Journal of Computer Science Review and Engineering | ***ISSN 1694-4488***
Published by AIR JOURNALS | *https://airjournal.org/jcsre*
16/18 Avenue des Longaniers, Quatre Bornes, Mauritius
airjournals@gmail.com; enquiry@airjournal.org

RESEARCH ARTICLE

Development of a Localized Dataset for Plant Disease Detection in Ohodo, Enugu State

Akobundu, Chinyere I.[1], Nwankwo, Kenneth O.[2] & Salaudeen, Habib L.[3]
Computer Science Technology, Federal Polytechnic Ohodo, Enugu State, Nigeria

Citations - APA

Akobundu, C. I., Nwankwo, K. O. & Salaudeen, H. L. (2025). Development of a Localized Dataset for Plant Disease Detection in Ohodo, Enugu State. *Journal of Computer Science Review and Engineering, 9(3) 1-9. DOI: https://doi.org/10.5281/zenodo.17087651*

ABSTRACT

In recent times in Nigeria, there has been incessant rise in food insecurity which has not only affected the individuals but has also drastically affected the stability of the nations economy. Plant diseases have been identified as one of the major challenges facing farmers worldwide leading to substantial crop losses and economic hardship. This research addresses the critical challenge of plant disease detection in Ohodo, Enugu state, Nigeria's agricultural sector by developing and curating a localized, high-quality dataset from Ohodo, Enugu State. Unlike previous studies that rely on generic, globally sourced datasets, this work emphasizes the need for an indigenous dataset that accurately represents the unique environmental conditions, crop varieties, and disease strains of a specific region. The core output of this foundational phase is a robust dataset consisting of 356 healthy and 366 unhealthy plant images, which provides a balanced and sufficient resource for training a machine learning model. The study proposes a Convolutional Neural Network (CNN) architecture designed to leverage this localized data. The methodology outlines a systematic approach to data collection, including stratified sampling, high-resolution image capture, and meticulous, expert-driven labeling to ensure data integrity. Ethical considerations, such as informed consent and community engagement, were central to the process, ensuring the research directly serves the needs of the local farming community. The primary finding confirms the successful establishment of this unique dataset, which serves as a critical first step towards creating a more accurate, generalizable, and practical plant disease identification system tailored to local agricultural realities, ultimately aiming to improve crop yields and farmer livelihoods in Ohodo.

Keywords: *Plant Disease Detection; Convolutional Neutral Network (CNN); Food Security; Localized Dataset*

Introduction

Agriculture is the backbone of Nigeria's economy, through crops, raw materials and also an essential component of human civilization (Tirkey, *et al, 2021*) with a significant portion of the population engaged in small-scale farming. However, crop diseases pose a major threat to food security and farmer livelihoods, which has affected crop yields drastically in Nigeria, discouraging the efforts of the farmers and affecting the food bank of the nation, because it is leading to substantial crop yield losses. Plant diseases have been identified as one of the major challenges facing farmers worldwide leading to substantial crop losses and economic hardship (Wang, *et al*., 2021).

Traditional methods of disease identification, which rely on manual inspection by agricultural experts, are often slow, costly, and inaccessible to many rural farmers. This has led to a growing interest in using computer vision and machine learning (ML) to automate the process of plant disease detection. Accurate and rapid plant disease detection is critical for enhancing long-term agricultural yield. Disease infection poses the most significant challenge in crop production, potentially leading to economic losses (Ramanjot, *et al*, 2023; Abbas, *et al., 2024).*

While numerous studies have demonstrated the potential of ML models for plant disease detection, most have relied on publicly available datasets, such as the PlantVillage dataset. These datasets are often collected in controlled laboratory environments or from different geographical locations, which can limit the generalizability and real-world performance of the resulting models. The unique environmental factors, plant varieties, and disease strains present in specific regions like Ohodo, Enugu State, may not be adequately represented in these generic datasets. This data-distribution shift is a critical challenge that hinders the effective deployment of ML solutions in local agricultural settings.

In Ohodo, the prevalence of plant diseases has had a detrimental impact on agricultural enterprises which is the region's major source of income. Currently, the methods of diseases identification is often relying on visual inspection by farmers and it is time-consuming, prone to errors and can delay appropriate intervention. This can result in further spread of the disease, exacerbating crop loses and reducing overall yields. There is no indigenious dataset from Ohodo Enugu on crop health and its improvement.

This research addresses this gap by focusing on the development and curation of a localized, high-quality dataset of healthy and unhealthy plants from Ohodo, Enugu State for plant disease detection in that community using machine learning technique, as machine learning techniques have emerged as promising tools for automating plant disease diagnosis.. This dataset will serve as a foundational resource for training and validating a machine learning model tailored to the specific agricultural context of the region. By gathering data directly from local farms, this study aims to create a dataset that accurately reflects real-world conditions, including varying lighting, complex backgrounds, and multiple disease symptoms. The ultimate goal is to enable the creation of a more robust and accurate plant disease identification system that can empower local farmers with timely and reliable information, leading to improved crop yields and more sustainable agricultural practices. This topic focuses on the foundational step of the broader project—the data collection itself. It highlights the unique challenges and importance of creating a specific, local dataset rather than relying on generic, publicly available ones.

The key concept in this first output of this research is the need for localized dataset, that is an indigenous dataset, which emphasizes that the data is not from a global repository but is specific to the environmental conditions, crop varieties, and disease prevalence in Ohodo, the study area. This makes the dataset more relevant and effective for a model intended for that specific region. positions the dataset itself as the primary research output, acknowledging its critical role in training an accurate and effective machine learning model for plant disease detection.

Review of Related Literature

Data repositories are uniquely positioned to support researchers in sharing scholarly outputs. As funding agencies develop and institute policies for research data access and sharing, institutional data repositories have emerged as a critical feature in ecosystems for data stewardship and sharing. It shows that data repositories can meet and exceed the requirements and recommendations of federal data policy, thereby maximizing the benefits of data sharing (Narlock, *et al.*, 2024). It has been a positive trend to pull dataset from these repositories.

It has been identified that there are emerging concerns that must be addressed to help preserve the integrity and scientific value of this otherwise positive trend. These include: the risk of 'paper mills' mass-producing superficial papers with questionable authorship practices; duplicate publications produced through republishing already available results or by multiple groups testing the same hypothesis using identical datasets and methods without awareness of each other's work; proliferation of false-positive findings due to inadequate adjustment for multiple testing in large datasets(Rudan *et al.*, 2025); which ushers the need for localized data

Localized data is essential for research because it provides nuanced insights into specific community conditions, needs, and contexts that national or global data often miss, leading to more relevant, actionable findings and effective interventions. This data enables researchers to identify unique local challenges, develop context-specific solutions, and ensure that research directly addresses the concerns and realities of the people it aims to serve, ultimately fostering trust, collaboration, and greater impact.

Local context refers to the collective understanding, beliefs, and perceptions held by communities to interpret their surroundings. It goes beyond widely circulated knowledge and is unique to specific local contexts. Researchers can benefit from incorporating local knowledge into their work as it provides alternative theories, unique perspectives, and invaluable insights. By tapping into the knowledge of local communities, researchers can enrich their research findings and make them more relevant and applicable to the local context (geopoll.com, 2023). integrating research within local contexts and involving stakeholders in fieldwork design is critical for achieving impactful and meaningful outcomes

According the study by Gohain, (2021), United States has the highest number of data repositories (1102) followed by Germany (433) and United Kingdom (296). If the dataset is pulled and using for training and testing the machine learning model in this research, the true picture of the situation of Ohodo farmers would not be drawn.

In recent years, machine learning techniques have emerged as promising tools for automating plant disease diagnosis. Part of this literature review explores the state-of-the-art research in this field, focusing on the use of machine learning algorithms for plant disease identification (Ristaino *et al*. 2021)

According to Ristaino *et al.* (2021), Plant disease outbreaks are increasing and threaten food security for the vulnerable in many areas of the world. In order to tackle these grand challenges, a new set of tools that include disease surveillance and improved detection technologies including pathogen sensors and predictive modeling and data analytics are needed to prevent future outbreaks. The United Nations declared 2020 the International Year of Plant Health. It is estimated that food production will need to increase by 60% by 2050 to feed the estimated 10 billion people expected on Earth (Fedoroff, 2015 and Food and Agriculture Organization of the United Nations, 2019).

An increase in production along with a reduction in food loss due to pests and pathogens and food waste will be needed to meet demand (Savary, *et al*. 2019 and Delegado, *et al.* 2017). Global yield losses due to crop pests and diseases on food crops are large, with mean losses ranging from 21.5% (10.1 to 28.1%) in wheat, 30.3% (24.6 to 40.9%) in rice, 22.6% (19.5 to 41.4%) in maize, 17.2% (8.1 to 21%) in potato, and 21.4% (11 to 32.4%) in soybean.

In some regions of the world, which Nigerian is mostly considered, plant diseases have cause significant preharvest losses for smallholder farmers, with over 50% of beans and maize farmers surveyed and over 50% of potato farmers

surveyed are experiencing loss (Delegado, *et al*. 2017). Plant diseases cause significant losses in food crop production that lead not only to lower yields but also to loss of species diversity, mitigation costs due to control measures, and downstream impacts on human heath.

The National Academy of Sciences (2019) published an ambitious agricultural research agenda that emphasized the need for breakthrough technology for the early and rapid detection and prevention of plant diseases. Emerging plant diseases are diseases that 1) have increased in either incidence, geographical, or host range; 2) have changed pathogenesis; 3) have newly evolved; or 4) have been discovered or newly recognized.

Machine Learning techniques present itself as such an emerging breakthrough technology for early and rapid detection and prevention of plant diseases. Among others, Convolutional Neural Networks (CNNs) have been widely adopted for plant disease identification due to their ability to extract complex features from images. Several studies have demonstrated the effectiveness of CNNs in accurately classifying various plant diseases. For instance, **Zhang et al. (2020)** proposed a deep CNN architecture to identify potato late blight, achieving high accuracy rates. **Wang et al. (2021)** used a transfer learning approach with pre-trained CNN models to classify different rice diseases

Methodology

The methodology for developing a localized plant dataset in Ohodo, which would be used for sampling the machine learning model, for the detection of plant disease in this area, focused on four key areas: sample collection, image capture, data management, and ethical considerations.

1. Sample Identification and Collection

This phase is the foundation of the dataset. The researchers worked with local farmers and agricultural extension workers in Ohodo to identify specific farms and plots for data collection. The focus was on collecting samples from the most common crops in the area.

Sampling Strategy: A stratified sampling approach was used to ensure representation across different crop varieties and disease types. The researchers identified both healthy and unhealthy plants. For unhealthy plants, they were categorized based on visible symptoms such as leaf spots, wilting, discolored leaves or stem.

Collection Process: Plant leaves were carefully collected, taking care not to damage them. Each sample was be assigned a unique identifier and immediately bagged to prevent contamination or degradation before image capture.

2. Tools and Techniques for Image Capture

The quality of the images is crucial for the success of the machine learning model. This phase involves using the right equipment and techniques to capture high-quality, consistent images.

Tools: A high-resolution digital camera and a smartphone with a high-quality camera was used. The researchers used a backdrop (a neutral-colored sheet) to isolate the plant part and a lighting source (a ring light) to ensure consistent illumination, reducing shadows and glare.

Techniques: Images were captured from multiple angles and at different distances to provide a comprehensive view of the plant's condition. This was included close-ups of specific symptoms and wider shots of the whole plant. The metadata for each image (healthy or unhealthy) was logged immediately.

3. Data Quality, Diversity, and Labeling

Ensuring the dataset is accurate and robust requires careful data management and labeling.

Data Labeling: This is the most critical step. Each image was meticulously labeled by a trained expert (a plant pathologist or an experienced agricultural extension worker). The label specified the plant health status such as 'healthy' or 'unhealthy'. For a more granular dataset, the collected data was double-checked, this was performed by a second expert to minimize errors.

Data Augmentation (Digital): To increase the diversity and size of the dataset, digital techniques like image rotation, flipping, cropping, and color adjustments was applied. This would help the future machine learning model become more robust and less sensitive to variations in lighting or orientation.

Data Diversity: The collection aimed for a balanced distribution of images across different classes (healthy and unhealthy) to prevent the model from becoming biased towards the most common class.

4. Ethical Considerations

Data collection was conducted ethically, respecting the rights and privacy of the local farmers.

Informed Consent: Researchers obtained explicit permission from the farmers to access their land and collect samples. The purpose of the research was clearly explained to them in a way they understood it, and they were informed about how the data would be used.

Community Engagement: The research was a collaborative effort. Farmers and local agricultural workers were involved in the process (by engaging and paying them stipends for the sample collection), and the potential benefits of the technology was shared with the community. This ensures the research is not only academically sound but also serves the needs of the people it is intended to help.

Proposed Architecture of the Model

This proposed architecture describes a typical Convolutional Neural Network (CNN), which is a type of deep learning model designed for processing visual data. It's a sequential arrangement of layers that transform raw image data

into a classification output, like identifying a plant disease. Figure 1 depicts

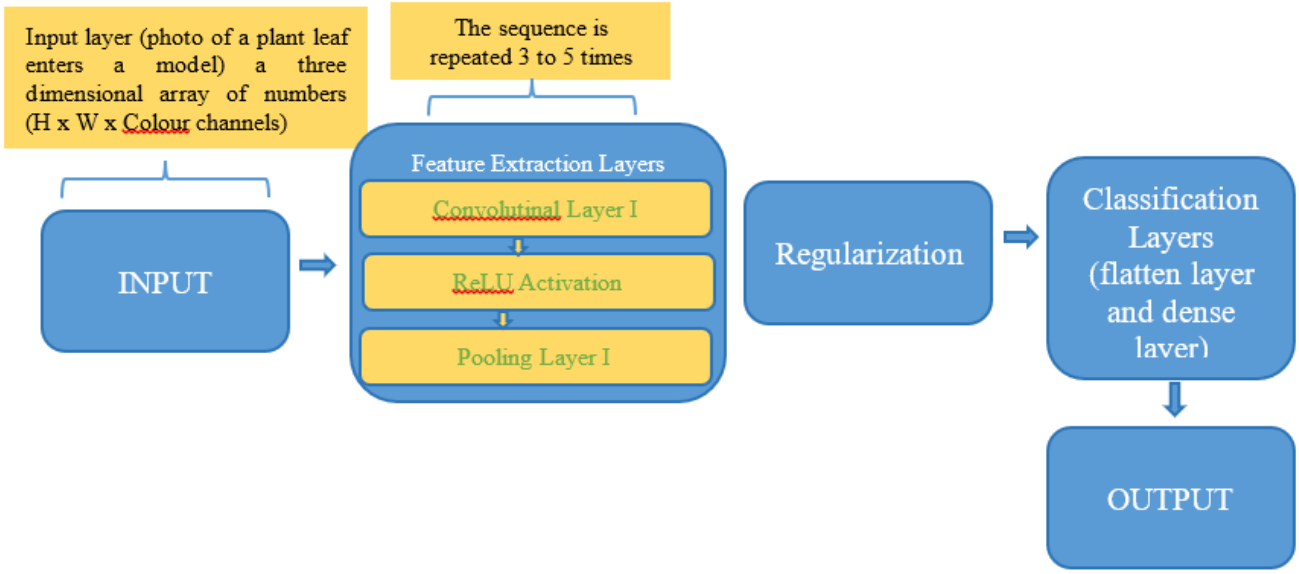

Fig. I: The proposed Architecture of the Model

Convolution → ReLU → Pooling → Dropout → Flatten → Dense → Output

Using 3–5 convolutional layers with increasing filter sizes

1. Convolutional Layers

This is the core of the CNN. A convolutional layer uses a small matrix called a filter (or kernel) to scan the input image. The filter slides over the image, performing a dot product with the pixel values it covers. This process creates a new representation of the image called a feature map. Each filter is designed to detect specific features, such as edges, textures, or patterns. The architecture suggests using 3-5 convolutional layers with increasing filter sizes. This is a common practice. Early layers would use smaller filters to detect simple features like lines or curves. As the data progresses through subsequent layers, the filters increase in size, allowing them to identify more complex, high-level features by combining the simple features from the previous layers.

2. ReLU (Rectified Linear Unit)

ReLU is an activation function applied to the output of each convolutional layer. It's a simple function that changes all negative values in the feature map to zero while leaving positive values unchanged. This non-linear transformation introduces non-linearity into the model, which is essential for learning complex patterns in the data. Without it, the network would only be able to learn linear relationships, which are insufficient for understanding images.

3. Pooling

A pooling layer reduces the spatial dimensions (width and height) of the feature maps. The most common type is max pooling, which takes the maximum value from a small window that slides across the feature map. Pooling helps to downsample the data, which reduces the number of parameters and computation. This makes the model more efficient and also helps to make the model more robust to slight shifts or distortions in the input image.

4. Dropout

Dropout is a regularization technique. During training, it randomly sets a fraction of the neurons in a layer to zero for a specific training iteration.This prevents the model from relying too heavily on a few specific neurons, forcing the network to learn more robust and distributed representations. It is a powerful method to prevent overfitting, where the model performs well on the training data but poorly on new, unseen data.

5. Flatten

The flatten layer takes the multi-dimensional output from the convolutional and pooling layers and converts it into a single, long vector. This prepares the data to be fed into the subsequent dense layers, which only accept one-dimensional data.

6. Dense

Dense layers, also known as fully connected layers, are the final stages of the network. Every neuron in a dense layer is connected to every neuron in the previous layer. These layers perform the final classification. The flattened vector is passed through one or more dense layers, where the network learns to weigh the features extracted by the convolutional layers to make a final prediction.

7. Output

The final layer is the output layer. The number of neurons in this layer corresponds to the number of classes the model needs to predict (e.g., healthy, disease A, disease B). An activation function like softmax is typically used to produce a probability distribution over the classes. This layer provides the final result, indicating the model's confidence that the input image belongs to a specific class.

Findings

This section detailed the output of the data collection efforts which is the first phase of this research. It is a formal report on the dataset itself, which is a major accomplishment of this research phase.

Dataset Size and Scope: A total of seven hundred and twenty-two (722) images were collected from Ohodo, Enugu State, consisting of: three hundred and fifty-six (356) images of healthy plants and three hundred and sixty-six (366) images of unhealthy plants. The collection of 356 healthy and 366 unhealthy images shows a significant finding in this phase of the research. It provides enough data variety for the convolutional neural network (CNN) to learn the distinct features that separate a healthy plant from an unhealthy one. This is a near-equal distribution between the two classes which is regarded as an excellent finding. It ensures that the model will not be biased toward predicting one class over the other, a common problem with imbalanced datasets. This balance contributes directly to the reliability of the future model's performance.

The Uniqueness of Localized Data: This finding goes beyond the numbers. The fact that the data was collected directly from Ohodo, Enugu State, is arguably the most important result of this phase. This dataset is not a generic collection; it is a unique resource that reflects the specific environmental conditions, lighting, backgrounds, and disease types prevalent in the case study area.

Relevance: The model developed will be trained on data that accurately represents its target environment. This drastically increases the likelihood that the final system will be effective and practical for the farmers intended to help, a key objective of this research.

Conclusion

The completion of this initial phase marks a significant and successful milestone in the research project. The core finding is the creation of a unique and valuable localized dataset that accurately reflects the specific agricultural conditions of Ohodo. This dataset's balanced nature and context-specific data overcome the limitations of generic repositories, establishing a robust foundation for building a reliable and relevant machine learning model.

By leveraging this homegrown data, the research is well-positioned to achieve its ultimate goal: empowering local farmers. The proposed CNN, once trained on this dataset, will be more likely to provide accurate and timely disease identification, which can help farmers optimize their strategies, minimize crop losses, and ultimately improve food security in the region. This research demonstrates that a community-centered approach to data collection is not only ethically sound but also essential for creating technological solutions that have a real and meaningful impact on the ground.

Acknowledgement

The researchers would like to express their profound gratitude to the Tertiary Education Trust Fund (TETFUND), Nigeria, for their invaluable financial support. This research would not have been possible without their sponsorship, which has been instrumental in funding the data collection, acquisition of software and hardware tools and foundational work for this research. Their commitment to advancing research and education in Nigeria is highly commendable.

References

Abbas, J., Nabila, B., Rizwan, A., Abolghasem, S., Daesik, J. (2024). Revolutionizing agriculture with artificial intelligence: plant disease detection methods, applications and their limitations. *Frontiers in plant science, Technical advances in plant sciences. 15(1)* 1-20. https://doi.ord/10.3389/fpls.2024.1356260.

Delegado L., Schuster M., Torero M. (2017). *The reality of food losses: A new measurement methodology*. IFPRI Discussion Paper 1686, International Food Policy Research Institute, Washington, DC

Fedoroff, N. (2015). Food in a future of 10 billion. *Agric. & Food Secur. 4, (11),* 31-37 https://doi.org/10.1186/s40066-015-0031-7

Gohain, Rashmi Rekha, "Status of Global Research Data Repository: An Exploratory Study" (2021). Library Philosophy and Practice (e-journal). 5193. https://digitalcommons.unl.edu/libphilprac/5193.

Narlock, M.R., Calvert, S., Taylor, S., Marquez, R.P. and Parkman, A. (2024) 'Knowledge Infrastructures Are Growing Up: The Case for Institutional (Data) Repositories 10 Years After the Holdren Memo', Data Science Journal, 23: 46, pp. 1–17. DOI: https://doi. org/10.5334/dsj-2024-046

National Academies of Sciences, Engineering, and Medicine (2019) *Science Breakthroughs to Advance Food and Agricultural Research by 2030.* National Academies Press, Washington, DC, 2019.

Ramanjot, Mittal, U., Wadhawan, A., Singla, J., Jhanjhi, N. Z., Ghoniem, R. M., Ray, S. K., & Abdelmaboud, A. (2023). Plant Disease Detection and Classification: A Systematic Literature Review. Sensors, 23(10), 4769. https://doi.org/10.3390/s23104769

Ristaino, J., Anderson, P., Bebber, D., Brauman, K., Cunniffe, N., Fedoroff, N., Finegold, C., Garrett, K., Gilligan, C., Jones, C., Martin, M., MacDonald, G., Neenan, P, Records, A., Schmale, D., Tateosian, L., Wei, Q. (2021). The persistent threat of emerging plant disease pandemics to global food security. *Proc Natl Acad Sci U.S.A*.

118(23):e2022239118. doi: 10.1073/pnas.2022239118. Erratum in: Proc Natl Acad Sci U S A. 2021 Oct 5;118(40):e2115792118. doi: 10.1073/pnas.2115792118. PMID: 34021073; PMCID: PMC8201941.

Rudan I, Song P, Adeloye D, Campbell H. Journal of Global Health's Guidelines for Reporting Analyses of Big Data Repositories Open to the Public (GRABDROP): preventing 'paper mills', duplicate publications, misuse of statistical inference, and inappropriate use of artificial intelligence. J Glob Health. 2025;15:01004

Savary S., Laetita, W., Sarah, J., Paul, E., Neil, M., Andy, N. (2019). The global burden of pathogens and pests on major food crops. *Nat. Ecol. Evol. 3(3),* 430–439.

Tirkey, D., Singh, K. K., & Tripathi S. (2023). Performance Analysis of AI-based Solutions for Crop Disease Identification, Detection and Classification. *Smart Agricultural Technology.* Volume 5, October, 2023, 100238 https://doi.org/10.1016/j.atech.2023.100238

Wang, X., Liu, Y., & Zhang, Y. (2021). Transfer learning for rice disease classification using pre-trained convolutional neural networks. *Computer Vision and Image Understanding*, *210 (1),* 1-10.

Wang, Y. P., Pan, Z. C., Yang, L. N., Burdon, J. J., Friberg, H., Sui, Q. J., & Zhan, J. (2021). Optimizing Plant Disease Management in Agricultural Ecosystems Through Rational In-Crop Diversification. *Frontiers in plant science*, *12*, 767209. https://doi.org/10.3389/fpls.2021.767209