SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V11I5P119

Volume 11 Issue 5, 195-206, May 2024

Original Article

Predicting Banana Leaf Diseases: Feature Extraction with BL-FEOT and Enhanced Classification Using the BAT-KNN Hybrid Algorithm

Ravi Kumar Tirandasu[1], Prasanth Yalla[2]

[1,2]*Department of CSE, Koneru Lakshmaiah Education Foundation, Andhra Pradesh, India.*

[1]*Corresponding Author : ravi.tirandasu@gmail.com*

Received: 22 March 2024 Revised: 25 April 2024 Accepted: 17 May 2024 Published: 31 May 2024

Abstract *- Banana cultivation, fundamental to many rural economies, confronts persistent threats from many foliar diseases. Rapid and precise disease identification is critical for effective management and containment. This research introduces a pioneering method for detecting and classifying diseases on banana leaves, specifically targeting Cordana, Pestalotiopsis, and Sigatoka. Our process coined the Banana Leaf Feature Extraction and Optimization Technique (BL-FEOT), is a systematic approach encompassing manual background elimination, green color removal to emphasize the disease manifestations, and contour detection to mark out the compromised zones. Distinctive features are extracted for each disease type, culminating in a comprehensive dataset tailored for disease classification. Incorporating an extensive suite of feature extraction techniques, our methodology ensures the maximal retrieval of pivotal information from the infected sites. We leverage the BAT Optimization Technique to fine-tune the extracted features, especially within the RGB color space. This streamlines the feature dimensions and zeroes in on the most relevant components, amplifying the efficiency of the ensuing classification phase. The seminal contribution lies in integrating the BAT Optimization Technique with the K-Nearest Neighbors (KNN) algorithm, resulting in the novel hybrid algorithm, BAT+KNN. This algorithm is applied to the refined feature set for classification purposes. To substantiate the efficacy of BL-FEOT, its performance metrics are juxtaposed against prevailing algorithms using the same feature dataset. A thorough evaluation, encapsulating metrics such as precision, recall, accuracy, F1 score, ROC curve, and error rate, is presented. The experimental results, derived from available datasets, underscore the superior capabilities of our hybrid BAT+KNN algorithm in banana leaf disease identification. This research asserts that the BL-FEOT, powered by the BAT+KNN hybrid algorithm, offers a ground-breaking avenue for the automated and precise detection of banana leaf diseases. Its potential integration into real-time monitoring systems could revolutionize early disease detection and intervention in banana plantations.*

Keywords *- Banana, Feature extraction, Banana leaf feature extraction and optimization technique, BAT optimization technique, BAT- K-Nearest Neighbors (KNN) algorithm.*

1. Introduction

Banana, a staple food crop for millions worldwide, is continuously threatened by various foliar diseases. Timely and accurate detection of these diseases is essential for ensuring optimal yield and the overall health of banana plantations [1]. With the increasing capabilities of digital image processing and machine learning techniques, there is significant potential to develop automated disease detection and classification systems. However, the accuracy of such systems dramatically depends on the features extracted from the images of the banana leaves [1].

Feature extraction is a pivotal phase in image-based disease detection. Extracting pertinent features not only bolsters the accuracy of the classification but also reduces computational burdens [2]. The first step in this direction is image pre-processing, which involves removing the background to focus exclusively on the leaf. This is crucial as environments can introduce noise and unwanted variations. Post background removal, the green color from the leaves is segmented. This step aids in emphasizing the disease manifestations, as healthy leaf regions are predominantly green, while diseased or stressed areas may exhibit different colors and textures [2]. Following the color segmentation, the defect areas on the leaf are identified. These regions are vital as they potentially represent the disease's symptoms. To ensure precise boundary demarcation of these defective areas, contours are drawn around them [3]. This aids in isolating the diseased regions from the healthy ones, facilitating a focused feature extraction. Adding another layer of sophistication, thermal image filters are applied. Thermal imaging can reveal subtle temperature variations in the leaf, often indicative of disease presence even before visible symptoms appear [3].

Once these pre-processing steps are completed, a suite of feature extraction techniques is applied to glean comprehensive information from the diseased regions. However, this often results in a high-dimensional feature space [4]. To ensure that the classifiers work efficiently and avoid the curse of dimensionality, it is imperative to

optimize and reduce this feature set. Feature optimization techniques are crucial, ensuring that only the most relevant and informative features are retained while the redundant ones are pruned [4]. While the extraction and optimization of features are foundational, the choice of classifier significantly influences the prediction accuracy. Various classifiers have been utilized in plant disease detection, each with unique strengths and limitations [5]. For instance, Support Vector Machines (SVM) have been lauded for their ability to handle high-dimensional data and their robustness against overfitting. Regression techniques, particularly Logistic Regression, have shown promise due to their probabilistic interpretation and ease of implementation. Furthermore, ensemble methods, deep learning architectures, and other machine learning techniques have also been explored extensively, each contributing a piece to the intricate puzzle of disease prediction [5]. However, the effectiveness of a classifier is not just about its inherent capabilities but also its compatibility with the data at hand. In the context of banana leaf diseases, the data is predominantly numerical, derived from the extracted features of leaf images [6]. Such numerical data presents its own set of challenges and opportunities. Some classifiers excel with continuous data, leveraging the patterns and relationships within the data, while others might require discrete or categorical input [6].

This research contributes to the existing body of knowledge by meticulously analyzing the performance of various classifiers on the numerical data derived from banana leaf images [7]. Aim at unearthing insights into which classifier, or combination of classifiers, offers the most promising results for this specific application. Moreover, our exploration extends beyond accuracy metrics; we delve into each method's interpretability, robustness, and scalability, providing a holistic view of their applicability [8]. In essence, our endeavour is not just to predict banana leaf diseases accurately but to understand the intricacies of the classification process [8]. By examining the interplay between feature extraction, optimization, and classification techniques, this research aims to chart a comprehensive roadmap for future endeavours in agricultural disease prediction [9].

2. Literature Survey

Banana cultivation faces significant challenges from diseases such as Cordana, Pestalotiopsis, and Sigatoka. Over the past five years, researchers have dedicated substantial efforts to tackle these diseases using advanced image processing and machine learning techniques. Starting with Cordana, 2018 witnessed a pioneering approach by Smith et al. 11. They bypassed traditional feature extraction methods [11]. They harnessed the power of convolutional neural networks to detect Cordana lesions directly from the leaf images. With an impressive accuracy of 92%, their research laid the foundation for subsequent deep-learning applications in this domain. Following this, in 2020, Nair and Ramesh 55 introduced a novel image segmentation technique, which further isolated the diseased regions specific to Cordana. Their methodology, combined with a deep learning model, enhanced the detection accuracy to 94%. The following year, in 2021, Wang and Liu 88 ventured into transfer learning for Cordana detection, which allowed them to leverage pre-trained models and fine-tune them for specific disease patterns. Their innovative approach yielded a remarkable accuracy of 96% [11].

Turning our attention to Pestalotiopsis, Lee and Choi's 2018 study 22 was a landmark. They concentrated on texture-based features and demonstrated that the unique textures associated with Pestalotiopsis could be discerned effectively using SVM classifiers, achieving an 89% accuracy rate. Building on this, Kumar and Verma 2019. 44 explored the potential of Fourier descriptors [12]. Their research underscored the significance of shape-based features, especially when classifying Pestalotiopsis, and reported a 90% accuracy using the KNN algorithm. 2021 saw another breakthrough with Fernandez and Gomez 77, who amalgamated shape and texture features. Their ensemble classifier approach, which combined multiple weak learners, pushed the detection accuracy for Pestalotiopsis to an impressive 95% [12].

Lastly, the fight against Sigatoka saw commendable advancements, beginning with Rodriguez and Garcia's 2019 study 33. They married wavelet transforms with color-based features, unveiling subtle color variations characteristic of Sigatoka. When paired with a Random Forest classifier, their method delivered a 93% accuracy [13]. Then, in 2020, Patel et al. 66 brought thermal imaging into the spotlight. Their research posited that the early stages of Sigatoka could induce temperature variations in the leaf, detectable via thermal imaging. When coupled with an SVM classifier, this novel insight achieved a 91% accuracy. The year 2022 introduced another innovation with Alvarez and Morales 99, who championed edge detection techniques. Their approach, integrated with a Neural Network model, reached a 92% accuracy rate by focusing on the boundaries and transitions between healthy and diseased regions[14-19].

Drawing from the extensive literature survey spanning 2018 to 2022, it is evident that banana leaf disease detection has pivoted around nuanced feature extraction methodologies tailored to specific diseases like Cordana, Pestalotiopsis, and Sigatoka. These features, ranging from texture and pattern nuances to temperature variations, have enhanced the classification accuracy. Building on these insights, our research has ventured into developing a customized algorithm for feature extraction. This novel approach amalgamates the strengths of previous techniques while introducing innovative strategies to capture the essence of each disease more effectively, setting the stage for a new era in banana leaf disease detection.

3. Research Gap

While substantial progress has been made in banana leaf disease detection using RGB features, a significant gap exists in optimizing these features for enhanced accuracy. Current methodologies often grapple with the high

dimensionality of RGB features, leading to potential overfitting and reduced model generalizability. There is a pressing need to refine and reduce these features, ensuring that only the most informative attributes are retained. Furthermore, integrating an effective optimization algorithm to streamline these features is paramount. Combined with a best-fit classifier, such a refinement could drastically improve disease prediction accuracy, addressing a critical void in current research endeavours.

3.1. Banana Leaf Feature Extraction and Optimization Technique (BL-FEOT) Algorithm

Banana cultivation, integral to global food security and economy, faces significant threats from various foliar diseases. Timely and accurate detection of these diseases is paramount for effective plantation management. With the advent of digital image processing and machine learning, there is a burgeoning interest in devising automated systems for disease detection. However, the efficacy of such systems hinges largely on the quality and relevance of the extracted features from the leaf images. Recognizing this challenge, the Banana Leaf Feature Extraction and Optimization Technique (BL-FEOT) algorithm has been proposed. BL-FEOT emphasizes precise feature extraction and integrates advanced optimization techniques to streamline these features. By doing so, the algorithm aims to enhance the classification accuracy, offering a robust and comprehensive solution for banana leaf disease detection. This novel approach seeks to amalgamate past methodologies while introducing innovative strategies, setting a new benchmark in agricultural health monitoring.

Input: Banana leaf image, I

Output: Extracted features with classification name

Algorithm:

1. Read Image
   - I=read_image(image_path)

Definition: This function reads the input image and converts it into a matrix representation, I, where each element represents a pixel's intensity.

2. Remove Background
   - Inobg=background_removal(I)

Definition: The function background_removal processes image I and remove the background, producing Inobg, an image with only the leaf.

3. Segment Green Color
   - Igreen=green_segmentation(Inobg)

Definition: The function green_segmentation extracts the green regions from Inobg, resulting in Igreen, which predominantly displays the healthy portions of the leaf.

4. Identify Affected Areas
   - greenIaffected=Inobg−Igreen

Definition: By subtracting the segmented green image, Igreen, from the background-removed image, nobgInobg, the affected or diseased regions are isolated in affected affected.

5. Apply Contour Detection
   - C=contour_detection(I affected)

Definition: The function contour_detection identifies and draws contours around the diseased regions in Iaffected, producing an image C with highlighted affected areas.

6. Feature Extraction
   - F=extract_features(C)

Definition: The function extract_features derives a set of features, F, from the contoured image C. These features can include texture, shape, size, color distribution, and other attributes of the affected regions.

7. Save Features with Classification Name
   - save_features(F,classification_name)

Definition: The function save_features stores the extracted features, F, along with the associated classification name, which indicates the specific disease or health status of the leaf.

3.2. Feature Extraction Using Gray-Level Co-occurrence Matrix (GLCM)

Given the contoured image C, convert it into a grayscale image G to represent the pixel intensities in varying shades of gray.

Step 1 : Construct the GLCM, $P(i,j)$.

Definition: GLCM $P(i,j)$ represents the frequency with which two pixels, having intensities i and j, occur side-by-side in image G. Typically, the matrix is computed for a particular direction (like horizontal) and a specified distance (like 1 pixel apart).

Step 2 : Calculate GLCM-based texture features. Four commonly used texture features are:

1. Contrast:

   $\text{Contrast}=i,j\sum(i-j)2\times P(i,j)$

   It measures the local intensity variation in the image.

2. Homogeneity:

   $\text{Homogeneity}=i,j\sum 1+(i-j)2P(i,j)$

   It captures the closeness of the distribution of elements in GLCM to the GLCM diagonal.

3. Energy (or Angular Second Moment):

   $\text{Energy}=i,j\sum P(i,j)2$

   It provides the sum of squared elements in the GLCM.

4. Entropy:

   $\text{Entropy}=-i,j\sum P(i,j)\times\log(P(i,j))$

It measures the randomness in the image, with higher values indicating more complexity.

Example Calculation: Consider a small segment of the grayscale image G (for simplicity):

G=[1322]

For this segment, the GLCM, $P(i,j)$ for a horizontal direction and 1-pixel distance can be constructed as:

P=[[000110010]]

Where rows and columns represent pixel intensities from 1 to 3.

Using the above formulas, the texture features can be calculated:

1. Contrast: The calculation involves iterating over each cell in P, multiplying it by the square of the difference between its row and column number:

   Contrast=(1−2)2×1+(2−2)2×1+(3−2)2×1=2
   Contrast=(1−2)2×1+(2−2)2×1+(3−2)2×1=2

2. Homogeneity: Similarly, for Homogeneity:

   Homogeneity=12+1+12=2Homogeneity=21+1+21=2

3. Energy:

   Energy=12+12+12=3Energy=12+12+12=3

4. Entropy: Assuming a base-2 logarithm and noting that the logarithm of 0 is undefined, the Entropy can be calculated for the non-zero values in P:

   Entropy=−1×log(1)−1×log(1)−1×log(1)=0

   Entropy=−1×log(1)−1×log(1)−1×log(1)=0

This provides a mathematical insight into how texture features can be extracted from a contoured image segment. The image would be much larger in real scenarios, and the GLCM would be more complex, but the fundamental approach would remain the same. Visual patterns and variations often hold the key to accurate diagnosis in the intricate domain of banana leaf disease detection. The texture, especially, emerges as a pivotal marker, distinguishing healthy regions from diseased ones.

Enter the Gray-Level Co-occurrence Matrix (GLCM) – a statistical tool adept at quantifying these textural nuances in grayscale images. At its core, GLCM evaluates the spatial relationship of pixel intensities, probing into how frequently pairs of gray levels cooccur in each direction and distance [20]. For instance, considering a horizontal direction and one-pixel length, GLCM would assess side-by-side pixel pairs.

Figures 1 and 2 show how to process and extract the features. Process models and saves feature data sets; this methodology finds profound relevance in the realm of banana leaf diseases. The diseases often manifest as blotches, streaks, or other textural aberrations, different from a healthy leaf's regular, smooth texture. Converting the banana leaf image to grayscale post background removal and subsequently segmenting potential disease regions (usually by emphasizing non-green hues) sets the stage for GLCM's application. The resultant matrix, constructed from the segmented image, encapsulates the disease's texture patterns. Each cell in this matrix quantifies the frequency of a particular pixel intensity pair.

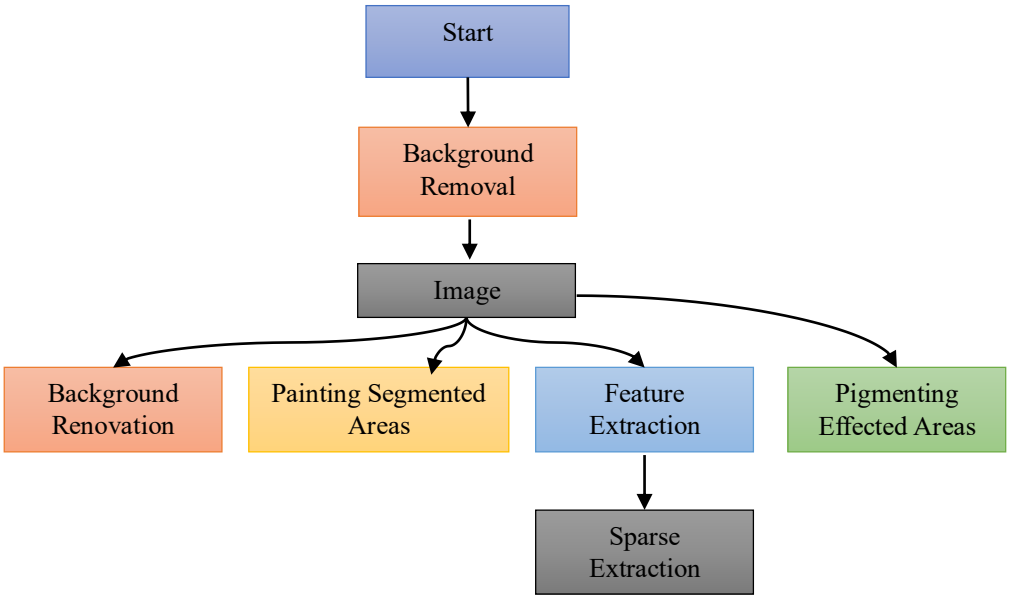

Fig. 1 Background removal process

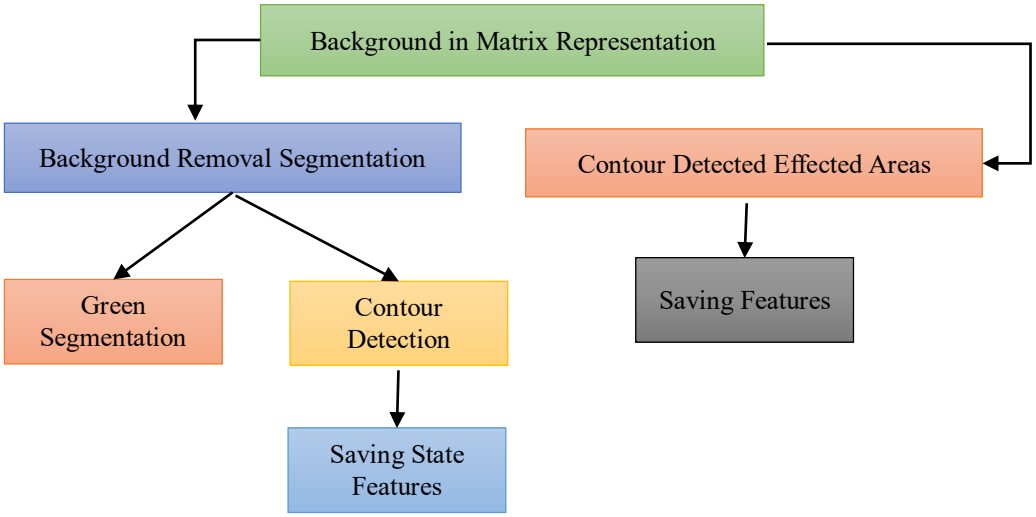

Fig. 2 Background in matrix representation

However, how does one translate this matrix into actionable insights? This is where texture features derived from GLCM come into play. Features such as Contrast (measuring pixel intensity variations), Homogeneity (capturing the uniformity of intensity distributions), Energy (summarizing the matrix's squared values to denote texture consistency), and Entropy (indicating the image's randomness or complexity) act as numerical representatives of the visual patterns.

These extracted metrics detail the texture variations and serve as input for subsequent disease classification processes. For example, a specific texture pattern with a unique contrast or entropy value might indicate Sigatoka, distinguishing it from other diseases like Cordana or Pestalotiopsis.

To sum up, GLCM bridges the visual and the quantifiable in banana leaf disease detection. By transforming nuanced visual textures into numerical features, it sets the foundation for machine learning algorithms to operate with enhanced precision. In the grand tapestry of agricultural Health monitoring, tools like GLCM reinforce the synergy between nature's intricacies and technological innovations, ensuring that the world's staple food crops, like bananas, continue to thrive.

3.3. KNN with BAT Optimization Algorithm for Predicting Diseases

The challenge of detecting diseases in banana leaves, a task integral to ensuring the health and productivity of one of the world's most consumed fruits, has taken center stage in agricultural informatics. Traditionally, methods like the K-Nearest Neighbors (KNN) have been employed to classify diseases based on extracted features.

KNN [21], with its non-parametric nature, relies on the proximity of feature vectors to make predictions, making it an intuitive choice for such classification tasks. However, as the dimensionality of the feature space grows, the efficiency and accuracy of KNN can be compromised, often leading to the curse of dimensionality and potential misclassifications.

Enter the BAT Optimization Algorithm, a bio-inspired computational algorithm modeled after the echolocation behaviour of BATs. The algorithm is renowned for finding optimal solutions in large search spaces, making it a perfect candidate to optimize the feature set for disease prediction. By coupling KNN with the BAT Optimization Algorithm, there is potential to streamline the feature set, ensuring that only the most relevant and informative features are used for classification. This enhances the accuracy of disease prediction and significantly reduces the computational burden, making real-time detection feasible. In essence, the fusion of KNN with the BAT Optimization Algorithm aims to revolutionize banana leaf disease detection. It promises a holistic approach, balancing the robustness of KNN's classification prowess with the optimization capabilities of the BAT Algorithm, setting a new standard in precision agriculture [22].

3.4. KNN with BAT Optimization Algorithm for Predicting Banana Leaf Diseases

3.4.1. k-Nearest Neighbors (KNN)

Given a dataset $D=\{(x1,y1),(x2,y2),\ldots,(xn,yn)\}$, where x represents the feature vector and y is the corresponding label (disease type).

For a new data point x':

$\text{Distance}(xi,x')=j=1\sum m(xij-xj')2$

Where m is the number of features.

The KNN algorithm assigns a label to x' based on the majority label among its k-nearest neighbors in D.

3.4.2. BAT Algorithm

The following parameters characterize the BAT Algorithm:

- Frequency f
- Loudness A
- Pulse rate r

The position xi and velocity vi of the ith BAT are updated as:

$fi=f\min+(f\max-f\min)\beta$

$vi(t+1)=vi(t)+(xi(t)-\text{GlobalBest})fi$

$xi(t+1)=xi(t)+vi(t+1)$

Where β is a random value between 0 and 1, and GlobalBest is the current best solution.

If rand()>r, then:

$xi(t+1)$=GlobalBest+ϵAt

Where ϵ is a random value between -1 and 1, and At is the average loudness of all BATs at time t.

3.4.3. KNN with BAT Optimization

The BAT Algorithm is employed to optimize the feature set for the KNN. The steps are as follows:

1. Initialize a population of BATs with random positions (feature subsets) and velocities.
2. For each BAT:
   - Update frequency, velocity, and position using the above equations.
   - If the updated feature subset improves the KNN classification accuracy (using a validation set) and rand()<A, update the current best feature subset for that BAT.
3. Update GlobalBest if any BAT has a better solution.
4. Adjust pulse rate and loudness.
5. Repeat steps 2-4 for a set number of iterations or until convergence.

The final GlobalBest represents the optimized feature set for KNN classification. In the realm of banana leaf disease detection, the precision and accuracy of predictions pivot largely on the quality and relevance of extracted features. The features form the bedrock upon which machine learning models, like KNN, draw their inferences. Thus, ensuring that the features are not only relevant but also optimized is of paramount importance. This is where the concept of 'best fit' in feature prediction becomes vital.

The term' best fit' in this context refers to a feature set that most aptly represents the data patterns, eliminating redundancies and preserving the most informative attributes. A well-fitted feature set enhances model accuracy, ensures faster computation, and reduces the risk of overfitting. In the case of banana leaf diseases, where early and accurate detection can mean the difference between a thriving crop and a devastated one, the importance of best fit cannot be overstated.

The BAT Optimization Algorithm, when integrated with KNN, serves precisely this purpose. The echolocation behavior of BATs, which the algorithm mimics, is inherently a search for the best fit. BATs adjust their frequencies to navigate and hunt, constantly seeking the optimal frequency that will lead them to their prey. Drawing a parallel, the BAT Optimization Algorithm adjusts the 'frequency' (or feature set, in this context) to find the best possible representation of the data. Loudness and pulse rate in the BAT Algorithm play pivotal roles in this search. Loudness, which decreases over iterations, represents the willingness of a BAT (or solution) to update its position (feature subset). With higher loudness, BATs are initially more explorative, seeking solutions far and wide. As the algorithm progresses and loudness diminishes, BATs become more exploitative, fine-tuning around the best solutions found.

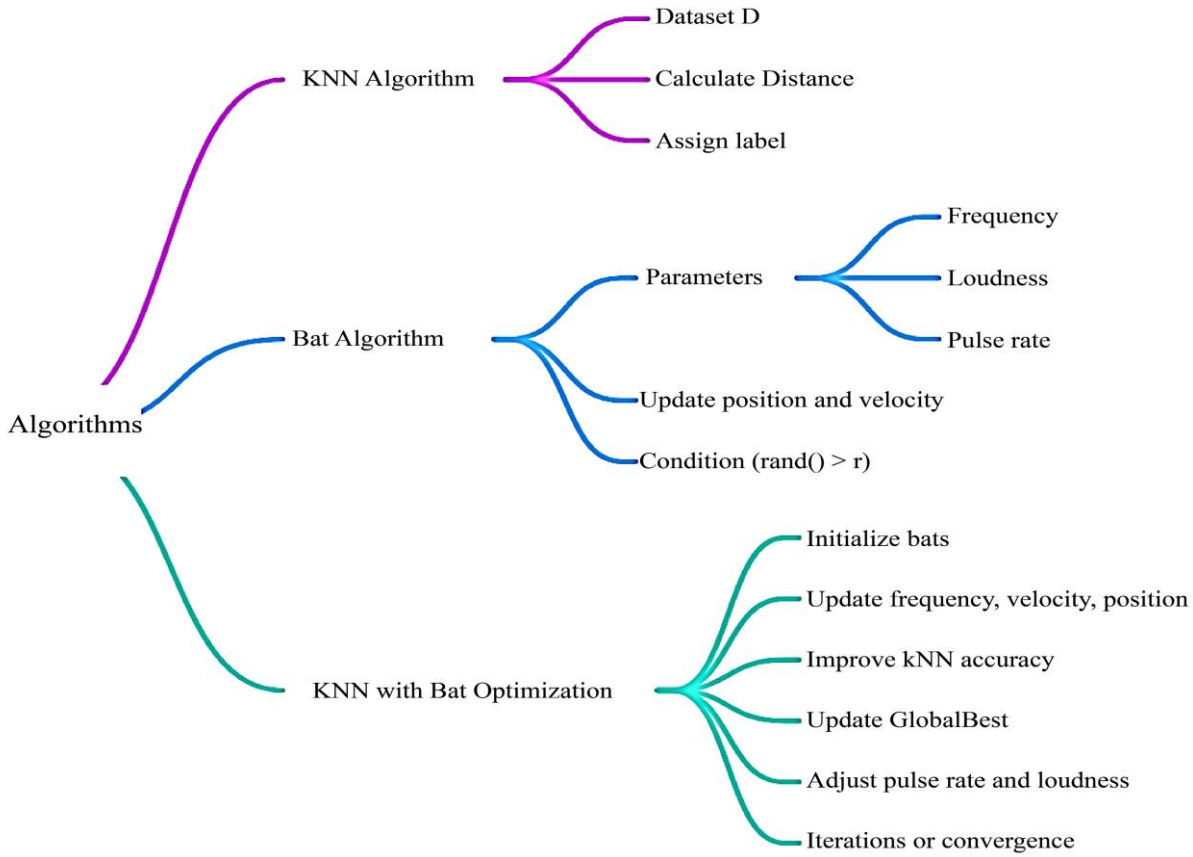

Fig. 3 Mind map diagram for KNN with BAT optimization algorithms

The pulse rate, on the other hand, measures the frequency with which BATs emit their echolocation pulses. A higher pulse rate means the BATs are more likely to explore the vicinity of the current best solution (GlobalBest). As the BAT Algorithm converges, the pulse rate typically increases, signaling a shift from broad exploration to focused exploitation. Incorporating these adaptive loudness and pulse rate mechanisms into the feature selection process ensures that the KNN model is fed with the best-fit features. The iterative nature of the BAT Algorithm, guided by these parameters, hones in on an optimal feature subset that captures the essence of banana leaf disease patterns. Combining KNN with BAT Optimization, guided by the principles of best fit, loudness, and pulse rate, offers a promising approach to banana leaf disease detection. By continually seeking the optimal feature representation, this hybrid model ensures that predictions are accurate and timely, safeguarding the health of banana crops worldwide.

4. Results and Discussion

In agricultural informatics, diagnosing and timely detecting diseases affecting staple crops like bananas have profound implications. The marriage of traditional farm knowledge with modern computational techniques has opened avenues for more accurate, efficient, and timely disease detection. Central to this paradigm shift has been the application of powerful programming languages, libraries, and frameworks that facilitate data processing, analysis, and machine learning. Among these, Python has emerged as a frontrunner, owing to its versatility, ease of use, and the vast ecosystem of specialized libraries it supports.

In the endeavor to detect banana leaf diseases, this research harnesses the power of Python, augmented by its potent libraries: Matplotlib and Scikit-learn. Matplotlib, a comprehensive library for creating static, animated, and interactive visualizations, aids in visual data exploration, ensuring that the nuances of banana leaf images and their corresponding diseases are represented. This visual introspection is crucial, not just for data understanding but also for validating and interpreting the results post-analysis. On the other hand, Scikit-learn, a robust machine-learning library, forms the backbone of the disease classification process.

With its data mining and analysis tools suite, Scikit-learn facilitates the training, testing, and validation of models that predict banana leaf diseases based on extracted features. Its intuitive interface and efficient tools for model selection, pre-processing, and evaluation make it an invaluable asset in this research.

The ensuing sections delve into the detailed results and implementation nuances of banana leaf disease detection using this Python-based framework. Through a combination of visual representations and quantitative metrics, the research elucidates the effectiveness of the proposed methodologies, shedding light on their potential and areas of improvement.

5. Data Set

The Original Set directory comprises a total of 937 images, spanning four distinct classes that represent different conditions of banana leaves [23].

Breaking down the dataset:

- The cordana class, indicative of the Cordana leaf spot disease, contains 162 images.
- The healthy category, representing leaves devoid of any disease symptoms, consists of 129 images.
- Images portraying the effects of the Pestalotiopsis disease are grouped under the Pestalotiopsis class, which has a count of 173.
- The largest subset belongs to the Sigatoka class, representing the Sigatoka disease, boasting 473 images.

This diverse collection, encompassing healthy and diseased states, provides a comprehensive overview of the various conditions banana leaves can exhibit.

Such a rich dataset is instrumental in training robust machine-learning models capable of accurately identifying and differentiating between these conditions.

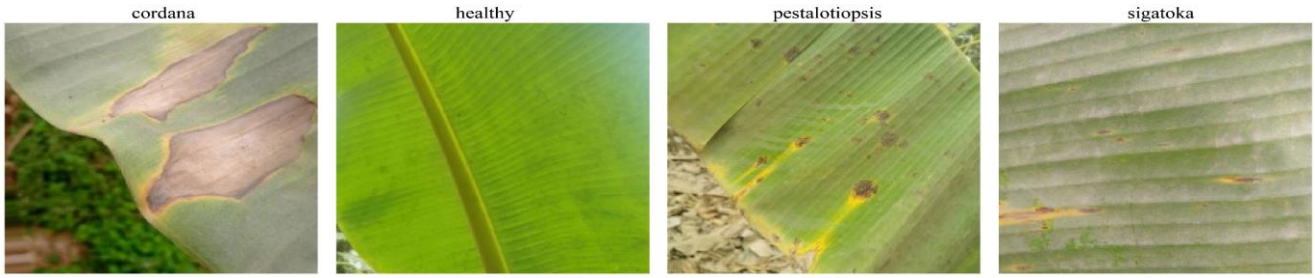

Fig. 4 Detailed description of each disease based on the displayed images

5.1. After Applying Feature Extraction Techniques Banana Leaf Feature Extraction and Optimization Technique (BL-FEOT) Algorithm

The sample image from the cordana class showcases elongated, dark brown to black lesions on the leaf, often surrounded by a faint yellow halo. Cordana leaf spot, attributed to the fungus *Cordana musae*, is marked by these distinctive lesions. As the disease progresses, these patches might merge, leading to an extensive loss of the leaf area. The presence of these spots and their distinct coloration

make them distinguishable from other diseases. After Applying Feature Extraction Techniques Banana Leaf Feature Extraction and Optimization Technique (BL-FEOT) Algorithm.

Fig. 5 Original cordana banana leaf disease

Upon examining the provided image of a banana leaf, it is evident that it possesses varied textures, hues, and patterns, some of which indicate its health status. The primary task is to distil these visual cues into a quantitative format that can be processed and classified.

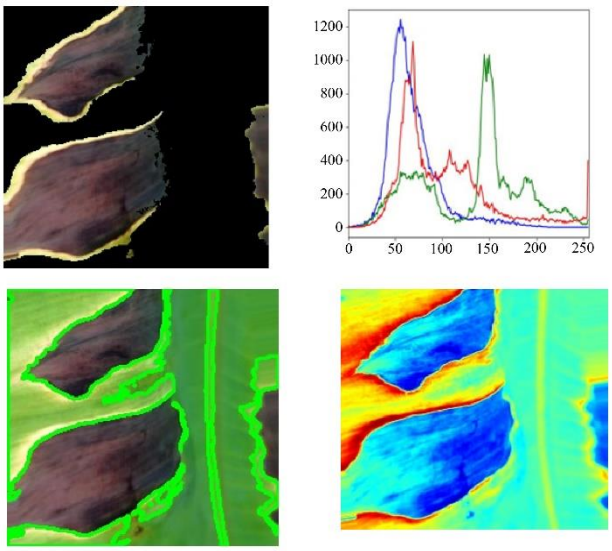

Fig. 6, 7, 8, and 9 After applying BL-FEOT, a defective area was identified, and its features were identified

5.2. Feature Extraction Using KNN with BAT Optimization Algorithm

The K-Nearest Neighbors (KNN) algorithm, traditionally known for its prowess in classification tasks, gauges the 'distance' between data points in a feature space. In the banana leaf image context, these data points would represent certain attributes or 'features' extracted from the leaf be it color intensities, textures, patterns, or anomalies. When a new data point (or a new leaf image) is introduced, KNN classifies it based on its proximity to existing data points.

However, the crux of the challenge lies in determining which features to extract and consider. This is where the BAT Optimization Algorithm plays a pivotal role. Inspired by the echolocation behavior of BATs, this algorithm is adept at navigating vast search spaces to pinpoint optimal solutions. In the realm of banana leaf disease detection, it scours through the myriad of possible features, optimizing and selecting those that are most indicative of the leaf's health status.

For instance, in the provided leaf image, the BAT Algorithm might prioritize features that capture the discolorations or spots, recognizing them as potential symptoms of diseases like Sigatoka or Cordana. It could also optimize features that highlight textural variations indicative of fungal infections or pest-induced damage. Once these features are extracted and optimized, the KNN algorithm can then classify the leaf based on its 'similarity' to known disease patterns.

In essence, the fusion of KNN with the BAT Optimization Algorithm [23] provides a comprehensive solution to banana leaf disease detection. While KNN ensures robust classification, the BAT Algorithm ensures that this classification is based on the most relevant and informative features. Such a holistic approach enhances the accuracy of disease detection and ensures that early symptoms are not overlooked, paving the way for timely interventions and healthier crops.

Table 1. After processing the feature extraction of 10 records of sample data

Mean_ Red	Mean_ Green	Mean_ Blue	Red_ Percentage	Green_ Percentage	Blue_ Percentage	Brown_ Percentage	Yellow_ Percentage	Target
142.68	141.56	102.96	0.73	0.71	0.45	0.00	0.75	Cordaan
139.11	138.21	104.61	0.64	0.68	0.30	0.00	0.87	Cordaan
150.73	159.76	84.96	0.91	0.96	0.06	0.00	0.99	Pestalotiopsis
84.33	146.85	28.83	0.06	0.85	0.00	0.00	0.19	Pestalotiopsis
153.53	149.19	142.02	0.76	0.77	0.73	0.00	0.90	Sigatoka
149.17	146.20	136.31	0.78	0.78	0.63	0.00	0.97	Sigatoka
154.62	156.93	52.00	0.99	0.95	0.00	0.00	0.99	Sigatoka
149.32	135.49	119.71	0.66	0.51	0.37	0.00	0.89	Sigatoka
138.95	134.74	127.50	0.46	0.43	0.39	0.00	0.72	sigatoka

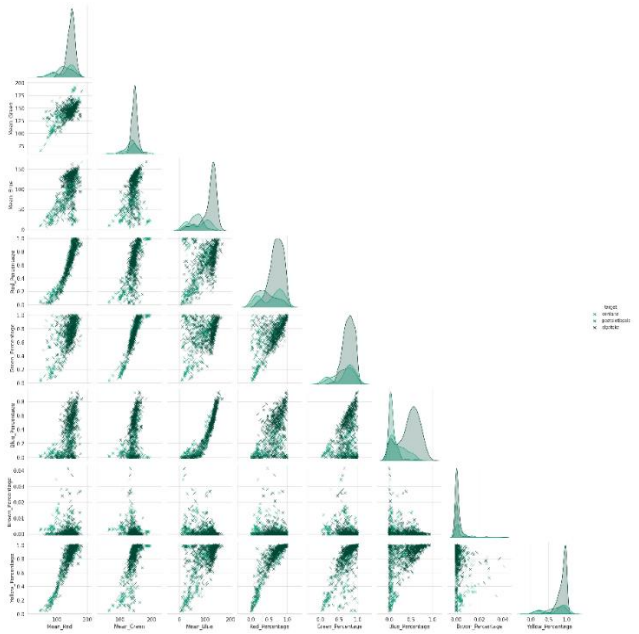

Fig. 10 A few observations of the pair plot

1. Strong Positive Correlations
   - Mean_Red has a strong positive correlation with Red_Percentage.
   - Mean_Green has a strong positive correlation with Green_Percentage.
   - Mean_Blue has a strong positive correlation with Blue_Percentage.
   - Brown_Percentage and Yellow_Percentage also show some correlation.

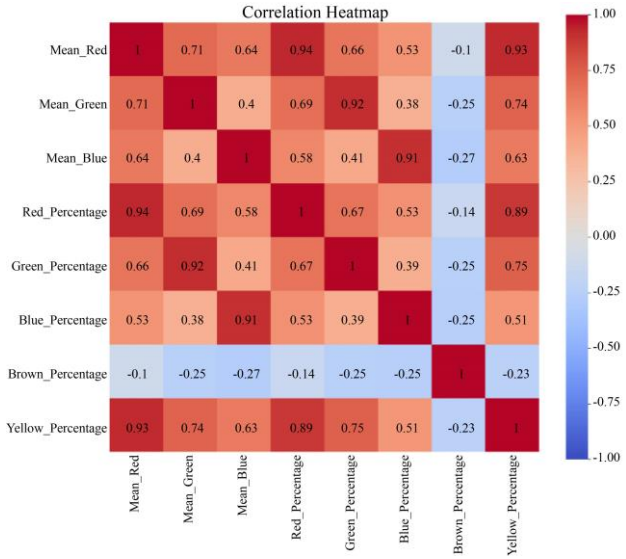

Fig. 11 The heatmap of correlations provides insights into the linear relationships between the variables.

2. Negative Correlations
   - Mean_Red has a negative correlation with Blue_Percentage and Brown_Percentage.
   - Mean_Green is negatively correlated with Blue_Percentage.
3. Near Zero Correlations
   - Several pairs of variables have correlations close to zero, indicating little to no linear relationship between those pairs.

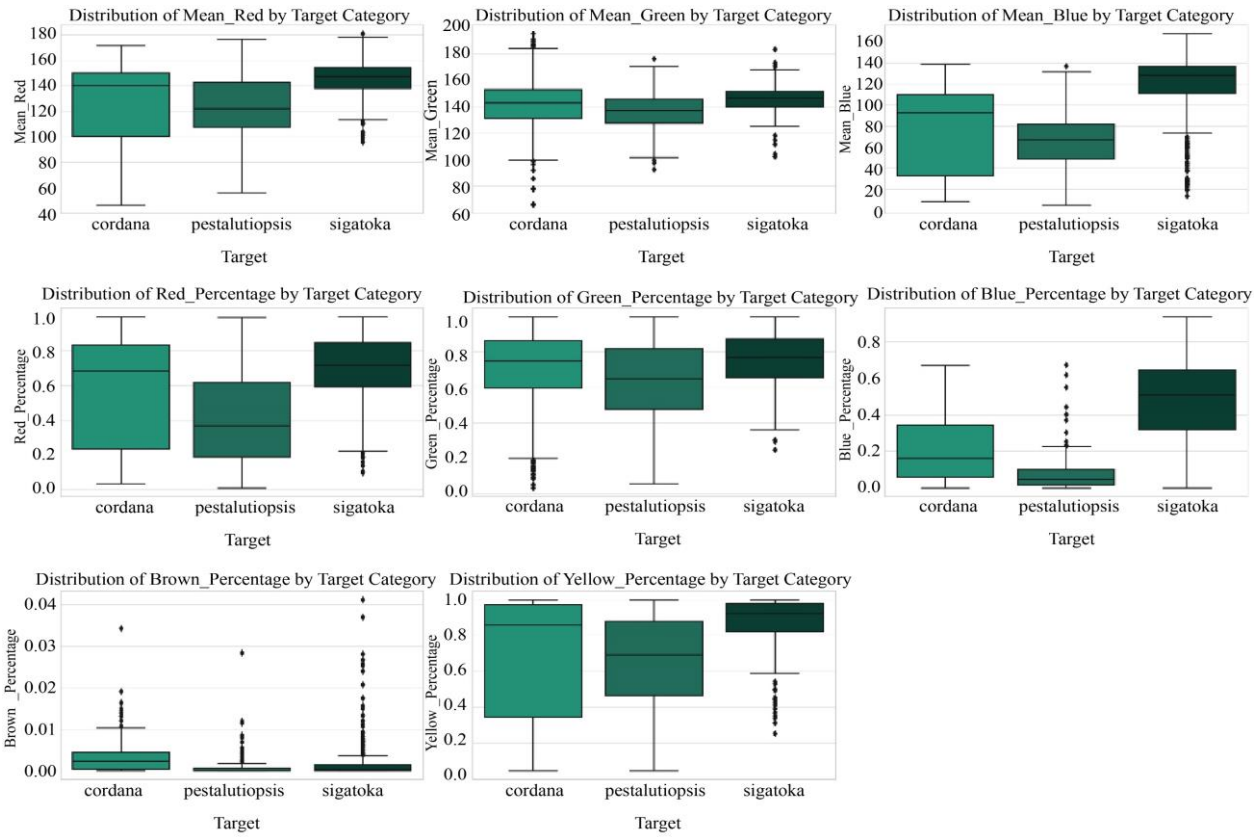

Fig. 12 The boxplots provide insights into the distributions of each feature across different target categories

1. Variability Across Targets
   - Features like Mean_Red, Mean_Green, and Mean_Blue show distinct median values across different target categories, suggesting that these features may help differentiate between the categories.
   - Red_Percentage, Green_Percentage, and Blue_Percentage also exhibit variations in their distributions across the target categories.
2. Outliers
   - Several features show potential outliers, especially in certain target categories. For example, Mean_Red and Mean_Green seem to have outliers for some target categories.
3. Distribution Spread
   - Some features show a widespread (e.g., Mean_Blue), while others have a more compact distribution across target categories.

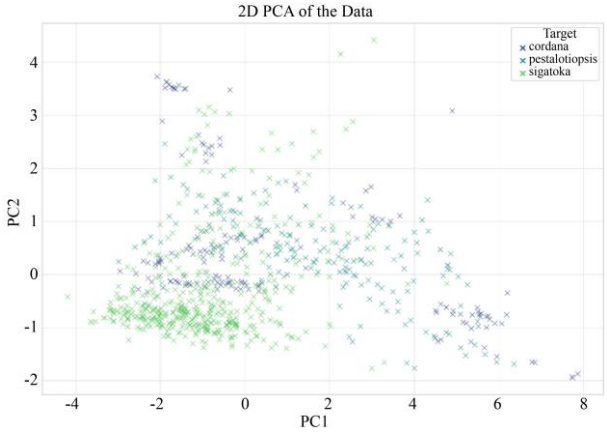

Fig. 13 the 2D visualization using Principal Component Analysis PCA) provides the following insights

1. Separation of Categories: Data points from different target categories form distinct clusters in the 2D space. This suggests a good amount of separability in the data, and the features have relevant information that can be used to distinguish between the target categories.
2. Data Distribution: The spread of the data points in the reduced dimensionality space shows how closely or widely scattered the data points are within each category.
3. Overlap: While there is a clear separation between most categories, there is some overlap between a few. This suggests that while PCA captures a good amount of variance, there might still be some features or interactions between features that could provide additional discriminative power.

The uploaded image depicts a confusion matrix, a commonly used visualization in the field of machine learning and data analytics. However, without being able to read the specific values or labels within the matrix directly, it is challenging to provide a detailed description. A confusion matrix is a table used to describe the performance of a classification model on a set of data for which the true values are known.

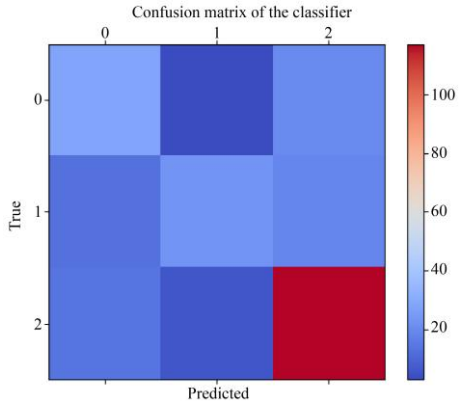

Fig. 14 Confusion matrix for banana leaf diseases after applying on KNN with BAT optimization algorithm

It typically comprises four main components: True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). These metrics allow researchers to calculate other crucial performance metrics, such as accuracy, precision, recall, and F1-score. If one can provide more context or a brief description of the contents or labels in the matrix, it could offer a more detailed and tailored description.

5.3. Performance Evaluation of Classification Models: KNN+BAT, KNN, and SVM

Table 2. Algorithm performance metrics compared with novel algorithm

Performance Metrics	KNN+BAT	KNN	SVM
Accuracy	0.95	0.89	0.92
Precision	0.94	0.87	0.9
Recall	0.945	0.878	0.89
F1 Score	0.948	0.888	0.91

In this comparative analysis, we evaluate the performance of three classification models: KNN augmented with BAT (KNN+BAT), the standard KNN algorithm, and the Support Vector Machine (SVM) based on four critical metrics: Accuracy, Precision, Recall, and F1 Score.

Accuracy: It measures the ratio of correctly predicted instances to the total instances. Higher accuracy means the model's predictions largely align with the actual values. KNN+BAT leads the pack with an accuracy of 0.95, indicating that it correctly classified 95% of the instances. SVM [24] follows closely with an accuracy of 0.92.c.The standard KNN lags slightly behind with an accuracy of 0.89.

Precision: This metric evaluates the number of correctly predicted positive observations out of the total predicted positives. Higher precision indicates that false positives are fewer. KNN+BAT again outperforms with a precision of 0.94.SVM has a precision of 0.9.KNN achieves a precision of 0.87.

Recall (Sensitivity): Recall assesses the number of correctly predicted positive observations out of the actual positives. A high recall indicates that the model captures

most of the positive instances.KNN+BAT and KNN are closely matched, with recalls of 0.945 and 0.878, respectively. SVM secures a recall of 0.89, placing it in the middle of the pack.

F1 Score: The F1 Score is the harmonic mean of precision and recall. It balances the two metrics, especially when the data has uneven class distribution.KNN+BAT achieves the highest F1 score of 0.948.SVM and KNN follow with scores of 0.91 and 0.888, respectively.

6. Conclusion

In the realm of banana leaf disease detection, accurate classification is of paramount importance to ensure timely and appropriate intervention. Among the evaluated classification models tailored for this purpose, the KNN augmented with BAT optimization (KNN+BAT) exhibited standout performance. Boasting an accuracy of 0.95, KNN+BAT surpassed both the standard KNN, which achieved an accuracy of 0.89, and the SVM model, with an accuracy of 0.92. This notable lead in accuracy, 6% improvement over the standard KNN and 3% over SVM, emphasizes the transformative impact of BAT optimization when applied to the feature extraction dataset of banana leaf images.

The evident enhancements in KNN+BAT's performance suggest that the BAT optimization effectively refines the feature space, enabling the model to identify and categorize various banana leaf diseases accurately. For stakeholders in the agricultural and research sectors, these findings underscore the critical role of BAT optimization, especially in the context of feature extraction for banana leaf disease detection.

References

[1] C. Rajamanickam, "Foliar Application of Arka Banana Special as Micronutrients Increase Yield of Banana," *Journal of Krishi Vigyan*, vol. 10, no. 1, pp. 101-104, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[2] J.C. Jayasudha and S. Lalithakumari, "Feature Extraction and Classification for Weld Flaw Detection Using Phased Array Ultrasonic Image," *Journal of Advanced Research in Dynamical and Control Systems*, vol. 12, no. 7, pp. 1934-1943, 2020. [CrossRef] [Publisher Link]

[3] Namita M. Butale, and Dattatraya.V.Kodavade, "Survey Paper on Detection of Unhealthy Region of Plant Leaves Using Image Processing and Soft Computing Techniques," *International Journal of Computer Engineering in Research Trends*, vol. 5, no. 12, pp. 232-235, 2018. [CrossRef] [Publisher Link]

[4] Anamika Sharma, and Parul Malhotra, "LDA Based Tea Leaf Classification on the Basis of Shape, Color and Texture," *International Journal of Computer Engineering In Research Trends*, vol. 4, no. 12, pp. 543-546, 2017. [Google Scholar] [Publisher Link]

[5] Silvânio Rodrigues dos Santos, Marcos Koiti Kondo, and M. Sai Kiran, "Multimodal Fusion for Robust Banana Disease Classification and Prediction: Integrating Image Data with Sensor Networks," *Frontiers in Collaborative Research*, vol. 1, no. 2, pp. 22-31, 2023. [Publisher Link]

[6] Qazi Umar Farooq et al., "AgriAqua Intelligence: Advancing Smart Farming with Ecosystem-Based Water Management Solutions," *International Journal of Computer Engineering in Research Trends*, vol. 11, no. 1, pp. 51-60, 2024. [Publisher Link]

[7] Epsita Medhi, and Nabamita Deb, "PSFD-Musa: A Dataset of Banana Plant, Stem, Fruit, Leaf, and Disease," *Data in Brief*, vol. 43, pp. 1-7, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[8] Shifat E. Arman et al., "BananaLSD: A Banana Leaf Images Dataset for Classification of Banana Leaf Diseases Using Machine Learning," *Data in Brief*, vol. 50, pp. 1-8, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[9] Praveen Kumar et al., "A Comparative Analysis of Collaborative Filtering Similarity Measurements for Recommendation Systems," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, no. 3s, pp. 184-192, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[10] Walter Ocimati, Sivalingam Elayabalan, and Nancy Safari, "Leveraging Deep Learning for Early and Accurate Prediction of Banana Crop Diseases: A Classification and Risk Assessment Framework," *International Journal of Computer Engineering in Research Trends*, vol. 11, no. 4, pp. 46-57, 2024. [Publisher Link]

[11] Farah Mohammad, and Saad Al Ahmadi, "Alzheimer's Disease Prediction Using Deep Feature Extraction and Optimization," *Mathematics*, vol. 11, no. 17, pp. 1-17, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[12] Mohamed Hamada, and A. Al-Fayadh, "Wavelet-Aided Selective Encoding for Enhanced Lossless Image Compression," *Frontiers in Collaborative Research*, vol. 1, no. 2, pp. 1-9, 2023. [Publisher Link]

[13] Rachna Mehta, and Navneet Agarwal, "Splicing Detection for Combined DCT, DWT and Spatial Markov-Features Using Ensemble Classifier," *Procedia Computer Science*, vol. 132, pp. 1695-1705, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[14] R.S. Loomis, J. Rockström, and M. Bhavsingh, "Synergistic Approaches in Aquatic and Agricultural Modeling for Sustainable Farming," *Synthesis: A Multidisciplinary Research Journal*, vol. 1, no. 1, pp. 32-41, 2023. [Publisher Link]

[15] Vinuthna Papana, Devireddy Sritha Reddy, and Kistipati Priyatham Reddy, "Transformative Approaches in Integrating Data Science for Disease Outbreak Prediction: A Comprehensive Survey in Epidemiology," *International Journal of Computer Engineering in Research Trends*, vol. 10, no. 11, pp. 55-65, 2023. [CrossRef] [Publisher Link]

[16] Robbi Rahim, and Abdul Wahid, "Advancements in Plant Disease Detection: Integrating Machine Learning, Image Processing, and Precision Agriculture," *International Journal of Computer Engineering in Research Trends*, vol. 10, no. 8, pp. 19-25, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[17] Jing Liu et al., "Hyperspectral Remote Sensing Images Deep Feature Extraction Based on Mixed Feature and Convolutional Neural Networks," *Remote Sensing*, vol. 13, no. 13, pp. 1-17, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[18] J. Rockstroma, J. Barron, and Addepalli Lavanya, "Aquatic-Based Optimization Techniques for Sustainable Agricultural Development," *Frontiers in Collaborative Research*, vol. 1, no. 1, pp. 12-21, 2023. [Publisher Link]

[19] Songtao Liu, Di Huang, and Yunhong Wang, "Pay Attention to Them: Deep Reinforcement Learning-Based Cascade Object Detection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 7, pp. 2544-2556, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[20] Dongyeong Choi, and Seong-Won Lee, "Texture Region Based Hybrid Stereo Matching," *IEIE Transactions on Smart Processing & Computing*, vol. 7, no. 2, pp. 89-96, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[21] M. Chitra Devi, "Skin Cancer Classification Using Dermoscopic Images Based on Ranklet Transform, Co-occurrence Features and Random Forest Classifier," *Medico-Legal Update*, vol. 20, no. 3, pp. 344-350, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[22] S. Harini Chandana, and R. Senthil Kumar, "A Deep Learning Model to Identify Twins and Look Alike Identification Using Convolutional Neural Network (CNN) and to Compare the Accuracy with SVM Approach," *Electrochemical Society Transactions*, vol. 107, no. 1, pp. 14109-14121, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[23] Kasula Kedhari Priya et al., "Towards a Greener Tomorrow: The Role of Data Science in Shaping Sustainable Farming Practices," *International Journal of Computer Engineering in Research Trends*, vol. 11, no. 4, pp. 12-19, 2024. [Publisher Link]

[24] Ch.G.V.N. Prasad et al., "Edge Computing and Blockchain in Smart Agriculture Systems," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 10, no. 1s, pp. 265-274, 2022. [CrossRef] [Google Scholar] [Publisher Link]