IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

Popular Datasets And Their Challenges In Plant Leaf Disease Detection

Jitender Singh, Research Scholar, Maharshi Dayanand University, Rohtak, Haryana
Gopal Singh, Associate Professor, Maharshi Dayanand University, Rohtak, Haryana

Abstract
Crop wastage as a result of diseases is a critical issue, increasing food costs and leading to enormous losses for farmers, hence making agriculture a high-risk profession. Therefore, PLD detection is relevant to the enhancement of food production and security. PLD detection is one of the growing interdisciplinary fields combining artificial intelligence and agricultural science; it has relatively been improved with the advent of transfer learning and deep learning. At the heart of each of these advancements has been access to meaningful datasets. It reviews some of the popular datasets used in PLD detection such as PlantVillage, PDDB, XDB, NLB, LWDCD2020, PDD271, and PlantDoc. This review has identified strengths and weaknesses associated with these datasets. Highlighting issues such as class imbalance, unrealistic laboratory conditions, and variability in data collection methods. Key papers on this topic are identified to show how machine learning and deep learning models have evolved in detecting PLDs. Future research may include data augmentation methods using GANs and the creation of a standardized data collection procedure. It will therefore help in raising the research community's awareness of the status of PLD datasets, underline continuous improvement for better accuracy and reliability in AI-based PLD detection models, and finally aid in sustainable agriculture and food security.

Keywords: Plant Leaf Disease, PLD Data sets, CNN, and Deep Learning.

Abbreviations

AI: Artificial Intelligence
CCTV: Closed Circuit Television
CNN: Convolutional Neural Network
Gen-AI: Generative Artificial Intelligence
LSTM: Long Short-Term Memory
LWDCD2020: Large Wheat Classification Dataset
ML: Machine Learning
NLB: Northern Leaf Blight
PLD: Plant Leaf Disease
UAV: Unmanned Aerial Vehicle

1. Introduction

Globally the crop wastage due to the diseases is a crucial problem that forces the buyer to purchase raw food at an increased cost due to higher demand and lower supply ratio. Farmers and agricultural stakeholders face losses that are difficult to cover and hence agriculture as a profession is associated with a risk. The diagnosis of plant diseases is therefore very crucial in the improvement of food production and food security for any country. Additionally, the plant leaf disease is a recently growing interdisciplinary field accommodating artificial intelligence and agricultural sciences. Researchers from both the fields have shown their interest in the last decade in this domain which was flourished after the introduction of transfer learning and deep learning based architectures. However, just like how human beings learn with experiences, without data, a machine or a model can't learn by themselves. Therefore, the availability of the publicly available data sets is the driving force to any AI based applications. In the field of PLD, the availability of the data set depends on multiple factors. However, the data sets posses their own pros and cons. This review highlights the available public and private data sets with their use cases and the challenges of training a model with these data sets. The research gaps highlighted in this review will further help in the selection of a suitable data set for the training of a PLD model.

2. Related Literature

P.S. Thakur et al. [1] examines the advances and challenges in using vision-based machine learning approaches for detecting plant diseases. The significance of plant diseases in global agriculture is emphasized, acknowledging that manual inspection methods are inadequate and automated solutions are needed. The authors went through 148 papers out of a pool of 1337, discussing downsides of contemporary methodologies particularly their dependence on public datasets collected under controlled environment that might not be applicable in real life situations. Some compact CNN-related techniques seem to have potential; nevertheless, they tend to be hampered due to limited data sets size as well as few diseases considered. Further, A. Ahmad et. al. [2] carried out an in-depth analysis with respect to the application of deep learning methods in agriculture for the diagnosis of plant diseases; on that note, they highlighted that strong disease management systems are important for agriculture to progress. The assessment includes seven important questions on dataset needs, image sensors, deep learning methods, model transferability, estimation of disease severity, and comparing the accuracy of deep learning with human performance. Points out how far deep learning has deviated from traditional image processing and machine learning techniques in highlighting the advantages that accruing to it in regard to automatic features extraction and accuracy. It has also highlighted the keynote aspects of early disease detection, diversified datasets like PlantVillage, the role of different imaging sensors and platforms in data collection, and the construction of deep structures with a keynote role of datasets leading to improved model efficiency. The authors concluded with suggesting that an effective plant disease control system must be devised to include early detection, precise assessment of severity, and the capability for identifying a range of crops so as to give farmers comprehensive and mechanized solutions. This provides a good, fundamental basis for the extension of deep learning applicability in precision agriculture to assist farmers in enhancing crop health and reducing losses in yield. Further, A. Abade et al. [3] reviews the application of CNNs in the identification and classification of plant diseases, focusing on precision agriculture. In the last decade, the review analyzed 121 studies about disease identification methods, dataset attributes, and crops or pathogens used. The findings of the research indicate that CNNs are very vital in improving the accuracy and speed of crop diseases detection, especially compared to the traditional manual and mechanized means that cannot cover large areas and give real-time data. Other benefits, among which is the capability of self-extracting features, include the provision of very high accuracy by CNNs, surpassing human performance on some large-scale recognition tasks. Along with data sets, the authors also had drawn attention towards state-of-the-art CNN architectures such as LeNet, AlexNet, VGGNet, GoogLeNet, InceptionV3, ResNet, and DenseNet in successfully diagnosing diseases. Contrarily, despite these progressions, difficulties still arise from genetic and phenotypic variations in crops, while the range of considered pests and diseases may cause overlapping symptoms and be caused by both biotic and abiotic factors that may result in misdiagnosis. It also points out the potential future research in the development of CNN designs applicable for most crops and diseases.

3. Popular Datasets in the Field of PLD

3.1.PlantVillage

The PlantVillage [4] is the earliest data set in this domain. For disease identification, PlantVillage dataset consists more than 50000 images of healthy and affected plant leaves selected individually. This assists in the problems that plant diseases bring to world food safety. The images which have been collected from the experimental research stations and which include 14 crop species and several disease kinds are to be used for the training of machines and crowdsourcing. The dataset of over 150 crops and 1800 diseases; it is an excellent example of crowdsourcing of solutions. PlantVillage demonstrates how the power of human intelligence combined with artificial intelligence can significantly improve disease diagnosis while sustaining greater reach and thus supporting sustainable agriculture as well as the food security crisis. This is clearly seen from the various translations of the different languages and also by the enhanced rate of people participation.

3.2.PDDB and XDB

Embrapa Agriculture Institute developed Plant Disease Detection database (PDDB) [5] consisting of 2,326 images of affected plant parts. It was later expanded into the XDB dataset, which contains 46,513 images of leaves from 18 plants whose leaves are affected by 93 different categories of disease. Furthermore, Digipathos is a new dataset that has 46,513 images representing 171 diseases on various crops. Only 2326 images out of these capture leaf diseases in controlled lab environments with consistent backgrounds or fields with complex backgrounds. The other remaining 45187 images are just for cropped disease lesions hence more than 95% do not represent real-field conditions. Even under this limitation; many studies have employed the Digipathos dataset to identify diseases in plants and it is available at Digipathos. The availability of these large datasets has greatly improved research on plant disease detection thereby providing an important tool for training and evaluating machine learning models. Such datasets have enabled researchers to develop more robust and accurate diagnostic tools required for improving crop health and agricultural productivity. As technology advances further improvements to existing datasets will be paramount in addressing emerging complexities in detecting plant diseases.

3.3.NLB

The study and proposed data set [6] is intended to address challenges that can be linked to complex field backgrounds and lighting configuration that prevents identification of diseases. To this end, based on deep learning, the researchers developed a CNN multi-scale feature fusion instance detection approach. It comes with new improvements in the data preprocessing with a faster Retinex algorithm for handling of low-flooding light intrusions, sharper tuning of the Region Proposal Network (RPN) for tuning the anchor boxes, integration of a transmission module for combining low- and high-level features The dataset proposed, which is the largest open dataset in NLP, is consists of a hand-held, boom-camera, and drone images, and among the hand-held Photometric and geometric augmentation techniques increased the data set from the original dataset to 8152 for train, validation and test data set implying a much larger data set . Finally, reaching 60000 iterations of the improvement in the model, the mean average precision (mAP) was attained to 91. 83% which was even higher and more accurate than the existing method for the identification of patients with a disease or infection. But it shows how deep learning can achieve accurate and fast NLB identification in the real-world working conditions.

3.4.LWDCD2020

The third most grown and utilized grain is the wheat and of this, a moderate portion is affected by diseases, so a new deep learning system for discriminating the type of wheat sickness into ten classes, with a test accuracy of 97%. 88% responded yes with the improvement made of 7.01% and 15. It was noted that there was a raised detection of 92% as compared to the achievements of the VGG16 and RESNET50 models.
The introduced LWDCD2020 [7] dataset contains about 12,000 images, which describe nine types of diseases in wheat and one type of healthy condition. It was noted that 40% of such pictures were taken in the

field and the other 60% of pictures were taken from the existing datasets available for public use. One of the pictures was rotated and all the pictures were also zoomed in the sense that an attempt was made to increase the amount of the available data. Again, the unpredictability of the backgrounds; miscellaneous capturing scenarios along with the different phases of the diseases also help in formulating a strong training profile for the model to serve as a stable base for the training.

3.5.PDD271

Some of the measures such as the transfer learning that involves application of adjustments to outstanding models for significant datasets have also enhanced the efficiency and accuracy. Some methods of weakly-supervised learning have been suggested to promote inclusion of areas that can be quickly identified in the diseases without additional pixel-level labels that can guide the model closer to the areas of the images required for diagnosis. The problem of different symptoms and complex background in plant disease pictures is solved through datasets like LWDCD2020 for classification of wheat disease and PDD271 [8] having 220,592 pictures for 271 types of plant diseases. The availability of these datasets has led to procedures that adjust the image's regions and the loss functions towards tuning to the disease region to enhance recognition effects. Emphasizing the further improvement of the signal we have got; of course, with the help of certain techniques, for example, extending weighed patch feature sequences via LSTM networks to detail feature representations we obtain the improvement of accuracy to an extent. Due to the integration of image processing and deep learning, the diagnostics of diseases in plants has improved; it contributes to the enhancement of accuracy in the agricultural operations and the enhancement of the results, crop yields, and stability of the food chain as well.

3.6.PlantDoc

To reach receptive readers, PlantDoc [9] contains 2,598 photographs of 17 illnesses in 13 different crops and overcomes challenges involved in using deep learning for disease detection. The dataset consists of images taken from various field environments – some having consistent backgrounds. There are also pictures showing different types of diseased leaves or entire plants which can make training the model harder. As a result only a few studies have used PlantDoc despite this lack of balance and few images per class (less than 200) limiting its usefulness because it has several problems that prevent accurate identification of diseases. To annotate this dataset required 300 human hours but it is promising because the classification accuracy can be increased by up to 31% with the aim to propagate early plant disease detection via computer vision techniques thus reducing India's enormous crop yield losses. In the absence of large non-laboratory datasets, developing a dataset is a response to the major challenge of early disease detection in agriculture. PlantDoc employs online images and detailed notes to improve machine learning models in real-world environments. This makes it useful beyond academic research and could help farmers and agronomists develop preventive disease control measures. Computer vision progress continues with such forms of datasets as PlantDoc playing an important role towards attaining affordable agricultural technology that can be applied on areas whose agricultural productivity is unstable hence ensuring sustainable food production.

Dataset	Year	No of Images	No. of classes	Condition Lab	In-wild Condition	Class Balance
PlantVillage	2015	54305	38	Yes	No	Unbalanced
PDDB (DigiPathos)	2018	2326	171	Yes	Yes	Unbalanced
XDB	2018	46513	105	Yes	No	Unbalanced
NLB	2020	105	1	No	Yes	Unbalanced
LWDCD2020	2021	12160	10	No	Yes	Unbalanced
PDD271	2021	220592	271	No	Yes	Unbalanced
PlantDoc	2021	2598	17	No	Yes	Unbalanced

4. Discussions and Future Scope

The datasets discussed in this paper are widely employed in this domain. Initially, the only availability for the researchers was the PlantVillage dataset and therefore most of the earlier literature in this domain involves training of models with this only dataset. With the growing time, researchers started collecting data for their own projects which facilitated the growing number of data sets.

The shortcomings of the available data sets include the class imbalance, the images not being captured in the real world conditioned, captured in a specific season and hence minimal generalization capabilities. Another issue with the mode of data collection is the images were captured using a UAV, handheld, drone, etc that introduces different complexities in the generalization and backgrounds of the images. While, only considering a laboratory-based image data set that consists of uniform background may speed up the training and classification, but may not be a robust model that can detect diseases in a real-world condition where a farmer is expected to capture an image find diseases.

The future scope of these data sets may include using GAN based CNN to generate images and minimize class imbalance along with common data augmentation strategies. Further, a steady or stable UAV based image collection techniques may be considered to gain uniformity in the data set. Counterwise, multiple permutation and combinations of the UAV based images, handheld cameras, CCTV, and drones can be taken in order to collect and design a more robust dataset.

5. Conclusion

The research field of PLD detection has grown tremendously due to the incorporation of Artificial Intelligence and Agricultural Sciences, with transfer learning and deep learning approaches. A part of this review is devoted to discussing the critical role of publicly available datasets in driving such progress, reporting their characteristics, advantages, and limitations for some of the prominent datasets like PlantVillage, PDDB, XDB, NLB, LWDCD2020, PDD271, and PlantDoc. Progress from dependency on early datasets like PlantVillage to more and more diverse and complete ones proves the maturation of the field. These improvements, however, are counterbalanced by perpetual problems: class imbalance; non-realistic conditions in the laboratory based images; and variability introduced by acquisition methods, which reduces generalization and lowers the robustness of PLD detection models, hence their practical applicability in real agriculture. Future research targeting these challenges should focus on advanced data augmentation techniques that involve class balancing through Generative Adversarial Networks, along with the development of more homogeneous and coherent data collection procedures. Therefore, images captured with different methods like UAV, handheld camera, and drone can be combined to create more robust datasets representative of real-world conditions. Finally, much improvement has been done on the detection of PLD, but still, there is a need for further efforts from the side of raising data set quality and diversity. The improvement in any component would make deep learning models more accurate and reliable. All these aspects help realize sustainable agriculture and food security by enabling effective and timely disease detection.

References

[1] P. S. Thakur, P. Khanna, T. Sheorey, and A. Ojha, “Trends in vision-based machine learning techniques for plant disease identification: A systematic review,” *Expert Systems with Applications*, p. 118117, Jul. 2022, doi: https://doi.org/10.1016/j.eswa.2022.118117

[2] A. Ahmad, D. Saraswat, and A. El Gamal, “A survey on using deep learning techniques for plant disease diagnosis and recommendations for development of appropriate tools,” *Smart Agricultural Technology*, vol. 3, p. 100083, Feb. 2023, doi: https://doi.org/10.1016/j.atech.2022.100083

[3] Abade, P. A. Ferreira, and F. de Barros Vidal, “Plant diseases recognition on images using convolutional neural networks: A systematic review,” *Computers and Electronics in Agriculture*, vol. 185, p. 106125, Jun. 2021, doi: https://doi.org/10.1016/j.compag.2021.106125.

[4] D. P. Hughes and M. Salathe, “An open access repository of images on plant health to enable the development of mobile disease diagnostics,” *arXiv.org*, Apr. 11, 2016. http://arxiv.org/abs/1511.08060

[5] J. Garcia Arnal Barbedo *et al.*, “Annotated Plant Pathology Databases for Image-Based Detection and Recognition of Diseases,” *IEEE Latin America Transactions*, vol. 16, no. 6, pp. 1749–1757, Jun. 2018, doi: https://doi.org/10.1109/tla.2018.8444395.

[6] J. Sun, Y. Yang, X. He, and X. Wu, “Northern Maize Leaf Blight Detection Under Complex Field Environment Based on Deep Learning,” *IEEE Access*, vol. 8, pp. 33679–33688, 2020, doi: https://doi.org/10.1109/access.2020.2973658.

[7] L. Goyal, C. M. Sharma, A. Singh, and P. K. Singh, “Leaf and spike wheat disease detection & classification using an improved deep convolutional architecture,” *Informatics in Medicine Unlocked*, vol. 25, p. 100642, 2021, doi: https://doi.org/10.1016/j.imu.2021.100642.

[8] X. Liu, W. Min, S. Mei, L. Wang, and S. Jiang, “Plant Disease Recognition: A Large-Scale Benchmark Dataset and a Visual Region and Loss Reweighting Approach,” *IEEE Transactions on Image Processing*, vol. 30, pp. 2003–2015, 2021, doi: https://doi.org/10.1109/tip.2021.3049334.

[9] D. Wang, J. Wang, W. Li, and P. Guan, “T-CNN: Trilinear convolutional neural networks model for visual detection of plant diseases,” vol. 190, pp. 106468–106468, Nov. 2021, doi: https://doi.org/10.1016/j.compag.2021.106468.