JURNAL EMACS
e-ISSN: 2686-2573
[Engineering, MAthematics and Computer Science] Vol.6 No.3 September 2024: 179-184 DOI: 10.21512/emacsjournal.v6i3.11480

Overcoming Overfitting in CNN Models for Potato Disease Classification Using Data Augmentation

Simeon Yuda Prasetyo

Computer Science Department, School of Computer Science,
Bina Nusantara University,
Jakarta, Indonesia 11480
simeon.prasetyo@binus.ac.id

Correspondence: simeon.prasetyo@binus.ac.id

Abstract *– Classification of diseases in potato plants is crucial for agriculture to ensure quality and yield. Potatoes being staple foods worldwide are vulnerable to diseases that cause significant production losses. Early and accurate disease identification is essential. This study evaluates the impact of data augmentation on reducing overfitting in deep learning models for potato disease classification. Various CNN architectures including VGG16, VGG19, Xception, and InceptionV3 were compared in transfer learning and fine-tuning phases. The "Potato Disease Dataset" consisting of 451 images across seven classes was used. The dataset was split into training, validation, and test sets, and augmentation increased the training set from 360 to 2160 images. The results indicate that models trained with augmented data exhibited improved performance in terms of accuracy, precision, recall, and F1-scores compared to those trained without augmentation. The learning curves show that data augmentation helps in reducing overfitting and enhancing model stability. Data augmentation is crucial for developing robust deep learning models for potato disease classification. This study also highlights the potential of data augmentation in addressing the challenges posed by small datasets in agricultural applications. The findings contribute to the growing body of research in applying deep learning techniques for more sustainable and efficient disease management in crops. Future work will explore advanced augmentation techniques and other architectures to enhance model performance.*

Keywords: *Agricultural Disease Management; Convolutional Neural Networks (CNN); Data Augmentation; Deep Learning; Potato Disease Classification*

I. INTRODUCTION

Disease classification in potato plants is an important part of agriculture that ensures crop yields are both high and consistent. Potatoes are a basic item consumed worldwide and are sensitive to a variety of illnesses, which can result in severe production losses and economic issues for farmers. Thus, early and effective disease detection is critical for sensitive crops such as potatoes (Hamza et al., 2022).

Diseases in potato plants can lead to severe reductions in yield and quality, affecting both the economic stability of farmers and the food supply chain. Key diseases include late blight, early blight, black scurf, and powdery scab, which can spread rapidly and devastate crops if not identified and managed promptly (Sharma et al., 2023). The conventional methods of disease identification often rely on manual inspection, which is time-consuming, prone to errors, and requires significant expertise (Moawad et al., 2023).

Deep learning algorithms have shown significant promise in image-based illness classification applications in recent years. Convolutional neural networks (CNN)-based studies have shown great accuracy in recognizing several potato illnesses. For example, a CNN model accurately classified potato illnesses into three categories: early blight, late blight, and healthy (Shobanadevi et al., n.d.). Another study employing a DenseNet201 model obtained an accuracy of 99% in diagnosing potato illnesses in tubers, including black scarf and green tuber (Moawad et al., 2023).

Hybrid techniques have been used to improve the accuracy of potato disease categorization. The combination of CNN with long short-term memory (LSTM) networks

and optimization approaches such as Adaptive Shark Smell Optimization (ASSO) has proven useful in identifying illnesses with an accuracy of up to 99.02% (M. A. Patil & M, 2023). Additionally, a study employing hyper-parameter tuning and deep learning achieved an accuracy of 99.42% (Sharma et al., 2023).

Sholihati et al. (2020) classified potato diseases with a 91% accuracy rate using the VGG16 and VGG19 architectures in another study. This suggests that the deep neural network approach to disease classification is feasible for four classes of diseases: early blight, late blight, black scurf, and healthy (Sholihati et al., 2020).

Additionally, transfer learning techniques have been used to enhance deep learning models' classification performance for potato diseases. A study by Thangaraj et al. (2020) used a modified Xception model, achieving an accuracy of 98.16% for five classes: early blight, late blight, black scurf, powdery scab, and healthy) (Thangaraj et al., 2020). Similarly, Charisma & Adhinata (2023) utilized the DenseNet201 architecture, achieving an accuracy of 92.5% for multiple classes of potato diseases (Charisma & Dharma Adhinata, 2023).

Additionally, a study using EfficientNet-V2 architecture achieved an accuracy of 98.12% in classifying various potato diseases, showcasing the potential of optimized deep learning models in agricultural applications (Nazir et al., 2023).

Further, the application of MobileNet architecture in potato disease detection achieved an accuracy of 99.83%, highlighting the efficiency of lightweight models for real-time applications (Mishra et al., 2021).

Finally, Hamza et al. (2022) reviewed numerous deep learning algorithms, including VGG19, VGG16, Google Net, and Alex Net, for potato disease identification, stressing the continual developments and use of deep learning in agriculture across multiple classes of illnesses (Hamza et al., 2022).

These studies demonstrate the effectiveness of deep learning methods in enhancing the accuracy and efficiency of potato disease classification, thereby aiding in better disease management and reducing economic losses in agriculture.

The purpose of this study is to assess the influence of data augmentation on minimizing overfitting in deep learning models used to classify potato illnesses. Comparing the performance of several convolutional neural network (CNN) designs, such as VGG16, VGG19, Xception, and InceptionV3, in both transfer learning (TL) and fine-tuning (FT) phases, this research seeks to identify optimal augmentation strategies that enhance model generalization and accuracy. Specifically, the study aims to determine the effectiveness of data augmentation techniques in reducing overfitting and improving the robustness of these models in accurately identifying key potato diseases, including black scurf, blackleg, common scab, dry rot, healthy potatoes, miscellaneous, and pink rot, thereby contributing to more reliable disease management in agriculture.

II. METHODS

The methodology for this study consists of several key stages, as illustrated in Figure 1. The process begins with data gathering, where images of potato diseases are collected from various sources. These images represent different classes, including Black Scurf, Blackleg, Common Scab, Dry Rot, and Healthy potatoes. Once the data is collected, the images undergo preprocessing to enhance their quality and ensure consistency for model training. Preprocessing steps may include resizing the images, normalizing pixel values, and removing noise.

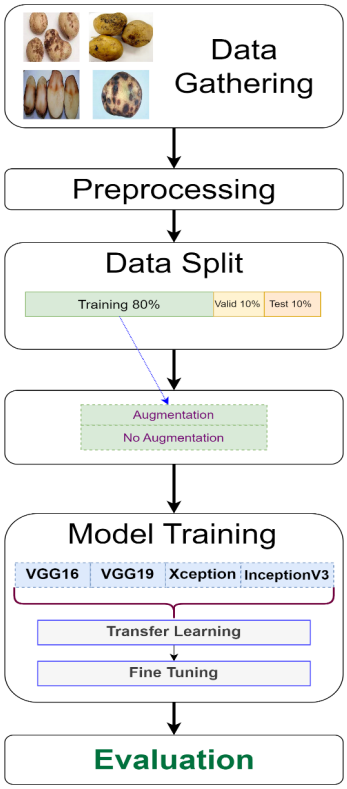

Figure 1. Experimental Design

The preprocessed data is divided into three subsets: 80% for training, 10% for validation, and 10% for testing. The training set is used to train the model, the validation set to tweak the parameters, and the test set to evaluate the model's performance. Data augmentation methods are used on the training set to increase the model's resilience and stop overfitting. These augmentation methods, such as rotations, shifts, flips, and zooms, generate additional training data by applying random transformations, thus increasing the diversity of the training set.

The model training phase consists of two parallel paths: one with enhanced data and one without. Training takes place using a variety of convolutional neural network (CNN) designs, including VGG16, VGG19, Xception, and InceptionV3. Each model goes through two phases: transfer learning and fine-tuning. During the transfer learning phase, the model is initialized with pre-learned weights from big datasets before being trained on the potato disease dataset. During the fine-tuning step, the whole model, including the pre-trained layers, is tweaked to enhance performance on the specific goal of potato disease classification.

Finally, the models are assessed on the test set using measures such as accuracy, precision, recall, and F1-score. The outcomes of the augmented and non-augmented training routes are compared to assess the efficacy of data augmentation and the best performing model architecture and training method. This comprehensive methodology, depicted in Figure 1, ensures a robust and reliable model for classifying potato diseases.

2.1 Potato Disease Dataset

The dataset used in this study, named "Potato Disease Dataset," was sourced from Kaggle (*Kaggle*, n.d.). This dataset comprises images of potato tubers affected by various diseases, as well as images of healthy potatoes. The dataset is categorized into seven classes, with a total of 451 images. The classes include Common Scab with 62 images caused by bacteria, Blackleg with 60 images also caused by bacteria, Dry Rot with 60 images caused by fungus, Pink Rot with 57 images caused by fungus, and Black Scurf with 58 images caused by fungus. Additionally, there are 80 images of healthy potatoes and 74 images categorized as miscellaneous.

To prepare the data for model training and assessment, it was divided into three subsets: 80% for training, 10% for validation, and 10% for testing. This divide guarantees that the model has enough data for training while maintaining distinct validation and test sets for unbiased model adjustment and performance evaluation. For the augmented data path, data augmentation techniques were applied to the original training set. The original training set of 360 photos was enlarged to 2160 images via augmentation. This was accomplished by performing numerous modifications to the original photos, including rotations, shifts, flips, and zooms, which increased the variety of the training data and helped to minimize overfitting.

In conclusion, Kaggle's "Potato Disease Dataset" offers a broad collection of pictures for training, validating, and testing models for potato illness classification. The use of data augmentation strengthens the training process, resulting in the construction of trustworthy and accurate models for detecting potato illnesses. Figure 2 shows examples of dataset pictures.

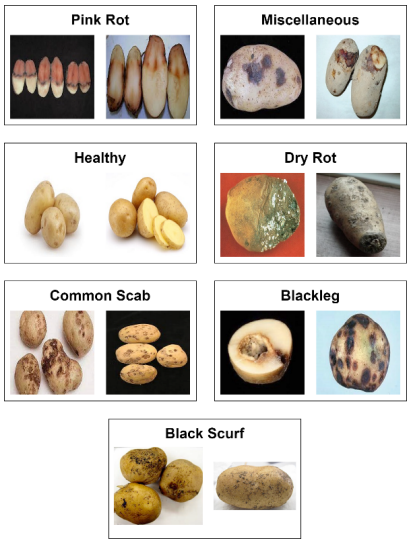

Figure 2. Potato Disease Dataset

2.2 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are a type of deep learning algorithm used largely to analyze visual input. They are especially useful for image classification problems because they can automatically and adaptively learn spatial hierarchies of features from input photos. A CNN's fundamental architecture consists of numerous layers, including convolutional layers, pooling layers, and fully connected layers, that work together to extract and learn characteristics from pictures.

VGG16 and VGG19 are two popular CNN architectures created by the Visual Geometry Group (VGG) at the University of Oxford. VGG16 and VGG19 have 16 and 19 layers, respectively. Both models rely on modest 3x3 convolution filters and are well-known for their simplicity and success in a variety of picture classification applications. A study by Sinha and Lalit (2021) compared VGG16 and VGG19 with other architectures such as ResNet50 and InceptionV3 for vision-based security systems, highlighting the computational expense of VGG models due to their large number of parameters (Sinha & Lalit, 2022).

Xception is another powerful CNN architecture that improves upon the Inception model by replacing the standard Inception modules with depthwise separable convolutions. This results in a more efficient model with fewer parameters and potentially better performance. A study by Patil et al. (2020) used Xception, along with other models like VGG16 and VGG19, to classify lung diseases from X-ray images, demonstrating its effectiveness in medical image analysis (N. Patil et al., 2020).

Inception, also known as GoogLeNet, is a deep CNN architecture designed to perform well even with fewer computational resources. It employs a combination of 1x1, 3x3, and 5x5 convolutional filters to capture various spatial features and reduce the number of parameters. In a study by Datt and Kukreja (2022), the Inception-v3 model was used for recognizing different phenological stages of apple crops, achieving high accuracy in comparison to other models like VGG16, ResNet50, and Xception (Datt & Kukreja, 2022).

In CNN applications, transfer learning is a popular technique used to fine-tune models that have already been pre-trained on huge datasets for certain tasks. Performance is greatly enhanced by this method, particularly when there is a shortage of data for the target task. Abdulsattar and Hussain (2022) highlighted the effectiveness of transfer learning and fine-tuning strategies using popular architectures such as VGG16, VGG19, InceptionV3, and Xception for facial expression recognition, demonstrating the adaptability and robustness of these models in different domains (Abdulsattar & Hussain, 2022).

2.3 Evaluation Metrics

Evaluation metrics such as accuracy, precision, recall, and F1-score are critical for measuring classification model performance. Accuracy quantifies the proportion of correct predictions among all forecasts made and is appropriate for balanced datasets. However, it may not be sufficient for unbalanced datasets since it might be deceptive. Precision (Positive Predictive Value) is defined as

the ratio of properly predicted positive observations to total expected positives, highlighting the significance of relevant outcomes. Recall (Sensitivity) is the ratio of accurately predicted positive observations to all observations in the actual class, demonstrating the capability to catch all relevant occurrences (Riyanto et al., n.d.).

F1-score is the harmonic mean of accuracy and recall, which balances the two measures and is especially effective for unbalanced datasets. This score is especially beneficial when we seek a balance between precision and recall (Sitarz, 2022). The equations for these metrics are as follows:

$$\textbf{Accuracy} = TP+TN/TP+TN+FP+FN \quad (1)$$

$$\textbf{Precision} = TP/TP+FP \quad (2)$$

$$\textbf{Recall} = TP/TP+FN \quad (3)$$

$$\textbf{F1-Score} = 2TP/2TP+FN+FP \quad (4)$$

Where, ***TP*** is True Positives, ***TN*** is True Negatives, ***FP*** is False Positives, and ***FN*** is False Negatives. These measures assist give a complete evaluation of model performance, especially in cases when class imbalance exists (Yacouby & Axman, 2020).

III. RESULTS AND DISCUSSION

The experimental environment utilized for this study was Google Colab Pro, leveraging the powerful computational capabilities of the V100 GPU and 25GB of RAM. This setup ensured efficient handling of the large dataset and complex computations required for training the deep learning models. The use of Google Colab Pro provided a robust platform to perform extensive data augmentation and train multiple convolutional neural network (CNN) architectures, including VGG16, VGG19, Xception, and InceptionV3.

The experiment findings, reported in Table 1, show the performance of models trained with and without data augmentation. The assessment parameters comprised accuracy, precision, recall, and F1-score. These metrics were calculated for both the transfer learning and fine-tuning stages of each model. The primary objective was to assess the impact of data augmentation on model performance and determine the optimal strategy for classifying potato diseases. Detailed experimental results can be found in Table I.

Table II. Experimental Results

Model	Phase	Metrics	No Aug	With Aug
VGG16	TL	Accuracy	73.33	62.22
		Precision	72.33	63.70
		Recall	73.33	62.22
		F1-Score	72.03	60.81
VGG16	FT	Accuracy	73.33	**75.56**
		Precision	74.93	69.07
		Recall	73.33	75.56
		F1-Score	73.15	71.75
VGG19	TL	Accuracy	57.78	66.67
		Precision	62.30	67.55
		Recall	57.77	66.67
		F1-Score	58.46	66.72
VGG19	FT	Accuracy	55.56	75.56
		Precision	56.24	74.88
		Recall	55.56	75.56
		F1-Score	55.18	74,04
Xception	TL	Accuracy	73.33	60.00
		Precision	78.86	74.58
		Recall	73.33	60.00
		F1-Score	71.99	56.98
Xception	FT	Accuracy	**77.78**	68.89
		Precision	78.22	71.74
		Recall	77.78	68.89
		F1-Score	77.05	67.22
InceptionV3	TL	Accuracy	77.78	57.78
		Precision	80.54	59.64
		Recall	77.78	57.78
		F1-Score	77.98	55.62
InceptionV3	FT	Accuracy	62.22	60.00
		Precision	64.7	62.43
		Recall	62.22	60.00
		F1-Score	61.60	58.20

The results in Table I indicate varying impacts of data augmentation on the performance of different CNN architectures.

For the VGG16 model, during the transfer learning phase, the model without augmentation showed better performance across all metrics, with an accuracy of 73.33%, precision of 72.33%, recall of 73.33%, and F1-score of 72.03%. However, during the fine-tuning phase, the model with augmentation performed better in terms of accuracy (75.56%) and recall (75.56%), although precision and F1-score were slightly lower compared to the model without augmentation. The learning curves for VGG16, as shown in Figure 3, illustrate that the model with augmentation has a smoother training process and less overfitting compared to the model without augmentation.

For the VGG19 model, a similar pattern was observed. During the transfer learning phase, the model with augmentation significantly outperformed the model without augmentation, with an accuracy of 66.67%, precision of 67.55%, recall of 66.67%, and F1-score of 66.72%. In the fine-tuning phase, the model with augmentation continued to show superior performance, highlighting the benefits of data augmentation in improving the generalization of the model.

The Xception model showed notable differences as well. Without augmentation, the model achieved an accuracy of 73.33%, precision of 78.86%, recall of 73.33%, and F1-score of 71.99% during the transfer learning phase. With augmentation, these metrics dropped, indicating that the model struggled with the increased data variability. However, during the fine-tuning phase, the model with

augmentation showed improved performance, particularly in precision and recall. The learning curves for Xception, as presented in Figure 4, demonstrate that data augmentation helped mitigate overfitting, leading to a more stable and generalizable model.

For the InceptionV3 model, the performance was better without augmentation during the transfer learning phase, with an accuracy of 77.78%, precision of 80.54%, recall of 77.78%, and F1-score of 77.98%. However, in the fine-tuning phase, the differences between the models with and without augmentation were less pronounced, indicating that the inherent architecture of InceptionV3 may already possess strong generalization capabilities.

Overall, the results indicate that data augmentation generally helps reduce overfitting and improves the robustness of the models, although the extent of improvement varies across different architectures. The learning curves and evaluation metrics underscore the importance of data augmentation in training deep learning models for image classification tasks.

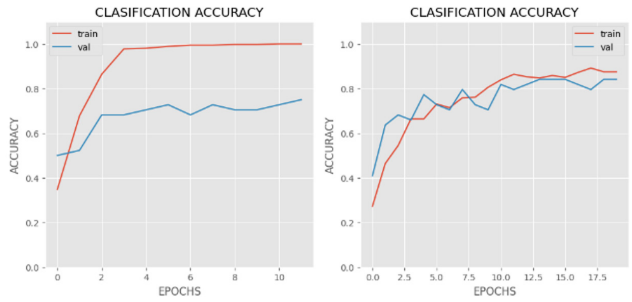

Figure 3. Comparison of Learning Curves for VGG16 FT Phase Without and With Augmentation

The learning curves for VGG16 (Figure 3) show that the model with augmentation has a more stable training process and less overfitting compared to the model trained without augmentation.

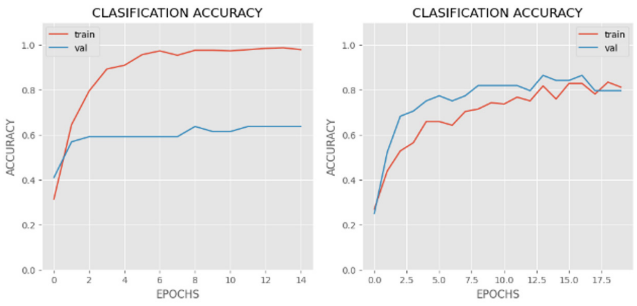

Figure 4. Comparison of Learning Curves for Xception FT Phase Without and With Augmentation

The learning curves for Xception (Figure 4) illustrate that data augmentation helped in reducing overfitting, leading to a more generalizable model.

In summary, the application of data augmentation techniques has shown to be beneficial in improving the generalization and robustness of deep learning models for potato disease classification, as evidenced by the results and learning curves presented.

IV. CONCLUSION

In conclusion, the results of this study demonstrate that data augmentation plays a crucial role in enhancing the performance of deep learning models for potato disease classification. By applying various augmentation techniques, the models trained on augmented data exhibited reduced overfitting and improved generalization, as evidenced by higher accuracy, precision, recall, and F1-scores in several cases. Specifically, models like VGG19 and Xception showed significant improvements in performance metrics when trained with augmented data, highlighting the effectiveness of data augmentation in creating more robust and reliable models. Overall, this study underscores the importance of employing data augmentation to develop accurate and generalizable deep learning models for agricultural applications.

For future works, several directions can be explored to further enhance the performance and applicability of deep learning models for potato disease classification. One potential avenue is to investigate the use of advanced data augmentation techniques and synthetic data generation to create even more diverse training datasets. Additionally, exploring other state-of-the-art architectures and ensemble methods could yield improvements in classification accuracy and robustness. Incorporating explainable AI techniques to interpret model predictions and provide insights into the decision-making process can also be valuable for gaining trust from agricultural practitioners. Moreover, expanding the dataset to include more diverse conditions and different stages of disease progression could help improve model generalization. Finally, deploying these models in real-time systems and validating their performance in practical field conditions would be crucial steps towards integrating these solutions into agricultural disease management practices.

REFERENCES

Abdulsattar, N. S., & Hussain, M. N. (2022). Facial Expression Recognition using Transfer Learning and Fine-tuning Strategies: A Comparative Study. *2022 International Conference on Computer Science and Software Engineering (CSASE)*, 101–106. https://doi.org/10.1109/CSASE51777.2022.9759754

Charisma, R. A., & Dharma Adhinata, F. (2023). Transfer Learning With Densenet201 Architecture Model For Potato Leaf Disease Classification. *2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE)*, 738–743. https://doi.org/10.1109/ICCoSITE57641.2023.10127772

Datt, R. M., & Kukreja, V. (2022). Phenological Stage Recognition Model for Apple Crops using transfer learning. *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, 1537–1542. https://

doi.org/10.1109/ICACITE53722.2022.9823711

Hamza, K., Un Nisa, S., & Irshad, G. (2022). A Review on Potato Disease Detection and Classification by exploiting Deep Learning Techniques. *J. Agri. Vet. Sci*, *01*(2), 2022–2079. https://doi.org/10.55627/Javs.01.2.0251

Kaggle. (n.d.). https://www.kaggle.com/datasets

Mishra, S., Singh, A., & Singh, V. (2021). Application of MobileNet-v1 for Potato Plant Disease Detection Using Transfer Learning. *2021 Workshop on Algorithm and Big Data*, 14–19. https://doi.org/10.1145/3456389.3456403

Moawad, N., Zaki, H., El Moniem Essa, T. abed, & Said, M. (2023). Detection of Potato Tuber Diseases Using Machine Learning Models. *2023 International Conference on Artificial Intelligence Science and Applications in Industry and Society (CAISAIS)*, 1–7. https://doi.org/10.1109/CAISAIS59399.2023.10269994

Nazir, T., Iqbal, M. M., Jabbar, S., Hussain, A., & Albathan, M. (2023). EfficientPNet—An Optimized and Efficient Deep Learning Approach for Classifying Disease of Potato Plant Leaves. *Agriculture*, *13*(4). https://doi.org/10.3390/agriculture13040841

Patil, M. A., & M, M. (2023). Potato Leaf Disease Identification using Hybrid Deep Learning Model. *2023 International Conference on Network, Multimedia and Information Technology (NMITCON)*, 1–9. https://doi.org/10.1109/NMITCON58196.2023.10276091

Patil, N., Ingole, K., & Rajani Mangala, T. (2020). Deep convolutional neural networks approach for classification of lung diseases using x-rays: Covid-19, pneumonia, and tuberculosis. *International Journal of Performability Engineering*, *16*(9), 1332–1340. https://doi.org/10.23940/ijpe.20.09.p2.13321340

Riyanto, S., Sitanggang, I. S., Djatna, T., & Atikah, T. D. (n.d.). Comparative Analysis using Various Performance Metrics in Imbalanced Data for Multi-class Text Classification. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 14, Issue 6). http://gcancer.org/pdr

Sharma, O., Rajgaurang, Mohapatra, S., Mohanty, J., Dhiman, P., & Bonkra, A. (2023). Predicting Agriculture Leaf Diseases (Potato): An Automated Approach using Hyper-parameter Tuning and Deep Learning. *2023 Third International Conference on Secure Cyber Computing and Communication (ICSCCC)*, 490–493. https://doi.org/10.1109/ICSCCC58608.2023.10176819

Shobanadevi, A., Shanthini, A., & Hsu, H.-C. (n.d.). *Potato Leaf Disease Classification using Deep Learning : A Convolutional Neural Network Approach*.

Sholihati, R. A., Sulistijono, I. A., Risnumawan, A., & Kusumawati, E. (2020). Potato Leaf Disease Classification Using Deep Learning Approach. *2020 International Electronics Symposium (IES)*, 392–397. https://doi.org/10.1109/IES50839.2020.9231784

Sinha, K., & Lalit, M. (2022). Comparative Analysis of Pre-Trained Deep Neural Networks for Vision-Based Security Systems on a Novel Dataset. *Proceedings of the 2021 IEEE/ACM 8th International Conference on Big Data Computing, Applications and Technologies*, 120–127. https://doi.org/10.1145/3492324.3494173

Sitarz, M. (2022). *Extending F1 metric, probabilistic approach*. https://doi.org/10.54364/AAIML.2023.1161

Thangaraj, R., P, P., Kaliappan, V. K., S, A., & P, I. (2020). Potato Leaf Disease Classification using Transfer Learning based Modified Xception Model. *Innovations in Information and Communication Technology Series*. https://api.semanticscholar.org/CorpusID:231783905

Yacouby, R., & Axman, D. (2020). Probabilistic Extension of Precision, Recall, and F1 Score for More Thorough Evaluation of Classification Models. In S. Eger, Y. Gao, M. Peyrard, W. Zhao, & E. Hovy (Eds.), *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems* (pp. 79–91). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.eval4nlp-1.9