

Optimized Light-Weight Deep Learning Model for Rice Disease Identification

Pardeep Seelwal*¹, Tilak Raj Rohilla²

Submitted: 19/07/2023

Revised: 08/09/2023

Accepted: 21/09/2023

Abstract: From past decade, deep learning models have gained the Convolutional neural network models have made substantial progression in the agricultural sector. But the utilization of the deep learning models is restricted confined because of enormous supernumerary and imperative parameters. In this article, the magnitude based pruning and dynamic range quantization have been employed to optimize the CNN model so as to be deployed on edge devices for the identification of four classes of rice leave i.e. brown spot, hispa, leaf blast and healthy. Experimental results show that classification accuracy achieved by baseline CNN Model for brownspot -97.15%, hispa- 97.03%, leaf blast- 96.94% and the healthy-96.9%. Overall test accuracy using baseline CNN model is 98.11%, using magnitude base pruning is 97.39% and using dynamic range quantization and pruning is 96.02%. The initial model size of the cnn model without pruning is 78.24 MB, model size with pruning is 25.743 MB, and with quantization model size achieved is 21.88 MB. The proposed work can deploy the models on edge devices that would be light weight with less memory consumption.

Keywords: Data augmentation, pruning, deep learning, compression, fine-tuning

1. Introduction

Rice, being the world's main staple food, plays a crucial role in feeding over half of the global population. However, rice diseases pose a significant threat to its yield and quality. These diseases exhibit diverse types, widespread occurrence, strong epidemics, and severe damage, leading to substantial annual losses. To address this issue and improve rice production, it is essential to diagnose rice diseases quickly and efficiently while implementing effective control measures. Thanks to advancements in computational intelligence and machine learning, image recognition technology, which has been successfully applied in various fields like facial recognition and disease classification, holds great potential for detecting and diagnosing rice diseases [1].

Deep Learning (DL), a powerful technique in machine learning, has facilitated significant progress in developing automated diagnosis of plant diseases. The collaboration between multiple disciplines has led to ingenious DL models, offering promising outcomes and substantial potential for addressing plant disorder issues. Consequently, there is a growing demand for an automated disease diagnosis approach to enhance measurement efficiency. A classification system based on convolutional neural networks (CNN) can greatly assist farmers in improving productivity and diagnostic accuracy. While CNN has demonstrated remarkable accuracy by leveraging

its robust feature extraction capabilities, it often requires significant computational resources and storage space, making it more suitable for high-performance devices. However, for mobile and low-cost devices, fast diagnosis is essential, necessitating the development of models that are efficient, accurate, and suitable for resource-constrained environments.

In light of the extensive rice production worldwide, it is crucial to devise an approach for accurate and rapid diagnosis that can operate on mobile and edge devices, minimizing waiting times. Few models are proposed on a pruned Convolutional Neural Network model, also known as a sparse model, to enable early detection of plant disease images, ensuring fast diagnosis and preventing the escalation of the situation. The key focus lies in leveraging the redundancy within the CNN's structure and parameters to compress the model effectively, resulting in a streamlined architecture with fewer parameters without compromising the automatic identification of rice diseases [2]. This approach holds immense value in reducing food losses and addressing the challenges associated with rice disease diagnosis.

1.1 Major Contribution

The key contributions of this article are as follows:

- (i) The primary task is to create a lightweight CNN model for rice disease diagnosis that can be further deployed on low-cost mobile and edge devices.

^{1,2} Baba Mastmath University, Rohtak, India

* Corresponding Author Email: pardeepselwal@gmail.com

- (ii) A comparison analysis is done made the baseline CNN model and magnitude based pruning used in this work.
- (iii) Compression technique is further applied to the pruned model for reducing the more network size without compromising the precision of the model.

The paper is organized as follows: (1) Introduction about the need for pruning, (2) Literature Review, (3) Material and Methodology, (4) Model Building (5) Model Compression and Quantization (6) Result Analysis and Discussion, and (7) Conclusion.

2. Literature Review

Convolutional Neural Networks have achieved considerable success in a variety of manual diagnosis mechanism, but their huge commutative and space complexity prevent their use in real-time remote sensing activities. In order to do this, filter pruning approaches that allow for the implementation of deep networks on remote sensing equipment while experiencing tolerable performance decreases have received a lot of attention. In this research, the author presented a novel method for accelerating and compressing conventional CNNs, called Pruning Filter with Attention Mechanism. To choose the filters to be pruned, a unique correlation based filter pruning framework is used, that uses an attention module to examine the far-reaching connections between filters. Three publicly available remote sensing datasets are used to test the proposed method, and the experimental findings show that it outperforms state-of-the-art standards. It is difficult to run computationally demanding machine learning computations on mobile monitoring components, making it difficult to detect microbiological ailments in rice fields. Farmers' fields Equations in rural areas with inadequate internet service make this problem worse. Therefore, in farms implementing digital agricultural practices, a plant-specific solution tailored to each individual plant is needed to detect environmental strain caused by crop viruses [3]. The potential of modern microbial disease detection technologies to make conclusions in real-time is lacking. In handheld technology with low processing abilities, it is necessary to have a mechanism suitable for rendering selections. In order to identify biological disorders in rice crops using the processing power of mobile devices, this research suggests RiceBioS, an AI-based deep learning-enabled handheld device. RiceBioS uses Edge-as-a-Service as a method to divide the rice plant image samples into two groups: infected and healthy. To reduced deep learning classification model, pruning is employed which uses the novel thresholding and progressive masking approaches to conduct dimension diminution, further identifies the biotic stress situation. The user-friendly

RiceBioS mobile application interface's real-time insights assist the farmers make informed decisions with this state-of-the-art technology [4]. These days plant diseases significantly reduce agricultural productivity and threaten food security. The solution of minimizing losses is the rapid and reliable identification of plant diseases. Deep neural networks have currently been widely used to identify diseases of plants, although these methods still have poor detection accuracy and require a lot of parameters. In order to identify diseases in endemic kinds, the authors offered the CACPNET model, which combines channel attention and channel pruning. On the public Plant-Village dataset, CACPNET's accuracy gets 99.7%, while on the dataset for the private peanut leaf disease, it reaches 97.7%. Furthermore, CACPNET performs better than the present methods in terms of throughput and inference time, achieving 22.8 ms/frame and 75.5 ms/frames, respectively. According to the findings, CACPNET is a promising candidate for implementation on edge devices to increase the effectiveness of sustainable farming in identifying illnesses in plants [5].

It is suggested to use a multi-crop disease detection algorithm which cans categories plant illnesses regardless of the crops. Complete Concatenated Block is utilized as a basic enity of operation in this structure. To limit the amount of parameters generation of the model, the convolution layer is placed prior to each convolution layer. It improves the use of feature maps and contributes to higher classification accuracy. Subsequently, the model has been trained with pruning to reduce model size. The Partial Standard Convolution approach model outgunned and produced 98.14% accuracy with a small model size of 10 MB in this recommended framework [6]. The production of rice crops is affected by a number of variables, including fertile soil, availability of water, climate changes, illnesses, and parasites. Finding the underlying reason for the decreased paddy yield is essential. With the help of convolutional neural networks, the study can help farmers in automatically diagnose leaf illnesses. Five classes of paddy leaf disease included by the authors comprising of blast, brown spot, tungro, bacterial blight, and healthy leaves. The use of an extended Huber loss function to minimize the loss is major contribution of the paper. It is also cross-compared with current loss functions. With five classes of paddy leaves, the suggested model had a 96.63% training accuracy and 86.61% validation accuracy [7].

3. Material and Method

3.1. Dataset

In this work, four classes of rice plants have been selected for the classification, including three diseases samples of rice i.e. brown spot, hispa, and leaf blast, as well as healthy

rice image samples [8]. The image samples were collected from kaggle which the source of publically available dataset for rice plant disease. Preprocessing techniques were applied to refine the quality of the image samples. Furthermore, data augmentation techniques were applied on the processed images before the training of the data samples. The pre-processing and the data augmentation applied on the dataset are further explained in the section. Fig. 1. shows the image samples of all the four classes used for model training where fig. 1(a) is the brownspot, 1(b) is leaf blast, 1(c) is hispa and 1(d) is healthy image sample of the rice plants.

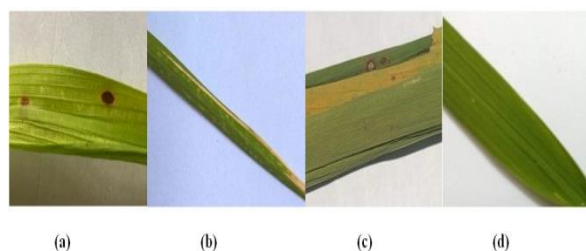


Fig. 1. Images samples of the four classes of the rice plant

3.1.1. Pre-processing and Data Augmentation

Images are pre-processed using different approaches before forwarding to the CNN model for training. Pre-processing can be able to reduce the computations and dimensions of the images in order to improve the final outcome. Primarily images were grey scaled with varying sizes, which can further uniform to $28 \times 28 \times 1$. The intensity pixel value range is changed by normalizing the image's pixel values. As previously stated, the computation is complicated since the greyscale channels lies between the ranges of 0-255, so the normalization can be employed to make range of image pixels between 0-1. In order to minimizing the risk of overfitting and getting the better results ,data augmentations technique is applied to expand the dimensions and enhance the quality measure of the dataset of rice leaves. To set up for the network to map input to output samples, training of CNN essentially entails adjusting parameters until the best point with the lowest model loss is reached [8]. The training complexity is directly correlated with the number of network parameters. As a result, training CNN needs a lot of sample data to get large number of parameters. The trained model performs better and has a bigger capacity for generalization as the number of data increases. The present work initially preprocesses rice disease data, using seven data augmentation functions operations, such as shearing, zca whitening, random rotation, horizontal flipping, height shift and width shift, brightness and vertical flipping. These techniques can to protect the model from overfitting, expedite up the convergence of the model, and boost the

robustness of the predictions [10]. Fig. 2 represented image samples of rice plants after applying the data augmentation techniques. Table 1 shows the number of original images of rice leaves and the images received after applying the data augmentation techniques.

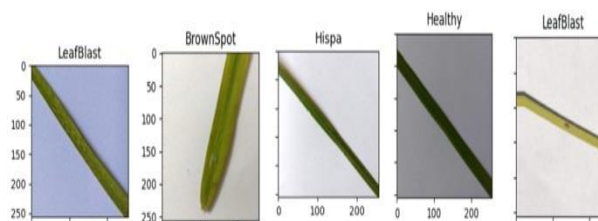


Fig. 2. Image samples of rice plants after data augmentation

Table 1: Dataset Count for Each Class of Rice

Rice Class	Original Samples	After Data-Augmentation
Brown Spot	523	3,661
Hispa	565	3,955
Leaf Blast	779	5,453
Healthy	1188	8,316

4. CNN Model Building

The network structure of the baseline CNN model is presented in Fig. 3, consisting of several blocks of convolution layers that include a convolution layer, flatten, max-pooling, Rectified Linear Unit activation function, and finally a dense layer. The final layer employs the softmax-function to classify four classes of rice plant diseases. The convolution layer is a critical component of the CNN-model that has common weights, responsible for learning the representation of the input features, and contains different feature-vectors. To extract local proprieties, the conformity of neurons is used in different positions of the preceding layer. The filters of size (3×3) are applied to the rice plant image to extracts the feature vectors, followed by the Relu activation function. The layers are organized like neurons in a typical neural network, and the output features are flattened to feed them into the output layer [11]. The Adam optimizer is used for training the neural network. The training and validation set with 8:2 ratios have been used. For training of the network model training and validation sets are utilized. While the test set evaluates the model's performance.

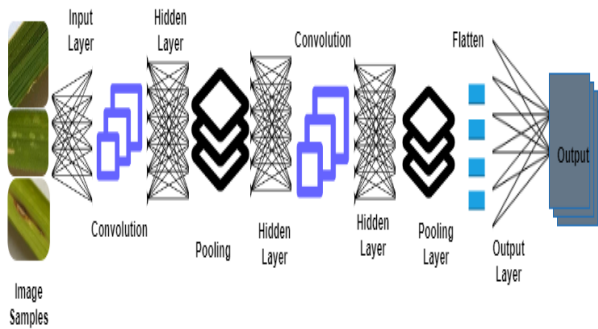


Fig. 3. Baseline CNN Model

4.1 Pruning

Pruning a Neural Network involves the removal of nodes (neurons) and/or edges (weights) from the network in order to reduce the number of parameters. This can help to decrease memory usage and prevent overfitting to training data. It is a careful and iterative process, ensuring that essential neurons and connections are not removed. By gradually identifying and removing less important nodes and edges, iterative pruning ensures that the critical elements of the network are retained [11].

4.2 Pruning in Deep Learning

Deep Neural Networks (DNNs) often require significant computational and memory resources, which can limit their use in the environment with low-resource. Extremely dense networks can also be over-parameterized and mostly susceptible to overfitting and traditional techniques like dropout may not be sufficient for improving computational and memory efficiency. However, by removing insignificant weights from a Neural Network, it is possible to increase the efficiency of learning and classification while requiring very few training samples. This technique can also help to prevent overfitting and improve the overall performance of the network.

4.2.1. Steps to Implement Pruning and Fine-tune pre-trained Model with Pruning

As discussed earlier, baseline CNN model is implemented first then input the rice disease samples to the baseline CNN model. The input samples are pre-processed using rescaling and normalization techniques. After that, train the base line model into 8:2 ratios. Next, magnitude based pruning and structural based pruning is employed and fine-tuned the model with pruning. In this work, Fine tuning is applied on pre-trained model instead of training the model from scratch. Network edges were pruned in a loop by employing the polynomial decay function. Level of Sparsity S is lined up iteratively which is represented by equation 1:

$$S = \xi_t + (\xi_\alpha - \xi_t) \left(1 - \frac{\beta - \alpha}{\alpha - \gamma} \right) \quad (1)$$

Where ξ_t is targeted sparsity; ξ_α is initial sparsity, α is initial iteration, γ is the last iteration, and β is the present iteration. With the help of the sparsity level step, the threshold of valuable links is determined. The threshold value is calculated by multiplying the total weights with the position's step sparsity level. As indicated in equation 1, all mask values associated with a weight are set to 0 when the true value in absolute terms of each weight filter is less than the threshold level. After wards, evaluating the pruning model against the baseline CNN model to get light weight network structure without compromising the accuracy and save the final pruned model. Fig. 4 represents all the steps to be followed to get the pruning results.

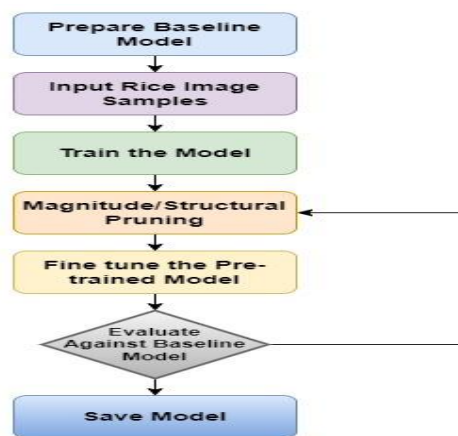


Fig. 4. Steps to be followed for Pruning the Baseline Model

4.3. Magnitude Based Pruning

Magnitude-based weight pruning is a technique that gradually sets certain weights in a neural network to zero during training to get the model sparsity. By compressing the model in this way, it becomes easier to store and execute, resulting in faster inference and better memory utilization. The technique has shown promise in a variety of applications, including speech recognition, medical diagnosis, and disease detection, as well as in vision and translation models. Layer-wise magnitude-based pruning (LMP) is a productive compression method that can prune connections in parallel, which is particularly useful for DNNs with a large number of connections. By setting layer-specific thresholds, LMP can achieve significant reductions in the number of parameters with only a relatively small loss of accuracy. Fig. 5 shows the configuration of the magnitude based pruning model [12].

```

Model: "sequential_2"
-----
Layer (type)                Output Shape                Param #
-----
prune_low_magnitude_reshape_2 (PruneLowMagnitude)  (None, 28, 28, 1)          1
prune_low_magnitude_conv2d_2 (PruneLowMagnitude)      (None, 26, 26, 12)        230
prune_low_magnitude_max_pooling2d_2 (PruneLowMagnitude)  (None, 13, 13, 12)        1
prune_low_magnitude_flatten_2 (PruneLowMagnitude)      (None, 2028)               1
prune_low_magnitude_dense_2 (PruneLowMagnitude)      (None, 10)                 40572
-----
Total params: 40,805
Trainable params: 20,410
Non-trainable params: 20,395

```

Fig. 5. Configuration of Magnitude Based Pruning Model

5. Model Compression and Optimization

Model size can be reduced using quantization by providing the less storage space, smaller bandwidth, less memory consumption at the cost of minor accuracy loss.

Less Storage Space: On devices small network model occupies less storage space like android applications takes smaller storage on mobile devices.

Smaller Bandwidth: Small DL models take comparatively less time and bandwidth to download on client devices.

- **Less Memory Consumption:** Small models on execution takes less RAM when, that can free up memory for other applications, and can give better performance and efficiency.

The amount of time taken to run a model with single inference is called the latency. Quantization can be employed to reduce the latency by simplifying the computation task during inference.

Post-training quantization task incorporates with some mechanism to reduce hardware and CPU latency, processing speed, and model size with small degradation in accuracy of the model. In our proposed work model weights can be converted to 8-bit integer for execution on CPU. Dynamic range quantization is a used by this work as it leads to less memory consumption and quick calculations task without representative dataset for calibration. It is quantizes the weights from floating point to integer during conversion time, that results into 8-bits of precision statically [13]. Fig. 6 represents the process of quantization and optimization after the applying the magnitude based pruning.

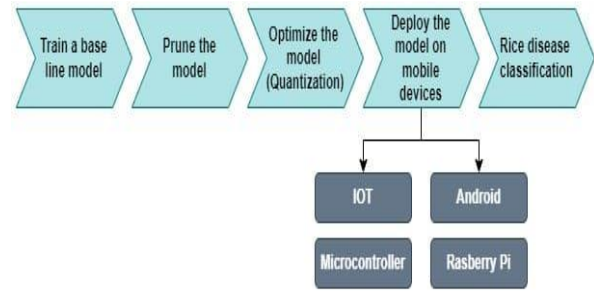


Fig. 6. Process of Quantization and Optimization

6. Result Analysis

The 80-20 cross-validation function is employed to evaluate and calculate the effectiveness of the model. The cross-entropy loss function is utilized to assess the model's efficiency. Adam optimizer is used to improve the cross-entropy ratio [14]. The Confusion matrix in table 2 captures the results of the proposed approach on the rice plant image dataset. Confusion matrix generated for four classes of the rice plant in which three classes belongs to infected class of rice plant diseases i.e. brown spot, leaf blast, hispa, and one class belongs to healthy plant leaves.

Table 2: Confusion Matrix Of Multi-Classification using CNN Model

Classes	Brown	Hispa	Leaf	Healthy
Brown Spot	691	23	32	26
Hispa	18	722	17	23
Leaf Blast	14	21	1012	18
Healthy	9	25	29	1597

The proposed work is evaluated on different parameters like precision (P), Recall (R), F1 score (F) and accuracy (Acc). Table 3 shows the final outcome of the multi-classification model.

$$P = \left[\frac{Tp}{Tp+FP} \right] \quad (2)$$

$$R = \left[\frac{Tp}{Tp+FN} \right] \quad (3)$$

$$F = \left[\frac{(2*PC*RC)}{PC+RC} \right] \quad (4)$$

$$Acc = \left[\frac{Tp+TN}{Tp+FP+FN+TN} \right] \quad (5)$$

Where Tp- True positive, Fp- False positive, TN- True Negative, FN- False Negative

Table 3: Evaluation Results of Multi-Classification Model

Classes	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Brown Spot	97.15	90	94	92
Hispa	97.03	93	91	92
Leaf Blast	96.94	95	93	94
Healthy	96.96	96	96	96

Fig. 7 represents the accuracy graph of baseline CNN Model generated during the training and validation. Fig. 8 represents the loss graph of baseline CNN Model generated over the training and validation phase. Similarly, Figure 9 represents the accuracy graph of pruning model generated over the training and validation process. Fig. 10 represents the loss graph of pruning model generated over the training and validation process.

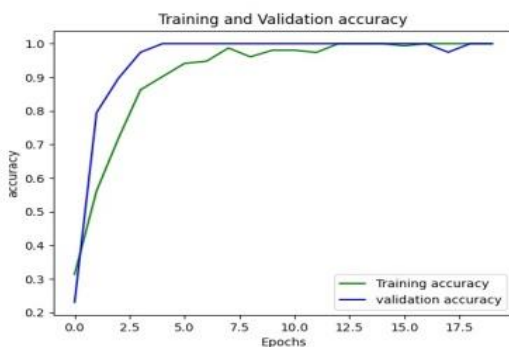


Fig. 7. Training and Validation Accuracy Graph of Baseline CNN Model

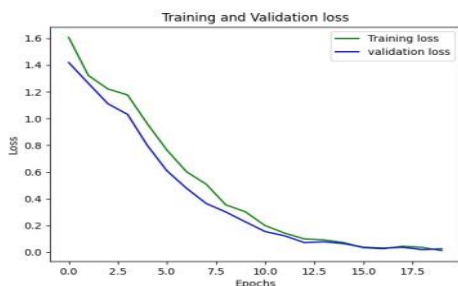


Fig. 8. Training and Validation Loss Graph of Baseline CNN Model

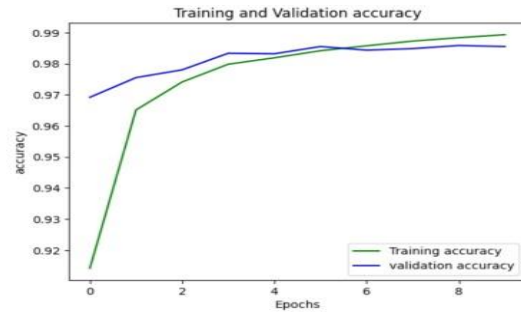


Fig. 9. Training and Validation Accuracy Graph of Pruning Model

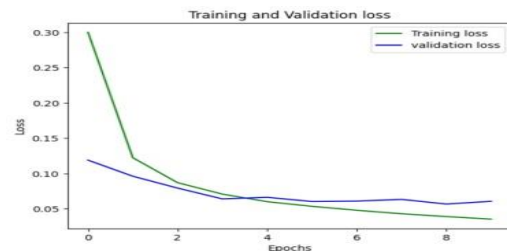


Fig. 10. Training and Validation Loss Graph of Pruning Model

Table 4 represents the test and validation accuracy, loss and the size of the models. It can be observed from the table 4 that test accuracy is baseline model is higher than the pruning model but the size of the pruning model is less as compared to the CNN model without compromising much accuracy. Test and validation loss is less in case of base line CNN model which is 0.0976 and 0.0984 respectively. It can be observed from the table 4 that test loss obtained by pruning model and quantized respectively. Dynamic range quantization provides lowest model size of 21.188 MB with slightly compromising the accuracy i.e. 96.24%.

The major applicability of the work:

- (i) The suggested work can help to recognize rice leaves disorders like brown spot, hispa , leaf blast for low resource and small scale mobile devices.
- (ii) Magnitude based pruning along with compression technique can able to get more light weight model which can help the farmers to detect the disorders on field using mobile devices in less time.
- (iii) This work help in developing light weight model for the edge devices with modest means in terms of space and computational ability.
- (iv) This work also helps in developing the model with low latency which makes faster inference for disease classification of rice plants regardless of network connection.

Table 4: Evaluation Outcomes of the Models

Model	Test Accuracy (%)	Validation Accuracy (%)	Test Loss (%)	Validation Loss (%)	Size (MB)
Baseline CNN Model	98.11	97.13	0.0976	0.081	78.24
Pruned Model	97.39	96.56	0.123	0.984	25.743
Dynamic Range Quantization	96.24	96.02	0.162	0.131	21.188

7. Conclusion

In this paper, optimized CNN Model is employed for lightweight rice plant disease image classification. The magnitude based pruning is applied to the CNN classifier to get the model with smaller size without compromising the accuracy. Overall accuracy achieved by the baseline model is 98.11% with model size of 78.24 (MB). Magnitude based pruning model can reduced the model size up to 3 times with the slight change in performance in terms of accuracy. The accuracy reduced by the pruning model is from 98.11% to 97.39%. Further pruning model is optimized using dynamic range quantization in order to get more light weight model which can provide less latency, less time and less memory consumption by simplifying the computation task during inference. The Post-training quantization task can be able to further reduce by 17.69% of the pruned model size. In future; models can be trained using edge computing, cloud and fog computing environment for faster analysis. Also, our technique can be applied on more rice diseases with supplementary data samples.

References

- [1] Dhiman, P. (2014, September). Empirical validation of website quality using statistical and machine learning methods. In *2014 5th International Conference-Confluence The Next Generation Information Technology Summit (Confluence)* (pp. 286-291). IEEE.
- [2] Arumuga Arun, R., & Umamaheswari, S. (2023). Effective multi-crop disease detection using pruned complete concatenated deep learning model.
- [3] Zhang, S., Wu, G., Gu, J., & Han, J. (2020). Pruning convolutional neural networks with an attention mechanism for remote sensing image classification. *Electronics*, 9(8), 1209.
- [4] Joshi, P., Das, D., Udutalapally, V., Pradhan, M. K., & Misra, S. (2022). Ricebios: Identification of biotic stress in rice crops using edge-as-a-service. *IEEE Sensors Journal*, 22(5), 4616-4624.
- [5] Chen, R., Qi, H., Liang, Y., & Yang, M. (2022). Identification of plant leaf diseases by deep learning based on channel attention and channel pruning. *Frontiers in Plant Science*, 13.
- [6] Arumuga Arun, R., & Umamaheswari, S. (2023). Effective multi-crop disease detection using pruned complete concatenated deep learning model.
- [7] Sowmiya, B., Saminathan, K., & Devi, M. C. Classification of paddy leaf diseases with extended huber loss function using convolutional neural networks.
- [8] Laut, S., Poapolathep, S., Piasai, O., Sommai, S., Boonyuen, N., Giorgi, M., ... & Poapolathep, A. (2023). Storage Fungi and Mycotoxins Associated with Rice Samples Commercialized in Thailand. *Foods*, 12(3), 487.
- [9] Dhiman, P., Kukreja, V., & Kaur, A. (2021, September). Citrus Fruits Classification and Evaluation using Deep Convolution Neural Networks: An Input Layer Resizing Approach. In *2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)* (pp. 1-4). IEEE.
- [10] Geng, Z., Xu, Y., Wang, B. N., Yu, X., Zhu, D. Y., & Zhang, G. (2023). Target Recognition in SAR Images by Deep Learning with Training Data Augmentation. *Sensors*, 23(2), 941.
- [11] Dhiman, P., Kaur, A., Hamid, Y., Alabdulkreem, E., Elmannai, H., & Ababneh, N. (2023). Smart Disease Detection System for Citrus Fruits Using Deep Learning with Edge Computing. *Sustainability*, 15(5), 4576.
- [12] Ghimire, D., & Kim, S. H. (2023). Magnitude and Similarity Based Variable Rate Filter Pruning for Efficient Convolution Neural Networks. *Applied Sciences*, 13(1), 316.
- [13] Dhiman, P., Poongodi, M., Lilhore, U. K., AlQahtani, S. A., Kaur, A., Iwendi, C., & Raahemifar, K. (2023).

PFDI: A Precise Fruit disease Identification Model based on Context Data Fusion with Faster-CNN in Edge Computing Environment.

- [14] Alsubai, S., Dutta, A. K., Alkhayat, A. H., Jaber, M. M., Abbas, A. H., & Kumar, A. (2023). Hybrid deep learning with improved Salp swarm optimization based multi-class grape disease classification model. *Computers and Electrical Engineering*, *108*, 108733.
- [15] P. Seelwal and A. Sharma, "Machine Vision Systems for Rice Diseases Detection: A Review," 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2022, pp. 1686-1689, doi: 10.1109/ICACITE53722.2022.9823713.
- [16] Prema, K. ., & J, V. . (2023). A Novel Marine Predators Optimization based Deep Neural Network for Quality and Shelf-Life Prediction of Shrimp. *International Journal on Recent and Innovation Trends in Computing and Communication*, *11(3s)*, 65–72. <https://doi.org/10.17762/ijritcc.v11i3s.6156>
- [17] Raj, R., & Sahoo, D. S. S. . (2021). Detection of Botnet Using Deep Learning Architecture Using Chrome 23 Pattern with IOT. *Research Journal of Computer Systems and Engineering*, *2(2)*, 38:44. Retrieved from <https://technicaljournals.org/RJCSE/index.php/journal/article/view/31>