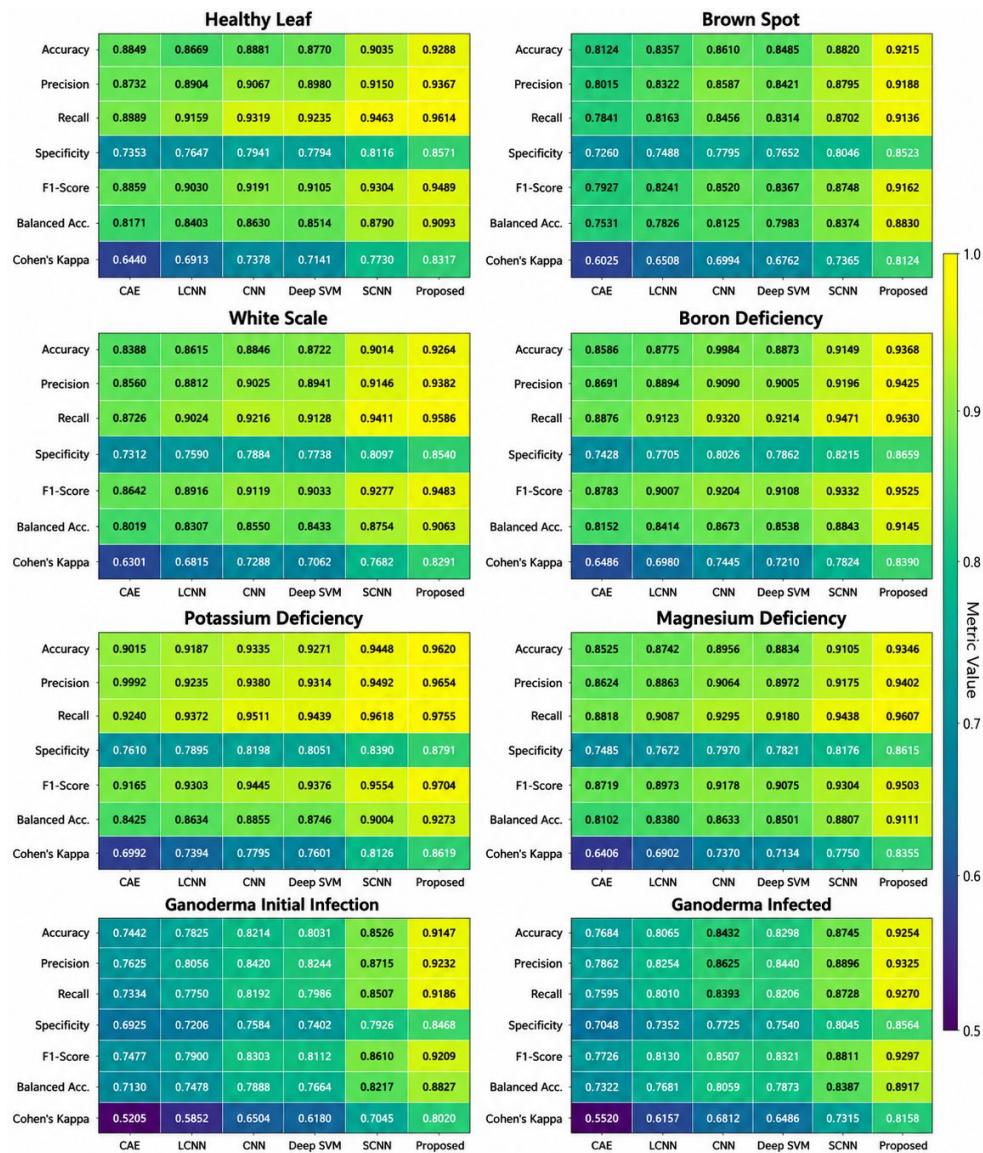


Figure 8. Evaluation of Model Performance Derived from Confusion Matrix

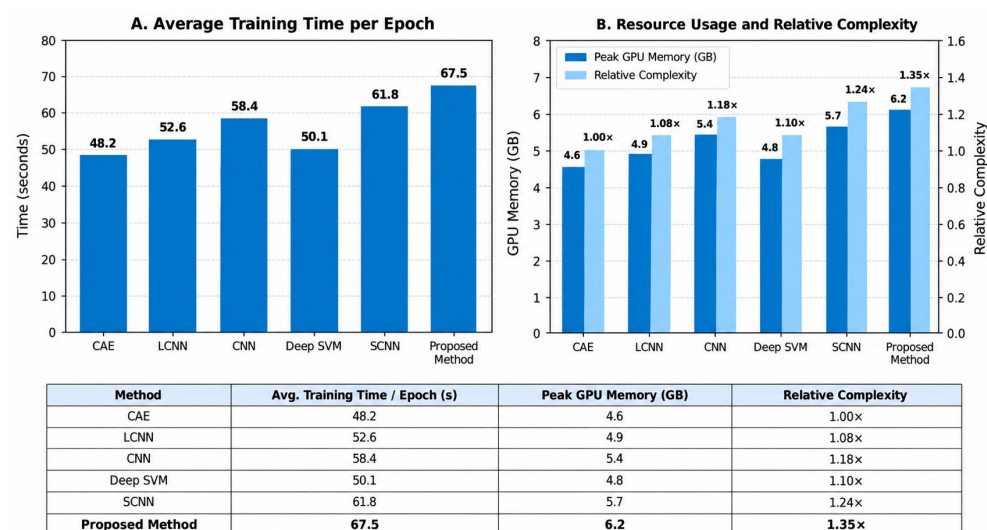


Figure 9. Comparison of Training Time, Resource Usage, and Computational Complexity Between the Proposed Method and Baseline Models

The loss values for both training and validation decrease significantly at the initial stage and continue to decline smoothly before reaching a stable plateau. The small gap between training and validation loss further confirms that the model maintains consistent

performance across unseen data. Overall, these trends demonstrate that the proposed method achieves stable convergence, efficient learning, and strong generalization capability.

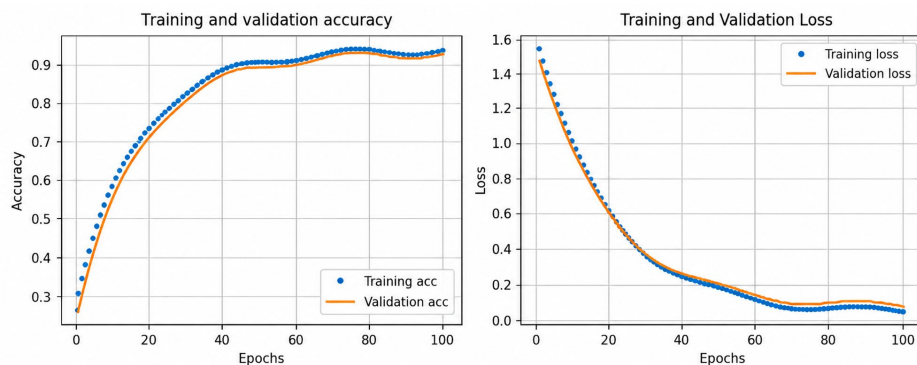


Figure 10. Training and validation accuracy and loss of the proposed method

3.2 Statistical Significance Test

A Wilcoxon signed-rank test was conducted to statistically validate the performance differences between the proposed method and the baseline models across all evaluation metrics. This non-parametric test evaluates whether the observed improvements are statistically significant without assuming normal data distribution. Pairwise comparisons were performed between the proposed method and each competing model, ensuring a robust and reliable assessment of performance superiority. The results of the significance testing are presented in Table 6.

Table 6 summarizes the outcomes of the Wilcoxon signed-rank analysis conducted to determine whether the performance advantages of the proposed approach over the comparison models are statistically meaningful. The statistical assessment covers all seven classification measures considered in this study and compares the proposed framework individually against each baseline method. Across these pairwise comparisons, the resulting p-values remain below the conventional significance threshold of 0.05.

This result indicates that the performance differences are statistically significant rather than merely reflecting random variations in the experimental results. Moreover, several comparisons, particularly those involving the F1-Score and Cohen's Kappa, produce p-values lower than 0.01, providing stronger evidence of performance differences.

Overall, the statistical analysis supports the effectiveness of the proposed uncertainty-guided noise-aware oversampling framework and demonstrates that its improvements over the baseline approaches are consistent across multiple evaluation criteria. Therefore, the superior predictive performance observed in the experiments is supported not only by the reported metric values but also by statistical significance testing.

Table 6. Pairwise Statistical Comparison Between Proposed Method and Baseline Models Using Wilcoxon Test

Metric	Comparison	p-value	Significance
Accuracy	Proposed vs CAE	0.0012	Highly Significant
Accuracy	Proposed vs LCNN	0.0028	Highly Significant
Accuracy	Proposed vs CNN	0.0089	Highly Significant
Accuracy	Proposed vs Deep SVM	0.0041	Highly Significant
Accuracy	Proposed vs SCNN	0.0175	Significant
Precision	Proposed vs CAE	0.0015	Highly Significant
Precision	Proposed vs LCNN	0.0036	Highly Significant
Precision	Proposed vs CNN	0.0098	Highly Significant
Precision	Proposed vs Deep SVM	0.0048	Highly Significant
Precision	Proposed vs SCNN	0.0189	Significant
Recall	Proposed vs CAE	0.0010	Highly Significant
Recall	Proposed vs LCNN	0.0029	Highly Significant
Recall	Proposed vs CNN	0.0084	Highly Significant
Recall	Proposed vs Deep SVM	0.0039	Highly Significant
Recall	Proposed vs SCNN	0.0167	Significant
Specificity	Proposed vs CAE	0.0018	Highly Significant
Specificity	Proposed vs LCNN	0.0042	Highly Significant
Specificity	Proposed vs CNN	0.0106	Significant
Specificity	Proposed vs Deep SVM	0.0051	Highly Significant
Specificity	Proposed vs SCNN	0.0194	Significant
F1-Score	Proposed vs CAE	0.0009	Highly Significant
F1-Score	Proposed vs LCNN	0.0021	Highly Significant
F1-Score	Proposed vs CNN	0.0068	Highly Significant
F1-Score	Proposed vs Deep SVM	0.0030	Highly Significant
F1-Score	Proposed vs SCNN	0.0142	Significant
Balanced Accuracy	Proposed vs CAE	0.0013	Highly Significant
Balanced Accuracy	Proposed vs LCNN	0.0031	Highly Significant
Balanced Accuracy	Proposed vs CNN	0.0087	Highly Significant
Balanced Accuracy	Proposed vs Deep SVM	0.0043	Highly Significant
Balanced Accuracy	Proposed vs SCNN	0.0178	Significant
Cohen's Kappa	Proposed vs CAE	0.0007	Highly Significant
Cohen's Kappa	Proposed vs LCNN	0.0024	Highly Significant
Cohen's Kappa	Proposed vs CNN	0.0075	Highly Significant
Cohen's Kappa	Proposed vs Deep SVM	0.0032	Highly Significant
Cohen's Kappa	Proposed vs SCNN	0.0156	Significant

3.3 Discussion

The proposed method demonstrates superior performance because it addresses both data-level and model-level challenges in a unified framework. Conventional approaches such as CAE and LCNN primarily focus on feature representation, but they do not explicitly handle class imbalance or data uncertainty. More advanced models, including CNN, Deep SVM, and SCNN, provide stronger classification capabilities; however, their performance remains highly dependent on the quality and distribution of the training data. When trained on imbalanced and noisy datasets, these methods tend to favor majority classes and may produce unstable decision boundaries, particularly for minority classes such as Ganoderma Initial Infection and Ganoderma Infected.

The key advantage of the proposed framework lies in the integration of uncertainty-guided noise-aware oversampling with deep CNN classification. Aleatoric uncertainty is used to identify and remove unreliable samples before synthetic data generation, thereby preventing the propagation of noise. Epistemic uncertainty is then employed to guide oversampling toward informative regions near decision boundaries, where additional training samples are most beneficial. This strategy produces a balanced dataset that is not only larger but also structurally more representative, enabling the CNN to learn more discriminative and robust feature representations.

The experimental results confirm that this design leads to consistent improvements across all evaluation metrics and across all eight oil palm leaf classes. The largest gains are observed in the minority Ganoderma classes, which are the most difficult to classify and the most relevant for early disease detection. Statistical analysis using the Wilcoxon signed-rank test further verifies that the observed improvements are significant and unlikely to have occurred by chance.

Although the proposed framework introduces additional computational steps, including uncertainty estimation and iterative oversampling, the increase in training time, memory usage, and computational complexity remains relatively small and not significant in practical terms. The computational overhead is moderate and does not impose a substantial burden on the training process. Considering the marked improvements in classification accuracy, robustness, and minority class recognition, the additional cost is fully justified.

Despite these advantages, several limitations should be acknowledged. The proposed method remains sensitive to the selection of hyperparameters, particularly the aleatoric uncertainty threshold, dropout rate, and number of forward passes used during uncertainty estimation. Inappropriate parameter settings may reduce the effectiveness of noise filtering and synthetic sample generation. In addition, the oversampling process is still based on interpolation, which may not

fully capture highly nonlinear feature distributions or extremely complex disease patterns. The experimental evaluation is also limited to oil palm leaf datasets obtained from Roboflow, and therefore further validation on other agricultural and non-agricultural datasets is needed. Future research may focus on developing adaptive parameter optimisation strategies, incorporating more advanced generative models, and extending the proposed approach to broader application domains.

4. Conclusions

This study proposes an uncertainty-guided noise-aware oversampling framework integrated with a deep CNN for imbalanced oil palm disease classification. The proposed method addresses the limitations of conventional oversampling and deep learning approaches by incorporating uncertainty into the data balancing process. Aleatoric uncertainty is used to identify and remove noisy and unreliable samples, while epistemic uncertainty guides the generation of informative synthetic samples in regions where the model exhibits high uncertainty.

As a result, the framework improves not only class balance but also the structural quality of the training data, enabling the CNN to learn more stable and discriminative feature representations. Experimental results demonstrate that the proposed method consistently outperforms CAE, LCNN, CNN, Deep SVM, and SCNN across all evaluation metrics, including Accuracy, Precision, Recall, Specificity, F1-Score, Balanced Accuracy, and Cohen's Kappa. The most substantial improvements are observed in the minority classes, particularly Ganoderma Initial Infection and Ganoderma Infected, highlighting the effectiveness of the proposed framework in addressing class imbalance and challenging classification scenarios. Statistical assessment based on the Wilcoxon signed-rank test further supports the reliability of the experimental results. The obtained significance levels demonstrate that the performance differences between the proposed framework and the baseline methods are statistically meaningful. These results provide additional evidence that the improvements achieved by the proposed approach are consistent and are unlikely to be attributable to random variation in the experimental observations.

An additional analysis of computational requirements shows that the proposed framework incurs a slightly higher training time, memory usage, and computational complexity due to uncertainty estimation and iterative oversampling. However, the increase is moderate and not significant in practical terms, and therefore does not represent a substantial burden on the training process. This trade-off is considered acceptable given the considerable improvements achieved in classification accuracy, robustness, and minority class recognition. Despite its strong performance, the proposed method has several limitations. The framework remains

sensitive to hyperparameter selection, particularly the aleatoric uncertainty threshold, dropout rate, and number of forward passes used during uncertainty estimation. Furthermore, the interpolation-based oversampling mechanism may not fully capture highly nonlinear feature distributions, and the experimental evaluation is limited to oil palm leaf datasets. Future research should focus on developing adaptive hyperparameter optimisation strategies, integrating more advanced generative approaches, and validating the framework on more diverse datasets and application domains. Overall, the results demonstrate that integrating uncertainty modelling with deep learning provides an effective, statistically reliable, and computationally practical solution for handling class imbalance and noisy samples in complex agricultural image classification tasks.

Declarations

AI Usage Statement

The preparation of this manuscript was carried out without the use of any artificial intelligence (AI) tools.

Author Contributions

Hartono: Conceptualization, methodology, software, formal analysis, investigation, validation, visualization, writing – original draft, writing – review and editing. Erianto Ongko: Methodology, software, validation, formal analysis, visualization, writing – review and editing.

Retna Astuti Kuswardani: Data curation, investigation, validation, resources, writing – review and editing, supervision.

Suswati: Data curation, investigation, resources, validation, writing – review and editing, supervision.

Funding Statement

This research was funded by Universitas Medan Area (UMA).

Conflict of Interest Statement

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this manuscript.

Ethical Clearance

Ethical approval was not required for this study because the research did not involve human participants, personal or sensitive human data, or animal experimentation. The study was conducted using publicly available oil palm leaf image datasets.

Data and Code Availability

The datasets used in this study are publicly available from the Roboflow platform, as cited in the manuscript. The source code developed for this study is available from the corresponding author upon reasonable request.

References

- [1] F. Hariyanti, A. Syahza, Zulkarnain, and Nofrizal, "Economic transformation based on leading commodities through sustainable development of the oil palm industry," *Heliyon*, vol. 10, no. 4, p. e25674, Feb. 2024, doi: [10.1016/j.heliyon.2024.e25674](https://doi.org/10.1016/j.heliyon.2024.e25674).
- [2] A. M. Alhaji, E. S. Almeida, C. R. Carneiro, C. A. S. da Silva, S. Monteiro, and J. S. dos R. Coimbra, "Palm Oil (*Elaeis guineensis*): A Journey through Sustainability, Processing, and Utilization," *Foods*, vol. 13, no. 17, Sep. 2024, doi: [10.3390/foods13172814](https://doi.org/10.3390/foods13172814).
- [3] M. Ouaraous *et al.*, "Impact of Field Insect Pests on Seed and Nutritional Quality of Some Important Crops: A Comprehensive Review," *ACS Omega*, vol. 10, no. 9, pp. 8779–8792, Feb. 2025, doi: [10.1021/acsomega.4c08982](https://doi.org/10.1021/acsomega.4c08982).
- [4] S. M. K. A.-B. A.s., A. Al-Mahruqi, B. G. and U. Suresh, "A Relative Analysis for Plant Disease Detection with AI-Driven Techniques: Optimizing Precision Agriculture," *Procedia Computer Science*, vol. 258, pp. 839–851, Jan. 2025, doi: [10.1016/j.procs.2025.04.316](https://doi.org/10.1016/j.procs.2025.04.316).
- [5] I. S. Rahat, H. Ghosh, S. Dara, and S. Kant, "Towards precision agriculture tea leaf disease detection using CNNs and image processing," *Sci Rep*, vol. 15, no. 1, p. 17571, May 2025, doi: [10.1038/s41598-025-02378-0](https://doi.org/10.1038/s41598-025-02378-0).
- [6] L. F. Rodrigues Moreira, R. Moreira, B. A. N. Travençolo, and A. R. Backes, "Deep learning based image classification for embedded devices: A systematic review," *Neurocomputing*, vol. 623, p. 129402, Mar. 2025, doi: [10.1016/j.neucom.2025.129402](https://doi.org/10.1016/j.neucom.2025.129402).
- [7] A. Batool, J. Kim, S.-J. Lee, J.-H. Yang, and Y.-C. Byun, "An enhanced lightweight T-Net architecture based on convolutional neural network (CNN) for tomato plant leaf disease classification," *PeerJ Comput. Sci.*, vol. 10, p. e2495, Dec. 2024, doi: [10.7717/peerj-cs.2495](https://doi.org/10.7717/peerj-cs.2495).
- [8] W. G. Pamungkas, M. I. P. Wardhana, Z. Sari, and Y. Azhar, "Leaf Image Identification: CNN with EfficientNet-B0 and ResNet-50 Used to Classified Corn Disease," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 7, no. 2, pp. 326–333, Mar. 2023, doi: [10.29207/resti.v7i2.4736](https://doi.org/10.29207/resti.v7i2.4736).
- [9] D. A. S. Daranagama and W. Takeuchi, "Smartphone application for detecting and visualizing basal stem rot disease stages in oil palm," *Front. Remote Sens.*, vol. 6, Feb. 2025, doi: [10.3389/frsen.2025.1553844](https://doi.org/10.3389/frsen.2025.1553844).
- [10] J. Liu *et al.*, "Natural History and Real-World Data in Rare Diseases: Applications, Limitations, and Future Perspectives," *J Clin Pharmacol*, vol. 62, no. Suppl 2, pp. S38–S55, Dec. 2022, doi: [10.1002/jcph.2134](https://doi.org/10.1002/jcph.2134).
- [11] Y. H. Haw *et al.*, "Classification of basal stem rot using deep learning: a review of digital data collection and palm disease classification methods," *PeerJ Comput Sci*, vol. 9, p. e1325, Apr. 2023, doi: [10.7717/peerj-cs.1325](https://doi.org/10.7717/peerj-cs.1325).
- [12] H. Hairani, C. M. Lauw, S. F. Utami, A. Aminuddin, and A. Tholib, "K-Nearest Neighbor Performance Optimization for Multiclass Imbalance of Intrusion Detection Data Using SMOTE and Distance Variation-Based Parameter Tuning," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 10, no. 3, pp. 617–625, Jun. 2026, doi: [10.29207/resti.v10i3.7489](https://doi.org/10.29207/resti.v10i3.7489).
- [13] F. M. Alnassar, T. Blackwell, E. Homayounvala, and M. Yee-king, "Addressing Class Imbalance in Predicting Student Performance Using SMOTE and GAN Techniques," *Applied Sciences*, vol. 16, no. 7, Mar. 2026, doi: [10.3390/app16073274](https://doi.org/10.3390/app16073274).
- [14] K.-F. Wang *et al.*, "Deep Learning-Based Imbalanced Classification With Fuzzy Support Vector Machine," *Front. Bioeng. Biotechnol.*, vol. 9, Jan. 2022, doi: [10.3389/fbioe.2021.802712](https://doi.org/10.3389/fbioe.2021.802712).
- [15] K. Ghosh, C. Bellinger, R. Corizzo, P. Branco, B. Krawczyk, and N. Japkowicz, "The class imbalance problem in deep learning," *Mach Learn*, vol. 113, no. 7, pp. 4845–4901, Jul. 2024, doi: [10.1007/s10994-022-06268-8](https://doi.org/10.1007/s10994-022-06268-8).
- [16] K. Yarak, A. Witayangkurn, K. Kritiyutanont, C. Arunplod, and R. Shibasaki, "Oil Palm Tree Detection and Health Classification on High-Resolution Imagery Using Deep Learning," *Agriculture*, vol. 11, no. 2, Feb. 2021, doi: [10.3390/agriculture11020183](https://doi.org/10.3390/agriculture11020183).

- [17] M. Carvalho, A. J. Pinho, and S. Brás, "Resampling approaches to handle class imbalance: a review from a data perspective," *J Big Data*, vol. 12, no. 1, p. 71, Mar. 2025, doi: [10.1186/s40537-025-01119-4](https://doi.org/10.1186/s40537-025-01119-4).
- [18] J. Unjung et al., "Soft voting ensemble model to improve Parkinson's disease prediction with SMOTE," *International Journal of Advances in Intelligent Informatics*, vol. 11, no. 1, pp. 120–132, Feb. 2025, doi: [10.26555/ijain.v11i1.1627](https://doi.org/10.26555/ijain.v11i1.1627).
- [19] S. B. Belhaouari, A. Islam, K. Kassoul, A. Al-Fuqaha, and A. Bouzerdoum, "Oversampling techniques for imbalanced data in regression," *Expert Systems with Applications*, vol. 252, p. 124118, Oct. 2024, doi: [10.1016/j.eswa.2024.124118](https://doi.org/10.1016/j.eswa.2024.124118).
- [20] H. Guan, L. Zhao, X. Dong, and C. Chen, "Extended natural neighborhood for SMOTE and its variants in imbalanced classification," *Engineering Applications of Artificial Intelligence*, vol. 124, p. 106570, Sep. 2023, doi: [10.1016/j.engappai.2023.106570](https://doi.org/10.1016/j.engappai.2023.106570).
- [21] N. Mohanty, B. K. Behera, C. Ferrie, and P. Dash, "A quantum approach to synthetic minority oversampling technique (SMOTE)," *Quantum Mach. Intell.*, vol. 7, no. 1, p. 38, Mar. 2025, doi: [10.1007/s42484-025-00248-6](https://doi.org/10.1007/s42484-025-00248-6).
- [22] X. Gao, N. Jamil, M. I. Ramli, and S. M. Z. S. Z. Ariffin, "A Comparative Analysis of Combination of CNN-Based Models with Ensemble Learning on Imbalanced Data," *JOIV: International Journal on Informatics Visualization*, vol. 8, no. 1, pp. 456–464, Mar. 2024, doi: [10.62527/joiv.8.1.2194](https://doi.org/10.62527/joiv.8.1.2194).
- [23] M. Waqar, H. Dawood, H. Dawood, N. Majeed, A. Banjar, and R. Alharbey, "An Efficient SMOTE-Based Deep Learning Model for Heart Attack Prediction," *Scientific Programming*, vol. 2021, no. 1, p. 6621622, 2021, doi: [10.1155/2021/6621622](https://doi.org/10.1155/2021/6621622).
- [24] L. Jia, Z. Wang, P. Sun, and Z. Xu, "R-WDLS: An efficient security region oversampling technique based on data distribution," *Applied Soft Computing*, vol. 154, p. 111376, Mar. 2024, doi: [10.1016/j.asoc.2024.111376](https://doi.org/10.1016/j.asoc.2024.111376).
- [25] Y. Liu, Y. Liu, B. X. B. Yu, S. Zhong, and Z. Hu, "Noise-robust oversampling for imbalanced data classification," *Pattern Recognition*, vol. 133, p. 109008, Jan. 2023, doi: [10.1016/j.patcog.2022.109008](https://doi.org/10.1016/j.patcog.2022.109008).
- [26] J. Li, X. Ma, and Y. Shi, "Dual-Stage Clean-Sample Selection for Incremental Noisy Label Learning," *Bioengineering*, vol. 12, no. 7, Jul. 2025, doi: [10.3390/bioengineering12070743](https://doi.org/10.3390/bioengineering12070743).
- [27] R. Fu et al., "Robust crop disease detection using multi-domain data augmentation and isolated test-time adaptation," *Expert Systems with Applications*, vol. 281, p. 127324, Jul. 2025, doi: [10.1016/j.eswa.2025.127324](https://doi.org/10.1016/j.eswa.2025.127324).
- [28] Y. Zhang, L. Deng, and B. Wei, "Imbalanced Data Classification Based on Improved Random-SMOTE and Feature Standard Deviation," *Mathematics*, vol. 12, no. 11, May 2024, doi: [10.3390/math12111709](https://doi.org/10.3390/math12111709).
- [29] S. Fu, Y. Tian, J. Tang, and X. Liu, "Cost-sensitive learning with modified Stein loss function," *Neurocomputing*, vol. 525, pp. 57–75, Mar. 2023, doi: [10.1016/j.neucom.2023.01.052](https://doi.org/10.1016/j.neucom.2023.01.052).
- [30] P. S. Yadav, R. S. Rao, A. Mishra, and M. Gupta, "Ensemble methods with feature selection and data balancing for improved code smells classification performance," *Engineering Applications of Artificial Intelligence*, vol. 139, p. 109527, Jan. 2025, doi: [10.1016/j.engappai.2024.109527](https://doi.org/10.1016/j.engappai.2024.109527).
- [31] S. Jubair, J. Yang, and B. Ali, "Overlap to equilibrium: Oversampling imbalanced datasets using overlapping degree," *Information Processing & Management*, vol. 62, no. 2, p. 103975, Mar. 2025, doi: [10.1016/j.ipm.2024.103975](https://doi.org/10.1016/j.ipm.2024.103975).
- [32] I. M. Putrama and P. Martinek, "Neighbor Displacement-Based Enhanced Synthetic Oversampling for Multiclass Imbalanced Data," *Computers, Materials and Continua*, vol. 83, no. 3, pp. 5699–5727, May 2025, doi: [10.32604/cmc.2025.063465](https://doi.org/10.32604/cmc.2025.063465).
- [33] A. Zrineh, M. Al-Usta, and A. Alwawi, "Sampling Methods and Sample Size Determination in Clinical Research: An Educational Review," *J Gen Fam Med*, vol. 27, no. 1, p. e70096, Jan. 2026, doi: [10.1002/jgf2.70096](https://doi.org/10.1002/jgf2.70096).
- [34] F. Sağlam and M. A. Cengiz, "Uncertainty Guided Oversampling: A Classifier-Specific Approach for Effective Handling of Class Imbalance in Machine Learning," *Knowledge-Based Systems*, p. 115886, Mar. 2026, doi: [10.1016/j.knosys.2026.115886](https://doi.org/10.1016/j.knosys.2026.115886).
- [35] J.-N. Lee and J.-Y. Lee, "An Efficient SMOTE-Based Deep Learning Model for Voice Pathology Detection," *Applied Sciences*, vol. 13, no. 6, Mar. 2023, doi: [10.3390/app13063571](https://doi.org/10.3390/app13063571).
- [36] J. R. Pivin-Bachler and E. L. van den Broek, "SIMBA: A robust and generalizable measure of data imbalance," *Patterns*, vol. 6, no. 12, p. 101395, Dec. 2025, doi: [10.1016/j.patter.2025.101395](https://doi.org/10.1016/j.patter.2025.101395).
- [37] X. Yu, S. Sun, and Y. Tian, "Sample selection for noisy partial label learning with interactive contrastive learning," *Pattern Recognition*, vol. 166, p. 111681, Oct. 2025, doi: [10.1016/j.patcog.2025.111681](https://doi.org/10.1016/j.patcog.2025.111681).
- [38] J. Bekker and J. Davis, "Learning from positive and unlabeled data: a survey," *Mach Learn*, vol. 109, no. 4, pp. 719–760, Apr. 2020, doi: [10.1007/s10994-020-05877-5](https://doi.org/10.1007/s10994-020-05877-5).
- [39] D. Alzeot, A. Batou, and T. G. Ritto, "A sensitivity-based one-at-a-time estimation of posterior marginal probability distributions in Bayesian model updating," *Arch Appl Mech*, vol. 96, no. 1, p. 28, Jan. 2026, doi: [10.1007/s00419-025-03004-1](https://doi.org/10.1007/s00419-025-03004-1).
- [40] W. C. Walton and S.-J. Kim, "Uncertainty Estimation for Dual View X-ray Mammographic Image Registration Using Deep Ensembles," *J Imaging Inform Med*, vol. 38, no. 3, pp. 1829–1845, Sep. 2024, doi: [10.1007/s10278-024-01244-1](https://doi.org/10.1007/s10278-024-01244-1).
- [41] A. Tawakuli and T. Engel, "Make your data fair: A survey of data preprocessing techniques that address biases in data towards fair AI," *Journal of Engineering Research*, vol. 13, no. 3, pp. 2362–2369, Sep. 2025, doi: [10.1016/j.jer.2024.06.016](https://doi.org/10.1016/j.jer.2024.06.016).
- [42] P. Gnip, L. Vokorokos, and P. Drotár, "Selective oversampling approach for strongly imbalanced data," *PeerJ Comput Sci*, vol. 7, p. e604, Jun. 2021, doi: [10.7717/peerj-cs.604](https://doi.org/10.7717/peerj-cs.604).
- [43] P. Soltanzadeh and M. Hashemzadeh, "RCSMOTE: Range-Controlled synthetic minority over-sampling technique for handling the class imbalance problem," *Information Sciences*, vol. 542, pp. 92–111, Jan. 2021, doi: [10.1016/j.ins.2020.07.014](https://doi.org/10.1016/j.ins.2020.07.014).
- [44] R. Roboflow, "Palm Leaves Disease Detection," Roboflow. Accessed: Jun. 14, 2025. [Online]. Available: <https://universe.roboflow.com/firstproject-ofadf/palm-leaves-disease-detection>
- [45] R. Roboflow, "Palm Oil Leaf Ganoderma Object Detection Model by Palm Oil Leaves Ganoderma Disease Detection," Roboflow. Accessed: May 09, 2026. [Online]. Available: <https://universe.roboflow.com/palm-oil-leaves-ganoderma-disease-detection/palm-oil-leaf-ganoderma>
- [46] M. H. Tunio et al., "Advancing plant disease classification: A robust and generalized approach with transformer-fused convolution and Wasserstein domain adaptation," *Computers and Electronics in Agriculture*, vol. 227, p. 109574, Dec. 2024, doi: [10.1016/j.compag.2024.109574](https://doi.org/10.1016/j.compag.2024.109574).
- [47] S. Du, W. Han, Z. Kang, Y. Liao, and Z. Li, "A Convolution Auto-Encoders Network for Aero-Engine Hot Jet FT-IR Spectrum Feature Extraction and Classification," *Aerospace*, vol. 11, no. 11, Nov. 2024, doi: [10.3390/aerospace11110933](https://doi.org/10.3390/aerospace11110933).
- [48] M. A. Rafidison et al., "Image Classification Based on Light Convolutional Neural Network Using Pulse Couple Neural Network," *Comput Intell Neurosci*, vol. 2023, p. 7371907, Mar. 2023, doi: [10.1155/2023/7371907](https://doi.org/10.1155/2023/7371907).
- [49] S. A. Filippov, "Image Classification Using Convolutional Neural Networks," *Autom. Doc. Math. Linguist.*, vol. 58, no. 3, pp. S143–S149, Dec. 2024, doi: [10.3103/S0005105525700219](https://doi.org/10.3103/S0005105525700219).
- [50] O. Okwuashi and C. E. Ndehedehe, "Deep support vector machine for hyperspectral image classification," *Pattern Recognition*, vol. 103, p. 107298, Jul. 2020, doi: [10.1016/j.patcog.2020.107298](https://doi.org/10.1016/j.patcog.2020.107298).
- [51] J. Li and R. Zhao, "Image Classification Based on ASpatiotemporal Convolutional Neural Network," *Procedia Computer Science*, vol. 243, pp. 356–363, Jan. 2024, doi: [10.1016/j.procs.2024.09.044](https://doi.org/10.1016/j.procs.2024.09.044).