

A Multi-District Field-Acquired Sugarcane Disease Image Dataset From Assam For AI-Based Crop Disease Detection

Ankita Goyal Agarwala^{1,2}, Mirzanur Rahman^{3*}, Jyotismita Das⁴, Yash Pareek⁴, Manash Choudhury⁴

¹Research Scholar, Department of IT, Gauhati University, Guwahati, Assam-781014

²Assistant Professor, Department of CSE, The Assam Royal Global University, Guwahati, Assam-781035

³Associate Professor, Department of IT, Gauhati University, Guwahati, Assam-781014

⁴B. Tech (CSE) 8th Semester, Department of CSE, The Assam Royal Global University, Guwahati, Assam-781035

*Corresponding author: mirzanurrahman@gmail.com

Abstract

Sugarcane is a major agricultural crop in Assam, India, where disease outbreaks significantly affect yield and farmer income. The development of automated disease detection systems requires region-specific, high-quality datasets. This study presents a curated dataset comprising 5,769 field-acquired sugarcane leaf images collected from multiple districts of Assam under natural environmental conditions. The dataset contains 5,769 field-acquired images categorized into nine classes representing eight disease types and healthy plants. Images were captured following standardized acquisition protocols and annotated through a structured labeling process. Quality control measures were applied to ensure clarity and class consistency. To evaluate dataset robustness, benchmark experiments were conducted using established convolutional neural network architectures. Performance metrics including accuracy, precision, recall, and F1-score are reported.

The dataset is made publicly available to support reproducible research and AI-driven crop disease monitoring systems tailored to humid agro-climatic regions.

Keywords Sugarcane disease; Image dataset; Deep learning; Precision agriculture; Assam.

1. Introduction

Sugarcane is one of the most important commercial crops globally and serves as a major cash crop in Assam, India, where it significantly contributes to the regional agricultural economy and associated agro-industries. A substantial proportion of farmers in Assam depend on sugarcane cultivation for their livelihood. However, crop productivity and quality are frequently compromised by various plant diseases, including mosaic, ring spot, leaf spot, red rot, smut, and rust. These diseases can spread rapidly under humid subtropical climatic conditions, resulting in considerable yield reduction and economic losses [1].

Timely and accurate disease detection is critical for effective crop management and minimizing financial impact. In recent years, machine learning and deep learning approaches have demonstrated significant potential for automated plant disease identification, often surpassing traditional image processing techniques [2]. Convolutional neural networks (CNNs), in particular, have shown strong performance in classifying plant diseases from leaf images under controlled conditions [3]. However, the generalization ability of such models depends heavily on the availability of high-quality and diverse training datasets.

Most existing plant disease datasets are developed at global or national scales and often fail to account for localized agro-climatic conditions, soil characteristics, and region-specific disease manifestations. Studies have shown that models trained on laboratory-based or homogeneous datasets may exhibit performance degradation when applied to real-world field conditions [4]. Consequently, there is a pressing need for region-specific, field-acquired datasets that capture natural variability in lighting, background, disease progression stages, and crop morphology.

To address this gap, the present study introduces a curated, multi-district image dataset specifically designed to represent sugarcane plant diseases prevalent in Assam. The dataset captures images under natural field conditions across diverse lighting environments, growth stages, and symptom severity levels. By incorporating variations in angles, backgrounds, and disease manifestations, the dataset aims to enhance robustness and real-world applicability for AI-based disease detection systems.

The development of a localized and publicly accessible dataset serves multiple stakeholders, including researchers, agricultural experts, policymakers, and farmers. It facilitates the design of reliable machine learning models for early disease diagnosis, supports precision agriculture initiatives, and contributes to sustainable crop management practices. Ultimately, this initiative aims to strengthen disease monitoring frameworks and improve sugarcane productivity in Assam, thereby supporting farmer livelihoods and regional agricultural sustainability.

Several plant disease datasets have been proposed in recent years for developing machine learning models for crop monitoring. However, many of these datasets are either captured under controlled laboratory conditions or represent crops from different geographical regions. Consequently, they often fail to capture the ecological variability and disease manifestations observed in the humid subtropical agro-climatic conditions of Northeast India. The proposed dataset addresses this gap by providing a field-acquired, multi-district sugarcane disease dataset from Assam, collected under real agricultural conditions. Compared with existing datasets, it incorporates diverse disease stages, natural lighting conditions, and geographically distributed sampling locations, making it particularly suitable for developing robust AI-based disease detection systems.

1.1. Contribution of this Work

This study introduces a novel field-acquired image dataset of sugarcane plant diseases collected from multiple districts of Assam, India. The key contributions of this work are summarized as follows:

- **Development of a region-specific dataset:** A curated dataset consisting of 5,769 high-resolution images representing nine classes (eight disease categories and one healthy class) was created from real sugarcane fields in Assam. To the best of our knowledge, this is the first open-access sugarcane disease dataset originating from this region.
- **Field-based data acquisition under natural conditions:** The images were captured using multiple camera devices under varying lighting conditions, plant growth stages, and disease severity levels. This ensures that the dataset reflects real-world agricultural scenarios, making it suitable for practical disease detection applications.
- **Balanced dataset with verified annotations:** The dataset maintains a balanced distribution of disease classes, and the labeling was performed based on visual

disease characteristics and further verified by experts from the Assam Agricultural University – Sugarcane, Medicinal and Aromatic Plant Research Station, Buralikson.

- **Comprehensive dataset validation:** Statistical analysis of pixel distributions, visualization-based inspection, and benchmark experiments using ResNet50, CSPDarkNet, and Vision Transformer models were conducted to evaluate the reliability and usability of the dataset for machine learning applications.
- **Support for AI-driven agricultural research:** The dataset is publicly released to support the development of machine learning models, intelligent disease monitoring systems, and precision agriculture tools aimed at improving sugarcane productivity and disease management.

2. Materials and Methods

This study presents the systematic development of a field-acquired image dataset of sugarcane plant diseases collected from multiple districts of Assam, India. The dataset was generated through structured field surveys conducted under natural environmental conditions to ensure real-world variability in lighting, background, and disease manifestation. Standardized protocols were followed for image acquisition, annotation, quality control, and data organization to ensure consistency, reliability, and suitability for artificial intelligence-based disease detection applications. The overall methodological framework is outlined in the following subsections.

2.1 Study Area

The dataset was collected from multiple sugarcane-growing locations across Assam, India, representing diverse agro-climatic and ecological conditions within the region. The selected field sites include Chukuniapara (25.9360683° N, 91.1423974° E), Dewbali (26.2516850° N, 92.5817496° E), Barman Bari (25.7471308° N, 93.1893323° E), Sontoli (26.1556492° N, 91.1282357° E), and Mahtali (25.5098481° N, 91.7979330° E). These locations span different parts of Assam and capture geographic variability across the Brahmaputra valley region.

Assam is characterized by a humid subtropical climate with high annual rainfall (approximately 1,500–3,000 mm), elevated humidity, and temperature variations ranging from about 10°C in winter to over 35°C in summer. Such environmental conditions create a favorable ecosystem for the development and rapid spread of fungal, viral, and bacterial pathogens affecting sugarcane crops. Variations in microclimate, soil type, irrigation practices, and cropping patterns across the selected locations contribute to differences in disease prevalence, symptom expression, and severity.

Images were acquired directly from open-field sugarcane plantations under natural lighting conditions and at different crop growth stages. By sampling across geographically distributed sites, the dataset captures ecological heterogeneity and real-world variability, enhancing its robustness for machine learning based disease detection systems. The inclusion of multiple locations improves model generalizability and supports the development of AI-driven diagnostic tools applicable across similar humid agroclimatic zones.

2.2 Data Acquisition

A comprehensive overview of the dataset creation workflow is illustrated in Fig. 1. The development process began with a preliminary field survey to identify active sugarcane

cultivation sites within the selected regions of Assam. Local stakeholders and farmers were consulted to obtain accurate information regarding field locations and disease prevalence. Based on multiple survey inputs, representative fields were selected as target sites for systematic data collection.

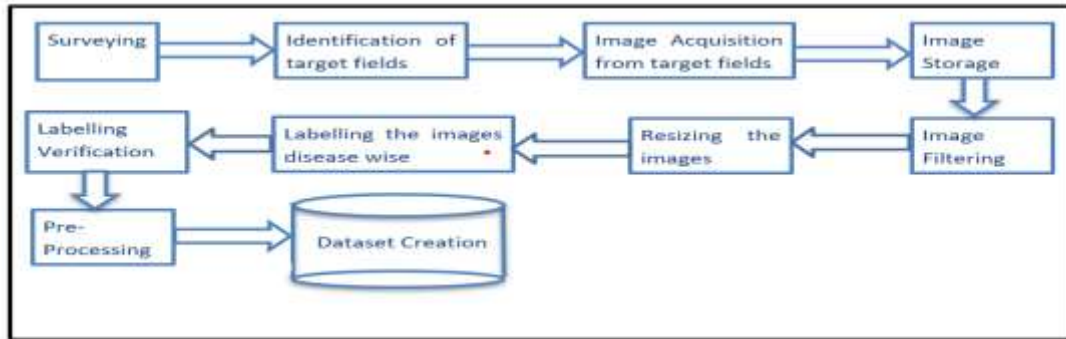


Fig. 1: Workflow of dataset creation

From the identified locations, a diverse set of images was captured under natural field conditions. The dataset comprises a total of 5,769 high-resolution images representing both infected and healthy sugarcane plants. The images were acquired using three high-end digital cameras: Panasonic DC-G95, Nikon Z 50, and Canon EOS 122D. Photographs were taken during daylight hours between October 2023 and May 2024, covering multiple crop growth stages and varying environmental conditions. Images were captured from different angles and distances to ensure clear visualization of symptomatic regions and to incorporate real-world variability in perspective and background.

Following acquisition, an image filtering process was conducted to remove blurred, noisy, or low-quality samples. Since images were captured using different camera devices, resolution and dimensions varied across samples. To ensure compatibility with deep learning frameworks, all filtered images were resized to a uniform dimension prior to model training. The resizing process preserved essential disease characteristics while standardizing input dimensions.

Manual annotation was performed based on visual symptom characteristics observed in the images. Each image was assigned to one of the nine predefined classes. The labeling criteria were based on established disease identification features reported in the literature [1]. To enhance reliability, the annotated dataset was externally verified by subject experts from the Assam Agricultural University: Sugarcane, Medicinal and Aromatic Plant Research Station, Buralikson, Assam (785622). This validation step ensured correctness and consistency in disease categorization.

The complete dataset occupies approximately 1.78 GB of storage and includes nine class labels, comprising eight disease categories namely grassy shoot, mosaic, red rot, red leaf spot, ring spot, wilt, yellow leaf, and pokkahboeng plus one healthy class. All images were stored in JPEG (.jpg) format. Representative samples of the dataset are shown in Fig. 2.

The structured pipeline from field survey and targeted acquisition to filtering, resizing, annotation, and verification ensures that the dataset is reliable, reproducible, and suitable for artificial intelligence-based disease detection applications.

2.3 Disease Categories

The Assam Sugarcane Disease Dataset consists of 5,769 high-resolution images categorized into nine distinct classes, representing eight major sugarcane diseases and one healthy plant class. The selected disease categories correspond to commonly occurring sugarcane diseases observed in the region and are frequently reported by agricultural experts and local farmers. These include grassy shoot, mosaic, pokkah boeng, red leaf spot, red rot, ring spot, wilt, and yellow leaf disease, along with a healthy class representing disease-free plants.

Each class contains a significant number of samples to ensure sufficient representation of disease characteristics for machine learning applications. Most classes contain 611 images, while certain classes such as red rot, ring spot, yellow leaf disease, and healthy plants include slightly higher sample counts to capture greater variability in disease symptoms and plant conditions. The inclusion of a healthy category is essential for enabling binary and multi-class classification models to differentiate between infected and non-infected plants.

Table 1 presents the class-wise distribution of images included in the dataset. The dataset is moderately balanced across disease categories, which helps reduce class bias during training of deep learning models. Images within each class capture variations in symptom severity, leaf orientation, lighting conditions, and background environment, thereby increasing the robustness and real-world applicability of the dataset. The balanced representation of disease classes makes the dataset suitable for multi-class classification tasks and comparative evaluation of deep learning models.



Fig. 2: Representative samples of sugarcane disease classes included in the dataset.

Table 1: Class-wise distribution of sugarcane disease images in the Assam Sugarcane Disease Dataset

Categories	Types of Diseases	No of image Samples
Infected	Grassy Shoot	611

	Mosaic	611
	Pokkah Boeng	611
	Red Leaf Spot	611
	Red Rot	678
	Ring Spot	679
	Wilt	611
	Yellow Leaf Disease	678
Healthy		679
Total	5769	

2.4 Data Annotation and Quality Control

Following data acquisition and labeling, several preprocessing steps were performed to ensure consistency and suitability of the dataset for machine learning and deep learning applications. Since the images were captured using three different camera devices, the raw images exhibited variations in resolution, size, and file characteristics. To maintain uniformity and facilitate efficient model training, all images were standardized through a structured preprocessing pipeline.

Initially, the collected images were manually inspected to remove low-quality samples such as blurred, noisy, or partially captured images that could negatively affect model performance. This filtering process ensured that only clear and informative images representing visible disease symptoms were retained in the dataset.

After filtering, the remaining images were resized to a uniform resolution to ensure compatibility with deep learning architectures that require fixed input dimensions. The resizing process preserved essential disease features while reducing computational complexity during model training.

In addition to resizing, the images were organized into separate directories corresponding to their respective class labels. This structured organization simplifies dataset accessibility and enables efficient integration with common machine learning frameworks.

The preprocessing pipeline, illustrated in Fig. 1, ensures that the dataset maintains consistency in image quality and structure while preserving the natural variations observed under real field conditions. These preprocessing steps improve the usability of the dataset for training, validation, and benchmarking of artificial intelligence models for

3. Technical Validation

To evaluate the quality, consistency, and suitability of the proposed dataset for machine learning applications, a combination of statistical analysis and visualization techniques was applied. Fig. 3 illustrates the results of the dataset analysis, including class distribution, pixel statistics, and sample visualizations.

All images in the dataset were resized to a uniform resolution of 128×128 pixels during preprocessing to ensure consistent spatial dimensions for deep learning model training. This standardization helps reduce computational complexity while preserving essential visual features required for disease recognition.

The class distribution analysis demonstrates that the dataset is well balanced across the nine categories. Each disease class contains approximately 611–679 images, resulting in minimal class imbalance. A balanced dataset is important for training reliable classification models

and reducing bias toward specific classes. The distribution plot confirms that the samples are evenly distributed across the disease and healthy categories.

To further examine the statistical characteristics of the dataset, pixel intensity statistics were computed for each color channel. The mean pixel values for the red, green, and blue channels were observed to be approximately 119.42, 126.56, and 83.44, respectively, while the standard deviations were 50.92, 51.48, and 48.89. These values indicate that the dataset maintains stable brightness and contrast levels across images. The relatively lower mean intensity in the blue channel suggests slightly darker blue components, which is consistent with the natural lighting conditions and shadow regions present in field-acquired images.

The pixel intensity histograms further confirm this observation. The red and green channels exhibit broader distributions, reflecting greater variability in vegetation color and disease symptoms. In contrast, the blue channel shows a comparatively narrower distribution, which is expected in outdoor agricultural imagery where green plant structures dominate the visual spectrum.

Dataset redundancy was also analyzed to ensure sample diversity. Out of the 5,769 total images, 4,773 were identified as unique samples, indicating that approximately 83% of the dataset consists of distinct images. This high proportion of unique samples confirms that the dataset contains minimal duplication and maintains sufficient variability for training robust machine learning models.

Loaded 5769 images with labels.

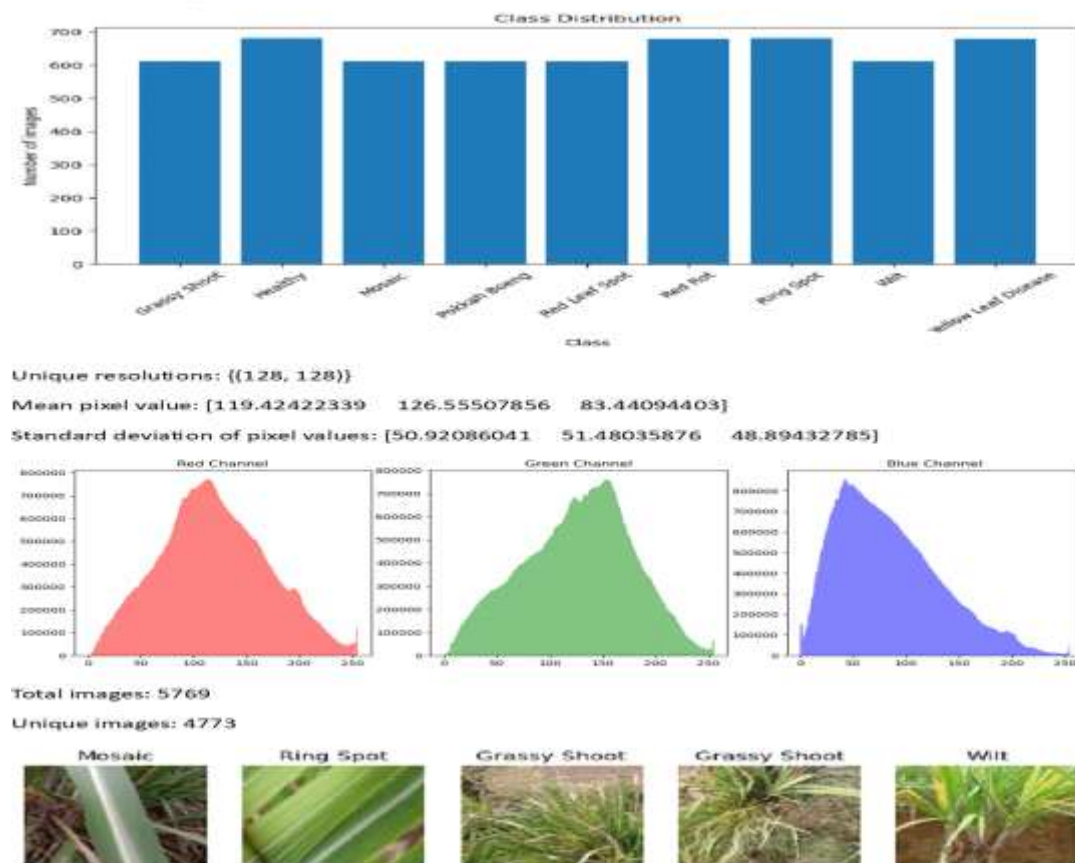


Fig. 3: Visualization and Statistical Analysis of the Dataset

Overall, the combination of balanced class distribution, consistent image resolution, stable pixel statistics, and high sample diversity demonstrates that the dataset is well structured and suitable for training and evaluating deep learning models for automated sugarcane disease detection.

4. Benchmark Deep Learning Evaluation

To further validate the effectiveness and usability of the proposed dataset for automated disease classification, benchmark experiments were conducted using three deep learning architectures: ResNet50, CSPDarkNet, and Vision Transformer (ViT). These models were selected to represent both convolutional neural network-based approaches and transformer-based architectures that are widely used in modern computer vision applications.

The models were trained and evaluated on the prepared dataset using standard classification metrics including accuracy, precision, recall, and F1-score. These metrics provide a comprehensive evaluation of the classification performance across the nine disease categories and help assess the dataset's suitability for deep learning tasks.

The experimental results are summarized in Table 2. Among the evaluated models, the Vision Transformer achieved the highest performance with an accuracy of 0.99, demonstrating its strong ability to capture complex visual patterns associated with sugarcane diseases. CSPDarkNet also performed competitively with an accuracy of 0.97, indicating effective feature extraction capabilities. The baseline ResNet50 model achieved an accuracy of 0.95, confirming that the dataset contains sufficiently discriminative features for accurate disease classification.

Overall, the high classification performance across different model architectures confirms that the dataset provides clear and informative visual representations of sugarcane diseases. These results demonstrate the dataset's potential to support the development of reliable artificial intelligence-based disease detection systems for agricultural applications.

Table 2: Model Performance Comparison

Model	Accuracy	Precision	F1 Score	Recall
ResNet50	0.95	0.95	0.95	0.95
CSPDarkNet	0.97	0.97	0.97	0.97
Vision Transformer	0.99	0.99	0.99	0.99

5. Data Availability

The Assam Sugarcane Disease Dataset developed in this study is publicly available to support research in plant disease detection, precision agriculture, and machine learning applications. The dataset contains 5,769 labeled high-resolution images categorized into nine classes, including eight disease categories and one healthy class. All images are organized into class-specific directories and stored in JPEG (.jpg) format to facilitate easy integration with common machine learning frameworks.

The dataset includes images captured under natural field conditions across multiple districts of Assam, India, using different camera devices. The dataset, along with metadata and

documentation describing the class labels and acquisition process, is available at the following repository:

Dataset repository: <https://data.mendeley.com/datasets/x98kpjvxxg/2>

Researchers can freely download and use the dataset for academic and research purposes, provided that appropriate citation of this dataset publication is given. The benchmark results indicate that the dataset can support the development of reliable disease classification models and can be integrated into mobile-based crop disease diagnostic systems or automated agricultural monitoring platforms to assist farmers and agricultural extension services

6. Reuse Potential

The Assam Sugarcane Disease Dataset provides a valuable resource for interdisciplinary research involving agriculture, computer vision, and artificial intelligence. Due to its field-acquired nature, balanced class distribution, and diversity of disease symptoms, the dataset can be used to develop and evaluate machine learning models for automated disease detection and classification in sugarcane crops.

Researchers can utilize the dataset for a variety of applications, including image classification, disease localization, object detection, and early warning systems for crop disease monitoring. The dataset can also serve as a benchmark for evaluating deep learning architectures such as convolutional neural networks and transformer-based models for agricultural image analysis.

In addition to machine learning research, the dataset may assist agronomists, plant pathologists, and agricultural extension workers in understanding visual disease characteristics and improving disease management strategies. The dataset may also support the development of mobile-based disease diagnostic tools and decision-support systems that can help farmers detect diseases at early stages and take appropriate preventive measures.

Future work may expand the dataset by including additional disease categories, more geographical locations, and seasonal variations to further enhance its diversity and applicability.

7. Conclusion

This study presents the first publicly available multi-district sugarcane disease image dataset from Assam, comprising 5,769 labeled images across nine classes. The dataset captures real field conditions and includes diverse disease manifestations, making it suitable for machine learning-based disease detection. Statistical analysis and benchmark deep learning experiments confirm the dataset's reliability and usefulness for developing intelligent agricultural monitoring systems.

Despite its strengths, the dataset has certain limitations. Although images were collected from multiple locations in Assam, the dataset primarily represents disease patterns observed during a specific cultivation season. Additionally, while the dataset captures several major sugarcane diseases, other less prevalent diseases were not included due to limited field occurrence during the data collection period. Future work will focus on expanding the dataset with additional disease categories, seasonal variations, and larger geographical coverage.

8. Funding Information

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

9. Credit Authorship Contribution Statement

Ankita Goyal: Conceptualization, Data curation, Methodology, Investigation, Writing - original draft preparation.

Mirzanur Rahman: Supervision, Validation, Writing – review and editing.

Jyotismita Das, Yash Pareek, Masash Choudhury: Dataset Curation

10. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

11. Acknowledgement

We are very grateful to the farmers and the nearby residents of the field who helped us in data collection. Also, we acknowledge the efforts of Dr. Tulshi Prasad Saikia, Chief Scientist and Dr. Tulika Medhi, Scientist of the Assam Agricultural University-Sugarcane, Medicinal and Aromatic Plant Research Station, Buralikson, Assam for being very cooperative and helping us with the data verification and guiding us throughout the classification process.

References

- [1] Oerke, E.-C., 2006. Crop losses to pests. *Journal of Agricultural Science*, 144(1), 31–43.
- [2] Mohanty, S.P., Hughes, D.P., Salathé, M., 2016. Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, 7, 1419.
- [3] Sladojevic, S., Arsenovic, M., Anderla, A., Culibrk, D., Stefanovic, D., 2016. Deep neural networks based recognition of plant diseases by leaf image classification. *Computational Intelligence and Neuroscience*, 2016, 3289801.
- [4] Barbedo, J.G.A., 2018. Factors influencing the use of deep learning for plant disease recognition. *Biosystems Engineering*, 172, 84–91.
- [5] Vikaspedia: Sugarcane Diseases and Symptoms. www.vikaspedia.in, <https://vikaspedia.in/agriculture/crop-production/integrated-pest-managment/ipm-for-commercial-crops/ipm-strategies-for-sugarcane/sugarcane-diseases-and-symptoms> (accessed 13 May 2024).
- [6] Sarmah, C. and Sarma, P., A dataset of Sattriya dance: Classical dance of Assam. *Data in Brief*, 52, p (2024).109878. <https://doi.org/10.1016/j.dib.2023.109878>
- [7] Thite, S., Suryawanshi, Y., Patil, K., Chumchu, P., 2024. Sugarcane leaf dataset: A dataset for disease detection and classification. *Data in Brief* 53, 110268.
- [8] Goyal Agarwala, Ankita; Rahman, Dr. Mirzanur; Sarwar Zaman, Dr Gaffar (2024), "Sugarcane Plant Disease Dataset of Assam", Mendeley Data, V1, doi: 10.17632/x98kpjvxxg.1