

Classification of Sri Lankan Paddy Varieties using Deep Learning Techniques

M.R.M. Aththas¹, M.N.F. Yusra², M.S. Sabrina³, W.A. Sanjeewa⁴, M. Janotheepan⁵ and A.R. Fathima Shafana⁶

^{1,2,3,4,5}Department of Computer Science and Informatics, Uva Wellassa University, Sri Lanka

⁶Department of Information and Communication Technology, South Eastern University of Sri Lanka, Sri Lanka

¹cst19030@std.uwu.ac.lk, ²cst19052@std.uwu.ac.lk, ³jit19062@std.uwu.ac.lk, ⁴aruna.s@uwu.ac.lk,

⁵janotheepan.m@uwu.ac.lk, ⁶arfshafana@seu.ac.lk

Abstract

Rice is a highly consumed staple food in Sri Lanka. From farming phase to distribution phase of paddy, classification of paddy is becoming vital as it provides efficiency to the planning, production, sales and consumption. In Sri Lanka, the evaluation of the classification of paddy varieties is typically overseen by the Rice Research and Development Institute (RRDI). Traditionally, paddy identification is done manually by human inspectors, ensuring some level of accuracy but requiring significant manpower, time, and subjective judgment. This research seeks to transform the categorization of paddy varieties in Sri Lanka. This paper provides an approach to identifying and classifying paddy variety in paddy sample with the help of image processing and CNN model. For this approach, 10 varieties of paddy samples were collected from Rice Research and Development Institute. With these samples a dataset of more than 10,000 images were captured and used in this research. Image preprocessing involved cropping, scaling, and noise removal to standardize the data. Experiments were conducted with nine different CNN models, iterating through various architectures and training parameters to optimize performance. The experiment was performed on ten rice categories to evaluate the suggested solution. The accuracy of classification is of 93.69%.

Keywords: Convolutional Neural Network (CNN), Deep Learning, Paddy Classification

I. INTRODUCTION

Rice is favorable and highly consumed food in Sri Lanka. Rice is one of the highly consumed and staple food of Sri Lanka. Around 3.1 million tons of rough rice (paddy) are grown every year to meet about 95% of the country's demand. Since more than 1.8 million farmers and their families depend on rice production, rice holds a unique importance

compared to other agricultural products in Sri Lanka (Anon, n.d.).

The accuracy of identifying paddy is one of the most important factors when classifying rice varieties. The use of paddy varieties differs depending on the purposes. Different varieties of rice are used for the production of many value-added products, including food varieties. Therefore, rice variety identification is very important for consumers (Golpour et al., 2014). In addition, the price and grade of rice are decided by its commercial value, genetic characteristics, and quality factors, which depend on the type of rice variety.

Currently, the classification of paddy is performed manually, typically through visual inspection by experienced and well-trained individuals. However, this approach has significant drawbacks, including time consumption and unreliability due to inconsistencies and the involvement of unskilled technicians. Moreover, results may vary from person to person leading to subjective results. Therefore, there is a pressing need for a more efficient and accurate method of paddy variety classification. The review of literature shows that both Machine Learning (ML) and Computer Vision (CV) have been extensively employed across various domains, offering a fast, accurate, nondestructive, and cost-effective substitute for automated paddy classification processes. Deep learning techniques have superseded statistical methods in computer vision due to their enhanced accuracy in tasks such as object identification and image recognition (Kiratiratanapruk et al., 2020).

While research on the classification of paddy varieties is limited, the identification of rice varieties has been extensively studied using external parameters such as shape, size, color, and texture (Cinar, 2019). For example, Singh and

Chaudhury (2020) classified rice grains based on morphology, color, texture, and wavelet features, using image pre-processing techniques followed by a cascade network classifier. Similarly, Nagoda and Ranathunga (2018) employed support vector machines (SVM) and image processing methods to classify rice samples based on physical properties like color and texture, achieving a segmentation accuracy of 96% and a classification accuracy of 88%. Cinar (2019) also identified seven morphological features for classifying two different rice species. Several machine learning models, including Logistic Regression (LR), Multilayer Perceptron (MLP), SVM, Decision Trees (DT), Random Forest (RF), Naïve Bayes (NB), and K-Nearest Neighbor (KNN), were tested for classification accuracy, with success rates ranging from 88.58% to 93.02%. Additionally, Chatnuntawech et al. (2018) proposed a deep CNN algorithm for classifying rice varieties, using spatial-spectral data from two datasets, and achieved a mean classification accuracy of 91.09%. Their study also employed hyperspectral imaging to examine rice seeds in a consistent orientation.

Despite the limited research on paddy classification, there is a clear need to focus specifically on classifying Sri Lankan paddy varieties. The review of existing literature highlights the scarcity of research on Sri Lankan paddy and the lack of application of emerging deep learning methodologies. Therefore, this study aims to evaluate the effectiveness of deep learning algorithms in classifying Sri Lankan paddy varieties.

II. LITERATURE REVIEW

Artificial Neural Networks (ANN) have significant role for rice classification. For instance, Pazoki et al. (2014) used ANN Multi-Layer perceptron (MLP) and neuro-fuzzy networks to classify five rice varieties in Iran along with UTA feature selection algorithm to fine-tune the classifiers. The analysis used 24 color features, 11 morphological properties, and four shape factors to classify rice grains. The screening is proved to have a rate above 99% for both approaches.

There are numerous ML techniques that are available for the classification purposes. Arora et al., (2020) used different image processing algorithms and ML algorithms for rice grain classification using various parameters of

individual rice grains like major axis, minor axis, eccentricity, length, breadth, etc. Relevant features of the rice grains have been extracted using various image processing algorithms. The rice grain images have been classified using different machine learning algorithms, such as LR, DT, NB, KNN, RF and Linear Discriminant Analysis (LDA) classifiers. They proposed future directions for incorporating additional features like chalkiness and moisture content analysis to ensure good quality rice is delivered.

While various ML algorithms have achieved significant classification accuracy, ensemble learning approach is also gaining momentum for classification problems. Ensemble Learning can achieve better performance than a single model alone by combining various models. It can be applied to various ML tasks including classification. Setiawan & None (2024) used ensemble learning methods to classify rice grains based on image features. The study compared various machine learning algorithms, ultimately finding that Bagging meta-estimator improved classification accuracy by combining predictions from multiple base estimators. They utilized Bagging meta-estimator to aggregate decisions from multiple base classifiers, reducing model variance and improving classification consistency. By applying this approach to various grain features, ensemble method achieved consistent classification accuracy across different paddy varieties.

Most studies on image-based paddy classification have primarily focused on color, morphology, and shape features. By using near-infrared hyperspectral imaging technology, both spatial and spectral information, as well as morphological features, can be captured. Jin et al. (2022) combined near-infrared hyperspectral imaging with traditional machine learning methods and deep learning models to classify rice seed varieties. This non-destructive imaging technique captures high-resolution spectra, enabling the detection of even subtle differences in paddy grain features, which leads to accurate classification across various rice varieties. Among conventional machine learning methods, SVM performed well, while in deep learning, LeNet, GoogLeNet, and ResNet models showed effective identification. Deep learning methods significantly outperformed conventional machine learning algorithms, with

most models achieving classification accuracies exceeding 95%.

In another study, Qiu et al. (2018) employed a near-infrared hyperspectral imaging system with two different spectral ranges (380–1030 nm and 874–1734 nm) to classify four rice seed varieties. The study compared the performance of various discriminant models, including KNN, SVM, and CNN. Models utilizing the spectral range of 874–1734 nm outperformed those built with the 380–1030 nm range, with CNN outperforming both KNN and SVM.

Rajalakshmi et al. (2024) achieved 97% accuracy in classifying 13 southern Indian paddy varieties — such as *Yanaikomban*, *Swarna Masoori*, *Sivapu Kowuni*, and *Mapillai Samba*—using a Deep Neural Network (RiceSeedNet) combined with traditional image processing techniques. They also demonstrated RiceSeedNet's potential to achieve 99% accuracy in classifying eight paddy grain varieties from a public dataset. The study utilized two datasets: one containing 13,000 images of southern Indian paddy varieties (1,000 images per variety), and another with 8,000 images from an open-source benchmark dataset (1,000 images per variety). In the research of Paddy seed variety classification using transfer learning based on deep learning, Jaithavil, D. et al. (2022) used three pre-trained models VGG16, InceptionV3, and MobileNetV2 to classify three paddy varieties. Compared with various other two models Inception-v3 showed the highest accuracy and least test loss with 83.33% and 28.41% respectively

Few other recent studies have been successful in classifying paddy varieties. For instance, Ansari, N. et al. (2021), presented a rapid inspection method to classify three paddy varieties using color, texture and morphological features and k-nearest neighbors, support vector machine, and partial least squares-discriminant analysis (PLS-DA) algorithm. Where the classification accuracy using PLS-DA, SVM-C, and KNN model was 83.8%, 93.9%, and 87.2% respectively. In another study, Uddin, M. et al. (2021) proposed a computer vision-based system for non-destructive paddy seed variety identification, crucial for maintaining seed purity in agriculture and industry. To address challenges like illumination variations during image capture, the study introduced a modified histogram-oriented gradient

(T20-HOG) feature. Combined with Haralick and traditional features, these were refined using the Lasso technique and used to train a feed-forward neural network (FNN) for accurate variety prediction demonstrated 99.28% accuracy in identifying paddy grain types.

Anami, B.S. et al. (2020) proposed a deep convolutional neural network (DCNN) framework for automatic recognition and classification of various biotic and abiotic stresses in paddy crops. The pre-trained VGG-16 CNN model was used to classify stressed images during the booting growth stage. The trained models achieved an average accuracy of 92.89% on the held-out dataset, demonstrating the technical feasibility of using the deep learning approach. The proposed work finds applications in developing decision support systems and mobile applications for automating field crop and resource management practices. The approach is applicable to 11 classes of biotic and abiotic stresses from five different paddy crop varieties.

III.METHODOLOGY

The methodology applied for this study is illustrated in Figure 01 below.

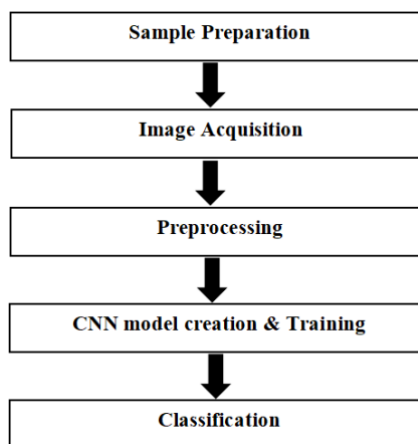


Figure 01: Applied Methodology

G. Sample Preparation

Although many different varieties of paddy are available today, ten common paddy grain samples were chosen for this study using a convenient sampling method. 100grams of paddy grain samples from eight common Sri Lankan Paddy varieties (At 309, At 362, At 373, Bg 300, Bg 352, Bg 359, Bg 374, Bw 367) and two Sri Lankan traditional varieties (Kahawanu, Madathawalu) were selected for the data set preparation. They

were obtained from Rice Research and Development Institute (RRDI), Bathalagoda, Ibbagamuwa (Figure 02).

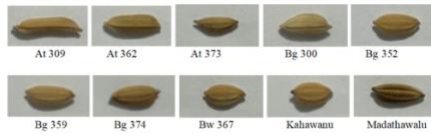


Figure 07: Samples images of ten paddy varieties

H. Image Acquisition

The paddy grains were first cleaned to eliminate impurities, and random samples from each variety were selected for image acquisition. 1000 images of each paddy varieties were captured in the same lightning condition and same fixed frame by an iPhone 14 pro camera. Each image of a paddy seed was acquired with the seed placed centrally and horizontally, with the seed body rotated along the horizontal axis.

I. Preprocessing

First, images of paddy were cropped and scaled to a uniform size of 500 x 250 pixels to standardize all the images and noise removal was done using bilateral and non-local filters where, bilateral filtering was effective for preserving edges and non-local filtering was effective for various noise types (Figure 03).



Figure 08: Preprocessed Images a) Raw image captured from camera b) Image after Cropping c) Image after Noise Removal

Bilateral filtering is an advanced image processing technique used to smooth images while preserving edges, making it ideal for applications where edge preservation is important. Bilateral filtering maintains the integrity of edges while reducing noise and smoothing the image.

Bilateral filtering is defined by:

$$BF[I]_p = \frac{1}{W_p} \sum_{q \in S} G_{\sigma_s}(\|p - q\|) G_{\sigma_r}(I_p - I_q) I_q$$

where:

$BF[I]_p$ is the filtered image at pixel p

I_p and I_q is are the intensities at pixels p and q respectively.

S is the spatial domain of the image.

$G_{\sigma_s}(\|p - q\|)$ is the spatial Gaussian kernel.

$G_{\sigma_r}(\|I_p - I_q\|)$ is the range Gaussian kernel.
 W_p is the normalization factor:

$$W_p = \sum_{q \in S} G_{\sigma_s}(\|p - q\|) G_{\sigma_r}(I_p - I_q)$$

The bilateral filter operates by combining both spatial and range kernels, where the spatial kernel depends on the Euclidean distance between pixels p and q , with σ_s controlling the spatial extent of the filter and the range kernel depends on the intensity difference between pixels p and q , with σ_r controlling the range of intensity values that influence the filtering. The filtered image at a pixel is computed as a weighted sum of neighboring pixels, with the weights determined by both the spatial distance and intensity difference. This method effectively smooths regions with similar intensities while maintaining the sharpness of edges. Bilateral filtering provides better edge preservation than other filtering methods like Gaussian filtering, which often blurs edges, or median filtering, which can lose finer details. By reducing noise while retaining sharp edges, bilateral filtering ensures that important features are maintained, leading to more accurate classification.

J. Model Creation and Training

Since CNN has shown proven accuracy in various image-based classification problems due to their ability to capture spatial hierarchies of features through convolutional layers, this study employed CNN for the classification of paddy (Alzubaidi et al., 2021). Single paddy image was used to train the CNN model for the paddy classification. Dataset was split into three sets: training, validation and testing for accurate evaluation. Several own baseline CNN Models were created and trained on the training set, with performance tracking and hyperparameter optimization guided by the validation set. Iterative refinement was done using the validation set to consistently improve the model's performance.

In the development of a CNN model, the initial attempts utilized basic architectures with dropout layers to prevent overfitting, followed by the addition of batch normalization for improved training stability. The third iteration introduced preprocessed images with noise removal. Subsequent models incorporated further refinements, including dropout, regularization

techniques, L2 regularization, and a learning rate scheduler to enhance model robustness and performance. In the final models, the primary focus was on significantly expanding the dataset to improve the model's effectiveness.

The best-performing CNN model, as determined by our experiments, was composed of convolutional layers, each followed by activation functions, batch normalization, and pooling layers. Initially the input layer of the architecture processed images of size 500x250x3. The first convolutional layer applied 32 filters of size 3x3 with ReLU activation, followed by a 2x2 max pooling layer. This structure was consistently applied across subsequent layers, with the number of filters progressively increasing to 64, 128, 256, and finally 512, enabling the model to extract increasingly complex features. Batch normalization was used after each convolutional layer to enhance training stability, and dropout layers were incorporated to mitigate overfitting. The model concluded with a fully connected layer comprising 512 neurons, regularized with L2, and a dropout rate of 0.5. The final softmax output layer classified the input into one of 10 categories.

K. Evaluation

In the development of a CNN models, each model designed with different architectural complexities. The comparison of these models was conducted to identify the most effective paddy identification task. Performance evaluation was carried out using performance metrics, including accuracy, precision, F1 score, recall and AUC.

Accuracy: The percentage of correctly classified instances in the dataset is measured generally as accuracy. The number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) is used to calculate it.

The accuracy is defined by:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Precision: Defined as the percentage of correctly identified positive instances among all predicted positive instances. It is particularly important when the cost of false positives is high, as it indicates the reliability of the positive predictions. The precision is defined by:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall: quantifies the percentage of real positives that the model accurately detected. When ignoring positive cases (false negatives) is more crucial than mislabeling negatives as positives, it is extremely significant.

The recall is defined by:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

F1 score: The average mean of recall and precision combined.

The F1 score is defined by:

$$\text{F1 Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

AUC: Area Under the Curve evaluates the model's ability to distinguish between positive and negative classes. It reflects the probability that a randomly chosen positive instance is ranked higher than a negative one. Higher AUC values indicate better performance, with 1.0 being perfect and 0.5 indicating no discrimination.

IV. RESULTS & DISCUSSION

A. Model 1

The global minimum of the loss function was reached at the 30th epoch. Despite this, the results were suboptimal; the validation accuracy was 74.2%. A significant oversight in this training phase was the use of raw data, rather than preprocessed data, which likely impacted the model's performance negatively. This experience underscores the importance of data preparation in building effective machine learning models, as preprocessing can significantly influence the accuracy and efficiency of the training process.

B. Model 2

The global minimum of the loss function was again reached at the 30th epoch. However, despite these adjustments, the results did not meet expectations. The validation accuracy was recorded at 62.5%. Similar to our initial attempt, the model was trained using raw data instead of preprocessed data, which adversely affected its performance. This experience further highlighted the critical importance of data preprocessing in training CNN models, as the lack of it can lead to significant discrepancies in performance metrics.

C. Model 3

The global minimum of the loss function was achieved at the 80th epoch. Despite these improvements in data preprocessing, the performance metrics indicated that the training was not optimal. The validation accuracy was relatively low at 47.5%. This suggested issues in the model architecture or parameter settings that were not addressed merely by preprocessing the data. The significant loss indicates that further model evaluation and adjustments are necessary to improve its effectiveness.

D. Model 4

with the global minimum of the loss function being reached remarkably early at the 10th epoch. Despite these extensive changes, the training outcomes were highly unsatisfactory. The model achieved a validation accuracy of only 17%. This performance indicates a significant misalignment in the model's training process or architectural setup. These results underscore the need for a thorough review and recalibration of the model's configuration and training strategy.

E. Model 5

The model achieved a validation accuracy of 83%, indicating a substantial enhancement in its ability to generalize from the training data to unseen validation data. The improvements in image quality, along with careful preprocessing and effective model architecture, contributed to the much-improved performance metrics. This iteration demonstrates the critical importance of high-quality data and appropriate model tuning in developing effective deep learning systems.

F. Model 6

The results led to a notable improvement in the model's performance, achieving a validation accuracy of 76%. This represents a significant enhancement, confirming the effectiveness of the learning rate scheduler in optimizing the training process and the L2 regularization in improving the model's performance. This iteration underscores the utility of adaptive learning rate mechanisms and regularization techniques in boosting the accuracy and efficiency of machine learning models, especially in scenarios involving complex datasets and model architectures.

G. Model 7

The increase in dataset size proved to be highly beneficial, as reflected by a validation accuracy of 75.42%, the highest achieved across all above iterations.

H. Model 8

Finally the same model 7 was trained using our whole dataset and the model achieved a validation accuracy of 93.69%. The substantial improvement in performance with the expanded dataset highlights the critical role of data volume in training machine learning models. A larger dataset provides a more comprehensive representation of the variability and complexity inherent in real-world data, thereby enhancing the model's ability to learn and generalize effectively. This milestone underscores the importance of both quality and quantity in dataset composition when aiming to improve model accuracy and robustness.

The performance of our models was affected by the quality of the dataset, image conditions, and the architectural choices made during model development. Models 1 and 2, which used raw, unprocessed images captured under varying lighting conditions and angles, struggled with noise and irrelevant features due to wide backgrounds and inconsistent image conditions, leading to poor identification and suboptimal results. In Model 3, preprocessing steps such as resizing and cropping were introduced, but the model still underperformed, indicating that the presence of wide backgrounds continued to overshadow the seeds. To address these issues, Models 5 through 8 utilized a standardized image capture process, where all images were centrally aligned, uniformly cropped, and preprocessed using bilateral filtering for edge detection and non-local means filtering for noise reduction. The architectural improvements in later models, including the addition of dropout, L2 regularization, and learning rate scheduling in Models 5 and 6, further helped to prevent overfitting and enhance generalization.

To ensure the trained model is correctly identifies the paddy seed, 30 paddy images were used for the prediction which are not used for train, test, or validate the model. The best performed model identified all images and other models identified few.

The result is summarized and presented below in Table 01 and Table 02.

Table 01: Summary of different CNN models which are trained using proper dataset

Model	Epoch	Learning rate	Validation Accuracy	Train Accuracy	Test Accuracy	F1 Score	Precision	recall	AUC
5	60	0.001	83%	99.5%	78.25%	0.11	0.14	0.11	0.51
6	60	Dynamically changed	76%	100%	55%	0.18	0.18	0.11	0.47
7	60	Dynamically changed	75.42%	100%	82.25%	0.11	0.11	0.12	0.51
8	60	Dynamically changed	93.69%	100%	89.75%	0.94	0.95	0.94	0.99

Table 02: Summary of different CNN models which are trained using improperly captured images

Model	Epoch	Learning rate	Validation Accuracy
1	100	0.0001	74.2%
2	40	0.001	62.5%
3	100	0.0001	47.%
4	60	0.001	17%

Figure 03 illustrates the prediction of paddy seeds by Model 8.

```

1/1 [-----] - 1s 63ms/step
Image: 362 (1).jpeg
Prediction Array: [[1.5889396e-05 9.9998415e-01 1.0628810e-11 3.7361032e-13 2.6187884e-13
1.2614057e-15 6.7037154e-11 6.6892455e-14 7.4628297e-19 3.9886514e-09]]
Predicted class: At 362
Confidence Score: 100.00%
1/1 [-----] - 0s 17ms/step
Image: md (1).jpeg
Prediction Array: [[2.26420778e-18 4.32060393e-13 1.10683946e-16 6.32299211e-15
7.0428942e-10 5.28779545e-15 1.70372715e-14 8.26221579e-18
1.9227280e-13 1.08080800e+00]]
Predicted class: Madathawala
Confidence Score: 100.00%
1/1 [-----] - 0s 18ms/step
Image: 352 (1).jpeg
Prediction Array: [[1.1551457e-10 5.9777302e-13 8.1712354e-15 3.3654701e-11 9.9999976e-01
2.9189147e-07 3.0443587e-08 8.9223295e-10 2.7009877e-08 2.8304097e-09]]
Predicted class: Bg 352
Confidence Score: 100.00%
1/1 [-----] - 0s 18ms/step
Image: 359 (3).jpeg
Prediction Array: [[1.9641088e-10 9.9807236e-09 3.7991930e-09 8.1479861e-08 1.1266263e-06
9.9999950e-01 2.5985314e-07 2.6493576e-09 2.0352255e-10 7.9110485e-08]]
Predicted class: Bg 359
Confidence Score: 100.00%
1/1 [-----] - 0s 18ms/step
Image: 359 (1).jpeg
Prediction Array: [[1.4786457e-10 5.6442135e-09 2.1270661e-09 6.1704355e-08 9.3205853e-07
9.9999881e-01 1.7159056e-07 2.2824875e-09 1.2905016e-10 4.7325081e-08]]
Predicted class: Bg 359
Confidence Score: 100.00%
1/1 [-----] - 0s 17ms/step
Image: kh (2).jpeg
Prediction Array: [[5.2067427e-14 1.2070995e-10 4.4925100e-10 4.3881765e-12 6.0628503e-07
6.8897011e-11 9.2541686e-10 3.2995886e-08 9.9999607e-01 1.4472382e-08]]
Predicted class: Kalamau
Confidence Score: 100.00%
1/1 [-----] - 0s 18ms/step
Image: 373 (2).jpeg
Prediction Array: [[4.91304597e-08 1.33561135e-11 1.00000000e+00 3.49475291e-15
3.97582184e-12 2.41737554e-12 1.96208102e-11 1.885754444e-09
3.10389709e-10 4.73261465e-11]]
Predicted class: At 373
Confidence Score: 100.00%
1/1 [-----] - 0s 18ms/step
Image: 374 (3).jpeg
Prediction Array: [[4.12737086e-13 5.86129145e-09 2.99479485e-09 9.96343985e-08
3.47335970e-07 3.26195959e-08 9.99999166e-01 1.20719501e-09
1.25251525e-08 3.28808525e-07]]

```

Figure 03: Prediction of paddy seeds by model 8

X. CONCLUSION

Deep learning technologies are now commonly used in various sectors of agricultural production and industrial food production. In this paper, we aim to develop CNN models to classify 10 paddy varieties from a dataset of nearly ten thousand images of paddy seeds. We investigated nearly 1000 data samples in each paddy variety for training and testing models. Several CNN models were evaluated and compared in order to obtain a model that had the best performance. The highest classification accuracy obtained was 93.69%. The preliminary work presented in this paper could be further enhanced by focusing on clustering to identify and classify different paddy varieties in a single image.

REFERENCES

Alzubaidi, L. et al. (2021b) ‘Review of Deep Learning: Concepts, CNN Architectures, challenges, applications, Future Directions’, Journal of Big Data, 8(1). doi:10.1186/s40537-21-00444-8.

Anami, B.S., Malvade, N.N. and Palaiah, S. (2020) ‘Deep Learning Approach for recognition and classification of yield affecting paddy crop stresses using field images’, Artificial Intelligence in Agriculture, 4, pp. 12–20. doi:10.1016/j.aiaa.2020.03.001.

Ansari, N. et al. (2021) ‘Inspection of paddy seed varietal purity using machine vision and multivariate analysis’, Journal of Agriculture and Food Research, 3, p. 100109. doi:10.1016/j.jafr.2021.100109.

Arora, B. et al. (2020) ‘Rice grain classification using Image Processing & Machine Learning Techniques’, 2020 International Conference on Inventive Computation Technologies (ICICT) [Preprint]. doi:10.1109/icit48043.2020.9112418.

Cinar, I. and Koklu, M. (2019) ‘Classification of rice varieties using artificial intelligence methods’, International Journal of Intelligent Systems and Applications in Engineering, 7(3), pp. 188–194. doi:10.18201/ijisae.2019355381.

Common rice diseases (no date) Department of Agriculture Sri Lanka. Available at: <https://doa.gov.lk/rrdi/index.php> (Accessed: 6 Dec. 2023).

Golpou, I., Parian, J.A. and Chayjan, R.A. (2014) ‘Identification and classification of bulk paddy, Brown, and white rice cultivars with colour features extraction using image analysis and neural network’, Czech Journal of Food Sciences, 32(3), pp. 280–287. doi:10.17221/238/2013-cjfs.

- Jaithavil, D., Triamlumlerd, S. and Pracha, M. (2022) 'Paddy seed variety classification using transfer learning based on Deep Learning', 2022 International Electrical Engineering Congress (iEECON) [Preprint]. doi:10.1109/ieecon53204.2022.9741677.
- Jin, B. et al. (2022) 'Identification of rice seed varieties based on near-infrared hyperspectral imaging technology combined with Deep Learning', ACS Omega, 7(6), pp. 4735–4749. doi:10.1021/acsomega.1c04102.
- Kiratiratanapruk, K. et al. (2020) 'Development of paddy rice seed classification process using machine learning techniques for automatic grading machine', Journal of Sensors, 2020, pp. 1–14. doi:10.1155/2020/7041310.
- Nagoda, N. and Ranathunga, L. (2018) 'Rice sample segmentation and classification using image processing and support vector machine', 2018 IEEE 13th International Conference on Industrial and Information Systems (ICIIS) [Preprint]. doi:10.1109/iciinfs.2018.8721312.
- Qiu, Z. et al. (2018) 'Variety identification of single rice seed using hyperspectral imaging combined with Convolutional Neural Network', Applied Sciences, 8(2), p. 212. doi:10.3390/app8020212.
- Rajalakshmi, R. et al. (2024) 'RiceSeedNet: Rice Seed Variety Identification Using Deep Neural Network', Journal of Agriculture and Food Research, 16, p. 101062. doi:10.1016/j.jafr.2024.101062.
- Robert Singh, K. and Chaudhury, S. (2020) 'A Cascade Network for the classification of rice grain based on single Rice Kernel', Complex & Intelligent Systems, 6(2), pp. 321–334. doi:10.1007/s40747-020-00132-9.
- Setiawan, R. and Hayatou Oumarou (2024) 'Classification of rice grain varieties using ensemble learning and Image Analysis Techniques', Indonesian Journal of Data and Science, 5(1), pp. 54–63. doi:10.56705/ijodas.v5i1.129.
- Uddin, M. et al. (2021) 'Paddy seed variety identification using T20-hog and Haralick textural features', Complex & Intelligent Systems, 8(1), pp. 657–671. doi:10.1007/s40747-021-00545-0.

Leveraging Big Data and Advanced Analytics for Enhanced Decision-making: Insights and Applications

W.C.K. Jayaweera¹, M.S. Shafana² and M.J. Ahamed Sabani³

^{1,2,3}Department of Information and Communication Technology, South Eastern University of Sri Lanka, Sri Lanka

¹charunikosala9723@seu.ac.lk, ²zainashareef@gmail.com, ³mjasabani@seu.ac.lk

Abstract

In today's era of exploding data volume, Big Data and its analytical tools are increasingly being embraced by organizations across various sectors to extract actionable insights for informed decision-making. This research paper investigates the critical role played by Big Data and analytics in driving strategic decisions across diverse domains. The multifaceted applications of Big Data analysis are examined in this paper, with a focus on customer behaviour analysis, marketing trend analysis, fraud detection and prevention, operational efficiency optimization, and risk management in decision-making. By organizations, deeper insights are gained into customer preferences, purchasing patterns, and consumer perceptions through the harnessing of Big Data, ultimately leading to an increase in customer loyalty. Big Data facilitates the identification of emerging market trends, enabling businesses to swiftly adapt their strategies, capitalize on new opportunities, and remain ahead of the competition. Anomalous patterns and suspicious activities are helped to be detected through advanced analytics techniques employed with Big Data, thereby fortifying organizations against fraud and minimizing financial losses. Additionally, operational processes are optimized through Big Data analytics, ultimately leading to cost savings and improved productivity. Furthermore, proactive risk identification, assessment, and mitigation strategies are enabled by Big Data analysis, empowering organizations to navigate uncertainties effectively and safeguard against potential threats. This paper sheds light on how valuable insights are provided for leaders seeking to leverage data for strategic decision-making and achieving sustainable success, with Big Data analytics transcending industries.

Keywords: Big data, Big data analytics, Decision making

I. INTRODUCTION

In today's increasingly interconnected and digitized world, the proliferation of data has been defined as a characteristic of our era. Big Data, characterized by its volume, velocity, and variety (the 3 Vs) (Elgendy, N. and Elragal, A., 2016), with some categorizing additional characteristics like value and veracity (5Vs) (Hiba, J. *et al.*, 2015) and even further extended lists including virality, volatility, visualization, viscosity, and validity (up to 8 Vs) (Kapil, G. *et al.*, 2016), has revolutionized how organizations operate and make decisions. The ability to harness and analyze vast amounts of data has been opened up by Big Data, from businesses and governments to healthcare and academia, for understanding complex phenomena, predicting trends, and driving innovation.

Big Data Analytics, a multifaceted discipline at the heart of the data revolution, refers to the process of large, complex datasets being collected, organized, and analyzed (Riahi, Y. and Riahi, S., 2018). It utilizes a range of techniques and technologies, including advanced algorithms, machine learning, and statistical methods, to unlock hidden value from these enormous datasets. This allows decision-makers to extract actionable insights from the vast amount of information available, empowering them to make informed choices (Elgendy, N. and Elragal, A., 2016). Across various domains and industries, a pivotal role is played by Big Data and its analysis in the realm of decision-making (Di Berardino, D. and Vona, S., 2023). Leveraging data-driven insights has become a cornerstone of success in the modern era, whether it's business strategies being optimized, customer experiences being enhanced, public services being improved, or scientific research being advanced.

Hidden insights from industrial data are unlocked by big data analysis in manufacturing, granting leaders a competitive edge through enabling

informed decisions in complex environments (Li, C. *et al.*, 2022). This aligns with findings from a separate study which highlights the link established between big data analytics and improved decision-making within businesses (Awan, U. *et al.* 2021). However, by the latter study, it is also suggested that a larger role may be played by business intelligence and data-driven insights than big data analytics capability itself.

Beyond manufacturing, big waves are being made by big data in healthcare. The potential of big data analytics in cardiology for improving the quality of care and reducing costs is pointed to by a study (Nazir, S. *et al.* 2019). Similarly, research on smart buildings explores how machine learning and big data analytics can be utilized to manage data and potentially improve decision-making (Qolomany, B. *et al.* 2019).

The importance of focusing on data quality over quantity for effective decision-making is emphasized by another study (Kościelniak, H. and Puto, A., 2015). A utilitarian decision-making model is proposed by this study, which considers the overall strategy of the enterprise, but acknowledges the need for further development in selecting important information from vast datasets.

Cloud computing and big data (Niu, Y. *et al.* 2021) emphasize the need for careful consideration before adopting cloud-based business intelligence. The entire+9 decision-making process can be potentially compromised by security risks associated with cloud storage and data transfer, if sensitive information is breached.

This literature review aims to comprehensively explore the pivotal role of big data and advanced analytics in enhancing decision-making processes. By examining the key applications, benefits, challenges, and best practices, this review seeks to provide valuable insights for organizations seeking to leverage these technologies for informed decision-making.

In Section II, we meticulously outline the methodology employed for our literature review, encompassing the search strategies, databases consulted, and the criteria for inclusion and exclusion. We also describe the data analysis techniques utilized to synthesize the reviewed literature. Section III elucidates the key findings

in alignment with our research objectives, providing comprehensive insights into the impact of big data and advanced analytics on decision-making processes. In Section IV, we delve into the broader implications of these findings for both theory and practice. Finally, Section V presents our concluding thoughts and suggests directions for future research in this evolving field.

II.METHODOLOGY

A multi-database approach was employed to identify relevant sources for this literature review. Initially, a broad search of Google Scholar was conducted to capture the available publications on the topic. Titles and research data were then retrieved from Google Scholar based on a developed search strategy. Subsequently, additional searches were conducted in databases including IEEE Xplore, JStor, ScienceDirect, and others.

A two-stage selection process was implemented to ensure the included sources were relevant and credible. During the initial screening, articles were selected for their apparent relevance to the research question, based on keyword matching. This initial selection was further refined through a full-text screening process utilizing pre-determined inclusion/exclusion criteria. Here, the focus shifted to the depth and detailed relevance of the content to the research question.

For this review, a variety of scholarly sources were considered, including research articles, conference publications, established literature reviews, and other credible review articles.

The selection of sources prioritized both relevance and credibility. Included articles directly addressed the research topic and originated from trustworthy and reliable sources. Google Scholar's comprehensiveness was initially valuable due to its broad search capabilities. However, the focus ultimately shifted to specialized databases such as IEEE Xplore for its curated content in engineering and technology. Additionally, resources from JStor, ScienceDirect, and other established academic databases were included to ensure a well-rounded selection.

The initial analysis of the selected resources began with a general overview being conducted for each source. Introductions, conclusions, and methodology sections were skimmed to grasp the

main points and research methods employed by the authors. This initial analysis will be followed by a more in-depth analysis involving a critical appraisal of each study. This appraisal will focus on the research methodology, potential biases, and the generalizability of the findings. Once each study has been critically evaluated, a process of synthesis will be undertaken. This synthesis will involve connections being drawn and patterns being identified across the studies, highlighting both agreements and contradictions. The goal of the synthesis is to provide a comprehensive understanding of the current state of knowledge on the research question.

III. BIG DATA ANALYTICS IN DECISION-MAKING

The power of big data analytics has led to a significant impact on the field of decision-making. The effectiveness of big data in this domain has been documented in numerous research articles. This review article focuses on identifying the key roles played by big data and big data analytics in decision-making. We will explore five key ways in which big data enhances decision-making processes.

I. Customer Behavior Analysis

The field of consumer behavior analysis is being revolutionized by the power of Big Data Analytics (BDA). Compelling evidence for BDA's ability to significantly enhance understanding of consumer behavior has been provided by numerous research frameworks. This has opened exciting opportunities for businesses to tailor their strategies and optimize customer experiences.

BDA is argued by Holmlund, M. *et al.* to be a useful tool for capturing and analyzing customer experience (CX) data (Holmlund, M. *et al.* 2020). However, improvement in CX is not solely achieved through having more data. Businesses need to focus on collecting the right data and utilizing it to generate actionable insights. The paper proposes a new framework for CXM that considers the various types of CX data and analytics that can be employed. The authors call for further research on how BDA can be used to improve CX in non-commercial settings, as well as how to develop better CX metrics and analytics tools.

BDA companies are becoming powerful allies for Consumer Goods and Retail Companies (CGRCs) in the realm of innovation (Mariani, M.M. *et al.* 2020). Faster innovation cycles for CGRCs can be fueled by BDA through bridging knowledge gaps. The research acknowledges limitations and emphasizes the need for further exploration across industries to solidify these concepts.

A promising Big Data application framework for analyzing consumer behaviors utilizes topological data structures, co-occurrence methodology, and Markov chain theory (Zin, T. T. *et al.* 2020). This framework operates in three layers: data organization, analysis and modelling, and prediction and inference. Studies have shown that this framework can effectively identify customer behaviors. For instance, it can be used to predict the most popular product combinations in a store, providing valuable insights into what products customers are most likely to buy together.

The proposed system addresses building decision trees for massive customer datasets using the C4.5 algorithm (Khade, A.A., 2016). Distributed processing for the ever-growing data volumes in today's cloud computing and big data world is catered to by leveraging the MapReduce framework. Traditional decision tree algorithms are simply not handled effectively by such large datasets.

Speed, reusability, and the familiar comfort of HTML elements alongside Scalable Vector Graphics (SVG) are offered by D3.js, which comes in for data visualization. The authors envision future improvements to be made to boost the system's efficiency and scalability, including the incorporation of realtime databases like Apache HBase or MongoDB, and the utilization of advanced distributed algorithms like ForestTree from Apache Mahout.

A mathematical and machine learning-based predictive model is shown to exist, with the capability to forecast consumer behavior using social media data from various platforms such as Facebook and YouTube (Chaudhary, K., *et al.* 2021). This model proves valuable for businesses by allowing them to understand how consumers might react to a product based on social media information. The findings demonstrate significant

variations in consumer behavior across different social media platforms, with a maximum deviation observed at 99.51%. The model's accuracy was also measured, achieving a maximum of 0.98. It utilizes machine learning techniques and big data analytics to analyze social media data such as likes, followers, and downloads to predict consumer behavior on different platforms.

Customer segmentation based on the Time-Frequency-Monetary (TFM) value model and the establishment of loyalty tiers were previously employed (Wassouf, W. N., *et al.* 2020).

Classification algorithms were then applied, using loyalty levels as the classification categories and selected customer attributes as features. The results were compared to identify the most accurate classification model. Subsequently, rules for loyalty prediction were derived from this model. These rules revealed the correlations between behavioral characteristics and loyalty levels, providing insights into the drivers of loyalty within each customer segment. Targeted marketing efforts with appropriate offers and services for each segment were enabled by this approach. An additional benefit of using classification algorithms was the development of a precise predictive model for classifying new users based on their loyalty potential.

J. Trend Analysis

Valuable trends can be uncovered by analyzing the vast amount of data on Twitter (big data) (Rodrigues, A. P., *et al.* 2021). Big data analytics techniques like LDA (topic modelling) and K-means clustering go beyond simply counting hashtags. Hidden themes, user groups with specific interests are revealed by these techniques, providing a more nuanced understanding of what's trending. This empowers businesses to target customers effectively, politicians to understand public sentiment, and movie studios to gauge audience reception – all with improved accuracy compared to traditional methods.

Various preprocessing techniques are applied to the data before analysis, such as converting emoticons to text, removing hyperlinks, punctuation, and white spaces, removing stop words, stemming, and lemmatization. Hashtag

counting was initially used in this study, but since it doesn't consider the actual tweet content for trend prediction, noun counting was also employed. Latent Dirichlet Allocation (LDA) was then used for clustering, followed by cosine similarity, K-means clustering, and Jaccard similarity for trend analysis. The analysis included both real-time and static data. Real-time streaming SPARK was utilized for real-time data analysis. In short, big data analytics unlocks a deeper level of trend analysis on Twitter, yielding actionable insights for a variety of stakeholders.

K. Fraud Detection and Prevention

The financial strain on healthcare systems in the US due to a growing elderly population and advancements in medical technology is highlighted in one of the articles (Herland, M. *et al.*, 2018). The article focuses on Medicare fraud, a significant issue that wastes billions of dollars. Traditionally, fraud detection relies on manual auditing, which is inefficient when dealing with vast amounts of data. The increasing availability of big data, like electronic health records, opens doors for using machine learning to improve fraud detection in Medicare. The Centers for Medicare and Medicaid Services (CMS) plays a role by releasing big datasets to aid in identifying fraud and abuse. A method for using big data and machine learning to identify fraudulent activity in Medicare claims is proposed by the authors. They compare the effectiveness of using individual datasets (Part B (physician and other supplier utilization and payment data), Part D (prescriber utilization and payment data), and DMEPOS (referring durable medical equipment, prosthetics, orthotics, and supplies utilization and payment data)) and a combined dataset. Their findings show that the combined dataset with Logistic Regression delivers the best overall performance in detecting Medicare fraud. The study paves the way for further research using data sampling techniques to improve fraud detection accuracy.

Several ways auditors leverage big data analytics to detect and prevent fraud are identified in one of the papers (Rosnidah, A. P., *et al.* 2021). A framework that considers technological, organizational, and environmental factors (TOE) is presented. Technological factors include the specific data analytics tools used, while organizational factors encompass the audit firm's size and management's attitude towards this approach. Finally, environmental factors include

the industry, competition, and government regulations that the firm operates within.

- Test data: Created by the auditor to test the client’s computer software controls.

Data mining, a broad concept used to find patterns and relationships within data, is highlighted as playing a crucial role. Text data mining is described as particularly valuable for fraud detection because much information is stored in text format. The paper goes on to explain that this process involves four key tasks,

- Classification: Sorting data into predefined categories.
- Clustering: Grouping similar data patterns together.
- Regression: Modeling data with minimal error.
- Association rule learning: Identifying how often specific patterns appear.

Fraud trends can be uncovered and the location’s role in suspicious activity can be understood through geospatial analysis. Large datasets like insurance claims or burglary reports can be analyzed to proactively identify fraud by finding patterns and clusters that might indicate fraud rings. While data integration and using automated tools remain challenges, exploring new models and systems to aid fraud detection and support decision-making is crucial. For successful implementation, obtaining data from various sources in a consistent format for unified analysis is essential.

Computer-Assisted Audit Techniques (CAATTs) are identified as valuable tools for audits of all sizes, not just large firms. Even with basic computer skills, CAATTs can be leveraged to improve productivity, accuracy, and client relationships. There are two main types of CAATTs:

- Audit software: Analyzes client data for control weaknesses and record integrity.

The paper also identifies key barriers to integrating big data analytics into audit practice, such as data overload, data availability and relevance/integrity, pattern recognition ability, ambiguity, and a lack of training and expertise among auditors. Solutions to improve data analytics procedures for preventing and detecting fraud are proposed by the authors, including operational analysis, strategic analysis, and deep neural networks.

A framework and tools for analyzing retail fraud detection are highlighted in one of the other review papers (Jha, B.K. *et al.*, 2020). This information is presented in Table 01.

L. Operational Efficiency Optimization

Systems to track employee performance and company success factors are being built by companies. These systems collect and analyze data to aid leaders in making informed decisions (Schl fke, M. *et al.*, 2012). Performance management is being extended beyond financials, with new metrics and ongoing advancements being embraced. Data analysis is being used by businesses to improve performance management by identifying cause-and-effect relationships and utilizing various data sources to inform better decisions. This approach necessitates a strong IT infrastructure and data analysis skills, but it can be very effective. This paper argues that performance analytics can be significantly improved by performance management systems (PMS) through the use of data to validate cause-and-effect relationships.

Table 01: Font format for this publication

Author (s)	Method	Application
Gadal, S.M.A.M. and Mokhtar, R.A., 2017	k-means clustering , Sequential Minimal Optimization (SMO)	Retail fraud detection
Zuech, R., <i>et al.</i> , 2015	Hadoop framework	Intrusion detection
Fan, Q., <i>et al.</i> , 2009	Polar Histogram Feature (PHF), Bag-of-Features (BOF)	Intrusion detection
Hoang Trinh, <i>et al.</i> , 2011	Finite State Machine (FSM)	Retail fraud detection
Coppolino, L., <i>et al.</i> , 2015	SEPA Direct Debit system	Online payment
Cantabella, M., <i>et al.</i> , 2017	Data gathering, investigation, imagination	Learning pattern analysis

Cui, H., <i>et al.</i> , 2016	Graph Mining with Frequent Pattern (GM-FP)	Healthcare fraud detection
Xing, E.P., <i>et al.</i> , 2015	Petuum framework	Large-scale Machine Learning
Kitts, B., <i>et al.</i> , 2013	Mix adjustment algorithm	Click fraud detection
Caldeira, E., <i>et al.</i> , 2012	Neural Networks, Random Fores	Transaction fraud detection
Leite, R.A., <i>et al.</i> , 2017	Financial Fraud Detection (FFD)	Banking fraud detection
Balasupramanian, N., <i>et al.</i> , 2017	Big Data analytics	Online Fraud Detection

The use of BDA to improve the sustainability of the mining supply chain in South Africa is examined by one study (Bag, S., *et al.* 2020). The mining industry, while crucial to the South African economy, has social and environmental impacts. BDA can be utilized to optimize business processes such as procurement and logistics, resulting in cost savings, waste reduction, and improved sustainability. The study explores the positive correlation between BDA expertise and employee development. A link between employee development, an organization's human capital, and positive supply chain sustainability outcomes is suggested by their findings. It is argued that product innovation not only fosters employee development but also leads to improved employee performance and overall innovation levels. Managers play a key role in optimizing employee performance through the creation of a supportive learning environment. Employee performance is seen to improve with a greater managerial emphasis on innovation, particularly through the path of green product design, which ultimately leads to a sustainable supply chain. The researchers acknowledge that success in supply chain management, a sequence of activities, can be achieved through both human-driven and data-driven approaches. The development of these skills and the closing of any skill gaps among employees are identified as critical roles played by training. They emphasize that, in today's world, every activity is scrutinized through a technological lens, particularly Big Data Analytics.

The Logistics and Transportation industries are identified as prime candidates for utilizing Big Data (Borgi, T. *et al.* , 2017). The constant movement of goods and people create massive datasets containing valuable information such as location, weight, and destination. These Big Data sets can be analyzed by logistics companies to improve service quality and efficiency. This study demonstrates the potential for big data technologies to be used in optimizing efficiency

within logistics and transportation. By leveraging big data, several milestones can be achieved, including last-mile delivery, route optimization, crowdsourcing and social transportation, smart logistics, and anticipatory logistics.

M. Risk Management

Big data empowers organizations to comprehensively assess and manage risks by revolutionizing the entire risk management process (El Khatib, *et al.*, 2023 & Doggalli, G., *et al.* 2024). Several studies have been conducted that demonstrate the use of BDA for mitigating risks in various sectors, including banking (Dicuonzo, *et al.*, 2019), transportation, and healthcare (Choi, *et al.*, 2017).

Big data analytics are shown to be beneficial for managing various risks, including financial, employee turnover, customer churn, and threats from partners, as evidenced by Apple's approach. Data from Siri and other sources are analyzed by Apple to identify, assess, and mitigate these risks before they occur. Additionally, this data is used to develop recovery plans and improve customer relationships. Due to its reliance on third-party vendors and various business operations, Amazon encounters a multitude of risks. However, these risks are effectively identified, assessed, and mitigated through the company's use of big data analytics. This big data is collected and analyzed through their cloud computing technology and AWS big data software, allowing data-driven decisions to minimize risks like fraud, employee churn, and operational issues. Similar to Apple and Amazon, Google utilizes big data analytics to manage internal and external risks. This big data, collected from various software sources, is analyzed to facilitate early identification and mitigation of risks. This allows Google to prevent fraud, manage risks from third-party companies, and reduce operational risks that could hinder its competitiveness (El Khatib, *et al.*, 2023).

Operational Risk Management (ORM) is defined as the process of identifying and mitigating risks in business operations. A study explores existing literature on ORM frameworks and their applications in various sectors such as transportation, emergency management, and healthcare. The findings indicate that ORM is a growing area of interest, with a focus on power and energy, healthcare, supply chain operations, and information systems.

Big data introduces new challenges to ORM. The value of information assets can be difficult to determine, and storing big data can be expensive. Additionally, cultural and political risks are associated with collecting big data, such as privacy concerns. To address these challenges, companies can estimate the value of information assets, utilize tiered storage for data, and establish risk tolerance levels. Furthermore, new frameworks, such as the Bayesian Markov chain Monte Carlo (BMCMC) framework, have been developed to incorporate big data into ORM. Data mining techniques can also be employed to analyze big data and identify operational risks. For instance, data mining techniques have been used to develop financial early warning systems, forecast customer loyalty, and assess the risk of management fraud (Choi, T.-M *et al.*, 2017).

IV.RESULT AND DISCUSSION

The findings of this literature review underscore the significant impact of Big Data Analytics (BDA) on decision-making across various domains. The analysis of existing research reveals that BDA enhances decision-making processes by providing organizations with deeper insights into customer behavior, emerging trends, fraud detection, operational efficiency, and risk management.

In customer behavior analysis, BDA allows organizations to understand and predict customer actions by analyzing vast datasets, enabling tailored strategies that improve experiences, loyalty, and satisfaction. In trend analysis, advanced techniques like topic modeling and clustering help identify subtle trends that traditional methods might miss, giving organizations a competitive edge in rapidly changing markets. BDA also significantly improves fraud detection and prevention,

particularly in healthcare and finance, by using machine learning and data mining to identify anomalous patterns, reducing losses and enhancing security. Additionally, BDA optimizes operational efficiency by analyzing supply chain data to identify inefficiencies and streamline processes, leading to cost savings and productivity gains. Finally, BDA empowers organizations in risk management by proactively identifying and mitigating potential threats and vulnerabilities, a capability especially crucial in high-stakes sectors like finance and manufacturing.

BDA provides a nuanced understanding of customer behavior, allowing businesses to develop targeted strategies and improve customer satisfaction. Integrate multimodal data sources (e.g., text, images, and sensor data) to enrich customer insights and enhance predictive accuracy. Develop real-time analytics capabilities to promptly respond to changes in customer behavior and preferences. Advanced analytical frameworks, such as topological data structures and Markov chain theory, effectively predict customer behavior and identify popular product combinations. Future direction may focus on improving scalability and efficiency of these frameworks to handle increasingly large datasets. Explore the integration of emerging technologies like AI to refine predictive models and enhance decision-making processes. Techniques like Latent Dirichlet Allocation (LDA) and K-means clustering provide deeper insights into trends and user interests, surpassing traditional hashtag counts. Future studies may lead to Advance real-time trend analysis technologies by enhancing streaming data platforms and incorporating advanced preprocessing techniques. Address ethical considerations and biases in trend analysis to ensure fairness and accuracy.

Machine learning and data mining techniques improve fraud detection, particularly in sectors like healthcare and finance, by identifying anomalous patterns and reducing losses. In future it is necessary to develop robust data privacy and security measures to protect sensitive fraud-related data. Explore cross-industry applications of fraud detection frameworks and tools to enhance effectiveness in various contexts. BDA optimizes operational efficiency by analyzing supply chain data and identifying inefficiencies, leading to cost savings and productivity gains. Future researches may think of Integrating BDA

with emerging technologies like IoT to enhance supply chain management and operational efficiency. Focus on human-centric design to ensure that analytics tools support and enhance human decision-making. BDA revolutionizes risk management by providing advanced tools for identifying and mitigating risks, as demonstrated by companies like Apple, Amazon, and Google. There is a room for addressing challenges in operational risk management (ORM) by developing new frameworks and utilizing data mining techniques to identify and manage risks. Enhance scalability and efficiency in ORM processes to handle large volumes of risk-related data.

V. CONCLUSION

This literature review underscores the transformative impact of Big Data Analytics (BDA) on decision-making across diverse sectors. By offering profound insights into customer behavior, emerging market trends, fraud detection, operational efficiency, and risk management, BDA equips organizations with the tools to make more informed and strategic decisions. The integration of advanced techniques, such as machine learning, data mining, and multimodal data analysis, enhances the value of Big Data, allowing businesses to navigate and capitalize on the complexities of today's data-driven landscape.

Despite these advantages, several challenges persist, including issues related to data quality, integration, privacy, and security. To harness the full potential of BDA, organizations must invest in robust data governance frameworks and cutting-edge analytical tools. Future research should focus on addressing these challenges, exploring the application of BDA in emerging fields, and developing scalable solutions that integrate new technologies like AI and IoT.

Big Data Analytics represents not merely a tool but a strategic asset capable of driving innovation, enhancing competitiveness, and ensuring long-term success. Organizations that adeptly leverage BDA will be well-positioned to thrive in the digital era, transforming data into actionable insights and fostering sustainable growth

REFERENCES

Awan, U., Shamim, S., Khan, Z., Zia, N. U., Shariq, S. M., & Khan, M. N., 2021. Big Data Analytics capability and decision-making: The role of data-driven insight on

circular economy performance, *Technological Forecasting and Social Change*, 168, p. 120766. doi:10.1016/j.techfore.2021.120766.

Bag, S., Wood, L.C., Xu, L., Dhamija, P. and Kayikci, Y., 2020. Big data analytics as an operational excellence approach to enhance sustainable supply chain performance. *Resources, conservation and recycling*, 153, p.104559.

Balasupramanian, N., Ephrem, B.G. and Al-Barwani, I.S., 2017, July. User pattern based online fraud detection and prevention using big data analytics and self organizing maps. In *2017 international conference on intelligent computing, instrumentation and control technologies (ICICT)* (pp. 691-694). IEEE.

Borgi, T., Zoghlami, N., Abed, M. and Naceur, M.S., 2017, July. Big data for operational efficiency of transport and logistics: a review. In *2017 6th IEEE International conference on Advanced Logistics and Transport (ICALT)* (pp. 113-120). IEEE.

Caldeira, E., Brandao, G., Campos, H. and Pereira, A., 2012, October. Characterizing and evaluating fraud in electronic transactions. In *2012 Eighth Latin American Web Congress* (pp. 115-122). IEEE.

Cantabella, M., de la Fuente, E.D., Martínez-España, R., Ayuso, B. and Muñoz, A., 2017, August. Searching for behavior patterns of students in different training modalities through Learning Management Systems. In *2017 International Conference on Intelligent Environments (IE)* (pp. 44-51). IEEE.

Chaudhary, K., Alam, M., Al-Rakhami, M. S., & Gumaiei, A., 2021. Machine learning-based mathematical modelling for prediction of social media consumer behavior using Big Data Analytics, *Journal of Big Data*, 8(1). doi:10.1186/s40537-021-00466-2.

Choi, T.-M., Chan, H.K. and Yue, X., 2017. Recent development in Big Data Analytics for Business Operations and Risk Management, *IEEE Transactions on Cybernetics*, 47(1), pp. 81–92. doi:10.1109/tcyb.2015.2507599.

Coppolino, L., D'Antonio, S., Romano, L., Papale, G., Sgaglione, L. and Campanile, F., 2015, November. Direct debit transactions: a comprehensive analysis of emerging attack patterns. In *2015 10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC)* (pp. 713-717). IEEE.

- Cui, H., Li, Q., Li, H. and Yan, Z., 2016, August. Healthcare fraud detection based on trustworthiness of doctors. In 2016 IEEE Trustcom/BigDataSE/ISPA (pp. 74-81). IEEE.
- Di Berardino, D. and Vona, S., 2023. Discovering the relationship between big data, big data analytics, and decision making: A structured literature review, *European Scientific Journal ESJ*, 19(19), p. 1. <https://doi.org/10.19044/esj.2023.v19n19p1>.
- Dicuonzo, G., Galeone, G., Zappimulso, E. and Dell'Atti, V., 2019. Risk management 4.0: The role of big data analytics in the bank sector. *International Journal of Economics and Financial Issues*, 9(6), pp.40-47.
- Doggalli, G., Kulkarni, S., Meti, S. C., Pattar, S., Aravind, S., Sankati, J., Krishnaveni, S., & Singh, S. S., 2024. Exploring Big Data Innovations in Food and Agriculture Research: An in-depth analysis, *International Journal of Research in Agronomy*, 7(3S), pp. 330–336. doi:10.33545/2618060x.2024.v7.i3se.471.
- Elgendy, N. and Elragal, A., 2016. Big data analytics in support of the decision making process. *Procedia Computer Science*, 100, pp.1071-1084.
- El Khatib, M., Al Shehhi, H. and Al Nuaimi, M., 2023. How Big Data and Big Data Analytics Mediate Organizational Risk Management. *Journal of Financial Risk Management*, 12(1), pp.1-14.
- Fan, Q., Yanagawa, A., Bobbitt, R., Zhai, Y., Kjeldsen, R., Pankanti, S., & Hampapur, A., 2009. Fast detection of retail fraud using polar touch buttons, 2009 IEEE International Conference on Multimedia and Expo [Preprint]. doi:10.1109/icme.2009.5202732.
- Gadal, S.M. and Mokhtar, R.A., 2017. Anomaly detection approach using hybrid algorithm of data mining technique, 2017 International Conference on Communication, Control, Computing and Electronics Engineering (ICCCCEE) [Preprint]. doi:10.1109/iccccee.2017.7867661.
- Herland, M., Khoshgoftaar, T.M. and Bauder, R.A., 2018. Big data fraud detection using multiple Medicare Data Sources, *Journal of Big Data*, 5(1). doi:10.1186/s40537-018-0138-3.
- Holmlund, M., Van Vaerenbergh, Y., Ciuchita, R., Ravald, A., Saran- topoulos, P., Ordenes, F. V., & Zaki, M., 2020. Customer experience management in the age of big data analytics: A strategic framework, *Journal of Business Research*, 116, pp. 356–365. doi:10.1016/j.jbusres.2020.01.022.
- Hiba, J., Hadi, H., Shnain, H.H., Hadishaheed, A., Haji, S., & Azizahbt, A., 2015. BIG DATA AND FIVE V's CHARACTERISTICS.
- Jha, B.K., Sivasankari, G.G. and Venugopal, K.R., 2020. Fraud detection and prevention by using Big Data Analytics, 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC) [Preprint]. doi:10.1109/iccmc48092.2020.iccmc-00050.
- Kapil, G., Agrawal, A. and Khan, R.A., 2016, October. A study of big data characteristics. In 2016 international conference on communication and electronics systems (ICCES) (pp. 1-4). IEEE.
- Khade, A.A., 2016. Performing customer behavior analysis using big data analytics, *Procedia Computer Science*, 79, pp. 986–992. doi:10.1016/j.procs.2016.03.125.
- Kitts, B., Zhang, J.Y., Wu, G. and Mahato, R., Click Fraud Botnet Detection by Calculating Mix Adjusted Traffic Value. *IEEE ISI 2013*, pp.4-7.
- Kościelniak, H. and Puto, A., 2015. Big data in decision making processes of enterprises, *Procedia Computer Science*, 65, pp. 1052–1058. doi:10.1016/j.procs.2015.09.053.
- Leite, R.A., Gschwandtner, T., Miksch, S., Kriglstein, S., Pohl, M., Gstrein, E. and Kuntner, J., 2017. Eva: Visual analytics to identify fraudulent events. *IEEE transactions on visualization and computer graphics*, 24(1), pp.330-339.
- Li, C., Chen, Y. and Shang, Y., 2022. A review of industrial big data for decision making in Intelligent Manufacturing, *Engineering Science and Technology, an International Journal*, 29, p. 101021. doi:10.1016/j.jestch.2021.06.001.
- Mariani, M.M. and Fosso Wamba, S., 2020. Exploring how consumer goods companies innovate in the digital age: The role of Big Data Analytics Companies, *Journal of Business Research*, 121, pp. 338–352. doi:10.1016/j.jbusres.2020.09.012.
- Nazir, S., Nawaz, M., Adnan, A., Shahzad, S., & Asadi, S., 2019. Big data features, applications, and analytics in Cardiology—a systematic literature review, *IEEE Access*, 7, pp. 143742–143771. doi:10.1109/access.2019.2941898.

- Niu, Y., Ying, L., Yang, J., Bao, M., & Sivaparthipan, C. B., 2021. Organizational Business Intelligence and decision making using Big Data Analytics, *Information Processing & Management*, 58(6), p. 102725. doi:10.1016/j.ipm.2021.102725.
- Qolomany, B., Al-Fuqaha, A., Gupta, A., Benhaddou, D., Alwajidi, S., Qadir, J., & Fong, A. C., 2019. Leveraging machine learning and big data for Smart Buildings: A comprehensive survey, *IEEE Access*, 7, pp. 90316–90356. doi:10.1109/access.2019.2926642.
- Riahi, Y. and Riahi, S., 2018. Big data and big data analytics: Concepts, types and technologies. *International Journal of Research and Engineering*, 5(9), pp.524-528.
- Rodrigues, A. P., Fernandes, R., Bhandary, A., Shenoy, A. C., Shetty, A., & Anisha, M., 2021. Real-time twitter trend analysis using Big Data Analytics and machine learning techniques, *Wireless Communications and Mobile Computing*, 2021, pp. 1–13. doi:10.1155/2021/3920325.
- Rosnidah, I., Johari, R. J., Hairudin, N. a. M., Hussin, S. a. H. S., & Musyaffi, A. M., 2022. Detecting and preventing fraud with Big Data Analytics: Auditing Perspective, *Journal of Governance and Regulation*, 11(4), pp. 8–15. doi:10.22495/jgrv11i4art1.
- Schläfke, M., Silvi, R. and Möller, K., 2012. A framework for business analytics in performance management. *International Journal of Productivity and Performance Management*, 62(1), pp.110-122.
- Trinh, H., Fan, Q., Jiyan, P., Gabbur, P., Miyazawa, S. and Pankanti, S., 2011, May. Detecting human activities in retail surveillance using hierarchical finite state machine. In 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 1337-1340). IEEE.
- Wassouf, W. N., Alkhatib, R., Salloum, K., & Balloul, S., 2020. Predictive analytics using big data for increased customer loyalty: Syriatel Telecom Company Case Study, *Journal of Big Data*, 7(1). doi:10.1186/s40537-020-00290-0.
- Xing, E.P., Ho, Q., Dai, W., Kim, J.K., Wei, J., Lee, S., Zheng, X., Xie, P., Kumar, A. and Yu, Y., 2015, August. Petuum: A new platform for distributed machine learning on big data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1335-1344).
- Zin, T. T., Tin, P., Toriu, T., & Hama, H., 2013. A Big Data Application Framework for consumer behavior analysis, 2013 IEEE 2nd Global Conference on Consumer Electronics (GCCE) [Preprint]. doi:10.1109/gcce.2013.6664813.
- Zuech, R., Khoshgoftaar, T.M. and Wald, R., 2015. Intrusion detection and Big Heterogeneous Data: A survey, *Journal of Big Data*, 2(1). doi:10.1186/s40537-015-0013-4.

Machine Learning-based Mobile Application for Weed Detection in Paddy Fields

S.E. Bhashana Ravisankha¹, K.K. Upeksha Hansani², W.A.K. Upeksha Randika³ and N. Kuruwitaarachchi⁴

^{1,2,3,4}Department of Information and Communication Technology, University of Sri Jayewardenepura, Sri Lanka

¹bhashana.ravisankha@gmail.com, ²upekshamaduhsani@gmail.com, ³forupeksha@gmail.com, ⁴kuruwita@sjp.ac.lk

Abstract

In the context of Sri Lanka, where agriculture, particularly paddy cultivation, plays a crucial role, farmers face significant challenges due to weed infestation. Unlike some other countries that have embraced machine learning technologies to address these issues, Sri Lanka has yet to adopt such advanced solutions. To tackle the pervasive weed problem, a research initiative was undertaken to develop a mobile application capable of identifying weed types. The methodology involved utilizing Convolutional Neural Network (CNN) pre-trained models, namely ResNet-50, Inception-v3, and VGG16, along with the Google Colab platform for training the dataset. Among the three models, VGG16 demonstrated the highest accuracy, making it the chosen model to further the research. The primary goal was to achieve a superior level of accuracy in detecting weed species in rice fields. The research team focused on delivering a mobile application with a high level of accuracy to identify and classify weeds in paddy fields. The integration of advanced technologies, such as IoT and machine learning, aimed to provide Sri Lankan farmers with an efficient and effective tool to combat weed-related challenges in their agricultural practices.

Keywords: Weed detection, CNN, VGG16, ResNet-50, Inception-v3, Weed control methods

I. INTRODUCTION

Technology is being developed rapidly, constantly offering new solutions across various industries. Digital devices such as computers and smartphones have become essential tools in addressing everyday challenges. In the modern world, researchers and tech enthusiasts are continually creating systems and innovations by leveraging the latest technological advancements. As an agricultural country, in Sri Lanka, most of the farmers are struggling with the weed problem.

The research is about the development of a machine learning based mobile application to identify the weed types in rice fields and provide weed controlling methods for the identified weed types. VGG16, InceptionV3, and ResNet 50 are the 03 Convolutional Neural Network (CNN) models that we used to train the image dataset. The mobile application was developed using the VGG16 model which gives the best accuracy in object detection. The research is based on the weed problem in Sri Lanka paddy cultivation and proposes a model with a higher level of accuracy for weed detection. In addition, it is planned to provide weed control methods for the identified weed species, and categorize weed control methods under organic (cultural, biological, physical, manual) and inorganic weed control methods with a short description of each weed species.

The target users are Sri Lankan farmers in both rural and urban areas, students, and the young generation interested in farming. In this research project, object detection accuracy is the key point and has the highest priority. Here, we used weed leaves in the image data set, and the accuracy level can be increased by using a large number of images.

In this work, the data set was collected by visiting the fields and capturing images of the found weed plants. The collected data set was trained on Transfer Learning technologies and three Convolutional Neural Network models (VGG16, Inception V3, ResNet 50) and found the best model as VGG16 according to the accuracy level and data loss.



Figure 01: Captured image data set

The technologies that were used for the project were Transfer Learning technologies, three

Convolutional Neural Network models (VGG16, Inception V3, ResNet 50), Keras, etc.

A. Convolutional Neural Network (CNN)

CNN is a part of the Artificial Neural Network (ANN) in deep learning. CNN can recognize the patterns of objects. CNN is designed explicitly for processing pixel data and has a large number of neurons that can self-optimize via learning. CNN consists of 5 layers.

B. Transfer Learning

Transfer learning is the use of a pre-trained model (a model that was trained previously on one problem) in some way on another second problem. In deep learning, transfer learning is a technique where one previously trained neural network model is used to solve another similar problem. Keras is a powerful open-source Python library for developing deep learning models that run on top of the machine learning platform [TensorFlow](#). So, many pre-trained models are available here. VGG16, Inception v3, and ResNet 50 are selected from them.

II. RELATED WORK

Research article shows that weeds reduce the yield and quality of the farm harvest (Kamath, Balachandra and Prabhu, 2020). But in most cases, weed management is not followed. The research contains a technique that can be used for automatic weed detection and identification. Their proposed system was a computer vision-based automatic weed detection. In their application, weeds can be detected and identified and classified from digital images. According to their review, computer vision can be defined as a process of analyzing images and videos into meaningful interfaces. The proposed system can be used in the agriculture sector to identify plant classification and crop disease identification. According to the paper, India lost INR 1050 million because of the harvest losses. They created the dataset according to types of weeds. They collected the dataset using two digital cameras facing down towards the ground. Images were acquired from a Raspberry PI (RP OV56647) camera and stored in RGB color space in JPEG format. They used MATHLAB (R2018) to process images. In their research, they used a sample of 300 datasets, dividing them into training and test sets. They predicted outcomes on the test data and measured diversity using Yule's statistic. The study developed two MCSS models to classify paddy crops and weeds from digital

images. Their proposed system can be applied to publicly available paddy crop data and aims to recommend appropriate herbicides for different weed types based on the classification results.

Bai et al. (2020) conducted their survey on object detection recognition and robot grasping based on machine learning. As this research is in the image processing field, machine learning plays a major role there. Convolutional Neural Networks realize the training of large scale image datasets. In this research, they applied machine learning, and machine vision to various image processing tasks, such as image detection, target detection, etc. The article contains information about how they use CNN to analyze and process. When compared with other image processing algorithms, CNN gives the advantage of having no processing requirements for detecting the target. Rechay et al. (2021) have focused on neural networks to detect disease in maize due to its economic significance. Weeds are a major problem in agriculture yield management. They proposed a smartphone application to identify maize crop disease in plants by using the dataset. As they mention, the detection accuracy is 83%. The work shows that they used the image Net dataset as a benchmark for computer vision. That dataset contains above 1000 items. Their defined model was developed using Python. The holdout methodology is used to evaluate results to separate the dataset to 80% training data and 20% for validating data. They did the testing part on a Ryzen 5 1500 6-core workstation. They develop the neural network model by developing two classes, healthy and diseased. They used the library TensorFlow Google developing model to design a deep learning model. The source library was written using Python. The backend was developed by using the TensorFlow backend. It is an open-source deep learning framework for developing mobile devices. The IDE that they used is Android Studio.

According to Roahn et al. (2011) in Sri Lanka, 50% of crop yields are reduced because of the weeds. Their research article shows how weeds affected paddy harvest and weed types. As they mentioned in the paper, there are 16 weed species.

According to Szegedy, Toshev and Erhan (2012) Image processing using Deep Learning Neural Networks is the most commonly used method in object detection. In the research article, the authors proposed a system to object detection in various classes using a formulation. Their proposed system is high-resolution object

detection that can be used to detect images using DNN. The authors focused on DNN for object detection in a larger number of datasets and formulated a DNN based regression to get a binary mask as the result. They used DNN mask detection in a multiscale fashion to increase accuracy. CNN layers were used to detect images from the dataset. CNN regression layer was used to generate the binary mask. To get the most accurate output, the dataset images were divided into N number of pixels. The fixed size of the N is equal to $d*d$. According to the paper, the authors faced some challenges because of the image sizes and the limitation of the output cell size. If the image size with $400*400$ and the $d=24$, the image can't be applied to the $16*16$ cell. That problem was fixed using the multi-mask Robust localization. The proposed system was trained using 5 masks. The mask size is always larger than the image size. They used 1000 data sets for the model training, and the image set was divided into 60% negative and 40% positive datasets. The future work is to formulate the proposed system to use for a larger number of classes.

Sambolek and Ivasic-kos (2021) proposed system is a model that can be used in SAR operations to detect persons. Their research is based on an automatic person detection system using CNN and YOLO V4. According to the authors, the automatic object detection of images and person detection are commonly used. So the article is mainly focused on technologies such as R-CNN and YOLO V4. Related work shows that the YOLO V4 is the fastest and has high accuracy with small false detections. The researchers used CNN YOLOV4 to detect people, and as the dataset, they used the SARD data set. The paper describes that YOLO was selected as the tester because of its high accuracy. Also, the authors checked the transferring setting of YOLO V4 and used YOLO and Deep CNN to get the results. So the images in the dataset were divided into $S*S$ frames and used the typical deep learning algorithms. The proposed system used a YOLO detector within a deep residual network with

53 layers. Also, the YOLO models were trained on the Google Colab service. In the experiment, they compared YOLOV4 with other testers, and the accuracy of the YOLO tester was 96%. Compared with the other testers, the best accuracy level is with the YOLO. So because of the accuracy level, they developed the system using the YOLO (SARD). The model was trained on $512*512$ image resolution, but the $832*832$ resolutions were used for the best results. For future work, they will develop a thermal camera to increase

detection performance and recognize human activities. Zhang et al. (2021) proposed a system that can be used to GPR-B scanned images from rail infrastructure methods for object detection using faster RCNN, SSD, and YOLO V2. The system accuracy level is 97%. The authors have proposed a GAN-based deep learning framework to detect the hyperbolas automatically. The proposed framework has two parts: data generation and object detection based on deep learning. The main purpose of the system is to generate an image when the random noise is input. In the object detection part, the authors have proposed a one-stage detection model. The article shows that the CNN YOLO can be used as an object detection system to get more secure. Classification of a single pixel converts the problem of object detection into semantic segmentation (Lin et al., 2020). The authors mentioned that YOLO V3 has more accuracy than other test methods because it depends on multi-scale fusion. The proposed system is to remove anchors that detect objects based on two key points. One stage method is used because of the problem that occurs in the two stage object detection model. In this work, YOLO is used as the tester, and the image is divided into $s*s$ cells. A cell center is called a grid cell. The grid cell was responsible for detecting the objects. Images were given with $511*511$ size, and the output is $64*64$ offsets. According to the authors CNN can be used to deep learning based well control object detection systems. The network framework that they used is like below.

From the research conducted by Kristo, Ivasic-Kos and Pobar (2020), they proposed a system of automatic person detection in thermal images. The used methodology to detect objects was CNN model training. According to this research, YOLO V3 was the fastest performance tester that can be used for object detection. To evaluate the best detection performance, they designed an original dataset and trained a deep learning model. The images were detected at the state of the art level. The proposed system was developed using an adaptive Boolean Based Saliency (ABMS) kernel with a YOLO detector. The data set was 4000 and the image size was $608*608$ pixels. The system was trained without multi-scale training. The resulting output was a 90% person class with RGB images, and the model was trained on a 3000 training data image set. The performance of the model train was succeeded with the YOLO V3. And the trained dataset was COCORGB. In future studies, the authors will plan to develop an application to detect persons and non-human objects in different weather conditions.

Ratnasekera (2015) Reviewed how weedy rice is a threat to rice production, distribution, and strategies for weedy rice management. The research identifies weedy rice as one of the four most harmful weeds affecting rice fields globally. In the mid-1990s, it was first recognized as a problem in the Vavuniya, Ampara, and Batticaloa districts. This paper is valuable for our research as it discusses the unique traits of weedy rice. It highlights that weedy rice is difficult to distinguish from cultivated rice at the seedling stage due to their similar appearance. The limitation of this research paper is it does not talk about the weedy rice management practices. But it gives a clear idea about how weedy rice has spread throughout Sri Lanka paddy fields and its morphological and genetic diversity. This research paper is not directly related to our research topic, but it is helpful to study weedy rice.

Even though many individuals have been trying to provide a solution in recognizing weeds in the crop using various methods for several years, no system has made a business breakthrough yet. Considering this situation, Jaiganesh et al. (2020) proposed a model for plant identification by plant leaves using a deep learning technique - CNN classifier. They have used a dataset that is available on Kaggle. The dataset includes around 960 distinct plants from 12 different species, captured at various growth stages. The model achieved an accuracy of 82% on the training set, with a validation accuracy (for plant identification) of 86%.

One of the best CNN algorithms, YOLO, is good at solving object detection in the most simple and highly efficient way (Du, 2018). The paper describes the new directions of the YOLO, YOLO versions (V1, V2, etc.), CNN, the layers of CNN, CNN algorithms, and the limitations of CNN. Classification and localization and detection are the tasks of the image processing technology. In image processing model training, the most occurring problems are accuracy, speed and cost. Until 2012, CNN reduced the error rate from 26% to 15.3%. Then CNN developed in two directions called normalization and optimization. Further, the researchers have compared Faster R-CNN and YOLO V2. The performance of the detection systems has been compared with mAP (mean average precision) and FPS (frames per second). Compared with Faster R-CNN, YOLO has more advanced applications in practice. Fast YOLO is the fastest general-purpose object detector. YOLO's FPS 155 and its mAP can also reach up to 78.6, surpassing the performance of Faster R-CNN greatly. The comparisons confirm that the

YOLO is a suitable algorithm for object detection research, and its performance, accuracy levels are higher than other detection systems. The limitation of YOLO is that YOLO struggles to generalize to objects in new or unusual aspects of ration or configuration. And there are shortages with its loss function errors. But YOLO can achieve high precision and keep real time for pictures with high resolution. YOLO is a unified object detection model. YOLO V2 provides state-of-the-art with the best tradeoff between the best accuracy and real time speed for object detection than other detection systems.

Overuse of herbicides in paddy fields leads to increased production costs and environmental pollution. To address this, it's essential to detect the location of rice seedlings and weeds for targeted weed management (Ma et al., 2019). The researchers propose a fast and robust image segmentation method for identifying rice seedlings and weeds at the seedling stage using SegNet, where these plants often overlap. The study focuses on two main objectives: introducing a semantic segmentation method based on encoder and decoder architecture, and comparing its performance with classical segmentation models, specifically FCN and U-Net. SegNet, a deep Convolutional Neural Network, is utilized for image segmentation, offering lower computational cost and higher precision compared to FCN. For the study, 28 RGB images were captured around 20 days after the rice seedlings emerged. The images, taken in paddy fields with weeds in early growth stages, were divided into smaller tiles, totaling 224 images. Of these, 80% were used for training and 20% for testing. The SegNet model was trained using transfer learning.

The results showed that SegNet achieved higher classification accuracy, with an average accuracy rate of 92.7%. In comparison, the FCN and U-Net models had average accuracies of 89.5% and 70.8%, respectively. In Sri Lanka, more than 142 weed species have been identified in rice fields. (Rao et al., 2017) The paper speaks about the methods (Manual, mechanical, tillage, mulching) of weed control used in South Asian countries. Manual weeding and submergence were the main weed control techniques used in Sri Lanka until the early 1960s. Then herbicides became more popular. As the single weed control approach is inefficient, Integrate Weed Management is needed to keep weeds below an economic threshold level.

The system proposed by Kulkarni and Angadi (2019) is about detecting weeds and crops using CNN and IoT. The proposed method is to train a large number of images of weeds and crops using

CNN. The trained CNN model is trained by getting images from the camera sent to the Raspberry pi. Raspberry performs image segmentation by dividing the image into small cells. Each CNN model classifies as weeds or crops. The system was trained using 250 image data sets. The accuracy rate is 85%, and the false ration is 7%. The proposed system consists of a Raspberry pi and a camera. The camera was used for image processing and segmentation. The results of the system are obtained with an average accuracy of 85%. So the proposed framework can be used by farmers to check whether the growth of weeds. The article shows that CNN can be combined and implemented in weed controlling to get excellent results. In the research of Liu et al. (2015) they used an algorithm consisting of two dimensional image information. The algorithm is focused on two main processes like convolution and sampling. The authors used the CNN subsampling method by sampling by time or space. The subsampling structure by time space was used to achieve some degree of scale and deformation displacement. The designed algorithm is based on gray image as input of 96*96 size, that turned in to 32*32 size of the images. The model was trained on the 7 convolutional layers and the results were more accurate.

The system proposed by Islam et al. (2021) is a machine learning algorithm to weed detection. The paper is organized with an overview of machine learning algorithms that can be used to weed detection in Australian Chili fields. The proposed system aims to detect weeds by using image processing and machine learning. The used data set was preprocessed using image processing, KNN based studies. Weed detection is the purpose of the proposed system. According to the paper, KNN offers a 63% percentage of RF 96%, and the SVM offers 94% accuracy in the weed detection proposed system. Their future work will be a deep learning algorithm to increase the accuracy of weed detection

III. EXPERIMENTAL WORKFLOW

In this section, the problem under investigation is explicated, providing context and emphasizing the study's significance. The existing knowledge gap is outlined, and the research objectives are articulated. By framing the problem, a foundation is laid for the subsequent sections, highlighting the relevance of the research.

A. Problem

In the present world, technology has involved every industry making their work more accessible and speedy. But when comparing to other industries, we noticed that no significant change can be seen in the farming industry. Especially in the Asia countries like Sri Lanka. Sri Lanka is considered as an agricultural country, and rice is the staple diet and the single most important crop in Sri Lanka (Senanayake and Premaratne, 2016). However, weeds in the rice fields are one of the major problems the Sri Lankan farmers face (Perera and Dahanayaka, 2015).

There are more than 120 weed species that can be identified that belong to 32 families (Gunasena, 1992). It is a challenging task to detect the weeds and select suitable weed control methods for each weed species. So, weeds have become a major problem to reduce the harvest of paddy cultivation. Further, the Sri Lankan younger generation is also interested in farming but they are not familiar with many weed species and the past weed controlling methods, and they have no knowledge or experience to continue their farming. So, we have identified that not having enough knowledge on weed detection and weed control methods is a reason behind this situation and the spread of weeds. Even though the Sri Lankan farmers are facing this problem from the earlier days, there is not enough support from modern technology to overcome this situation. So, it is clear that there should be more support from the technology to the Sri Lankan farmers and young generation, students, and researchers to identify the weed species in rice fields and suggest suitable weed controlling methods. The research is based on the above problems and proposes a model with a higher level of accuracy to weed detection. In addition, it is planned to provide weed control methods for the identified weed species.

B. Data Set Creation

The dataset was created using a Redmi Note8 48mp camera, Samsung S7 12mp camera, Samsung m12 48mp camera, and Nikon D750 camera. There were 3933 images belonging to 20 weed types. The data set was 3933 images belonging to 20 classes. In the dataset training process, the same data set was trained with the same parameters with all the 03 models. Google Colab idle environment was used as the idle environment to train the data set. The data set was created by cropping all images as squares and setting the pixel size to 224 x 224 for all and setting to auto-arrange white balance and high ISO

normalisation. After that, the images were divided into 20 classes, and then those sets were divided into 2 groups as train and test datasets.

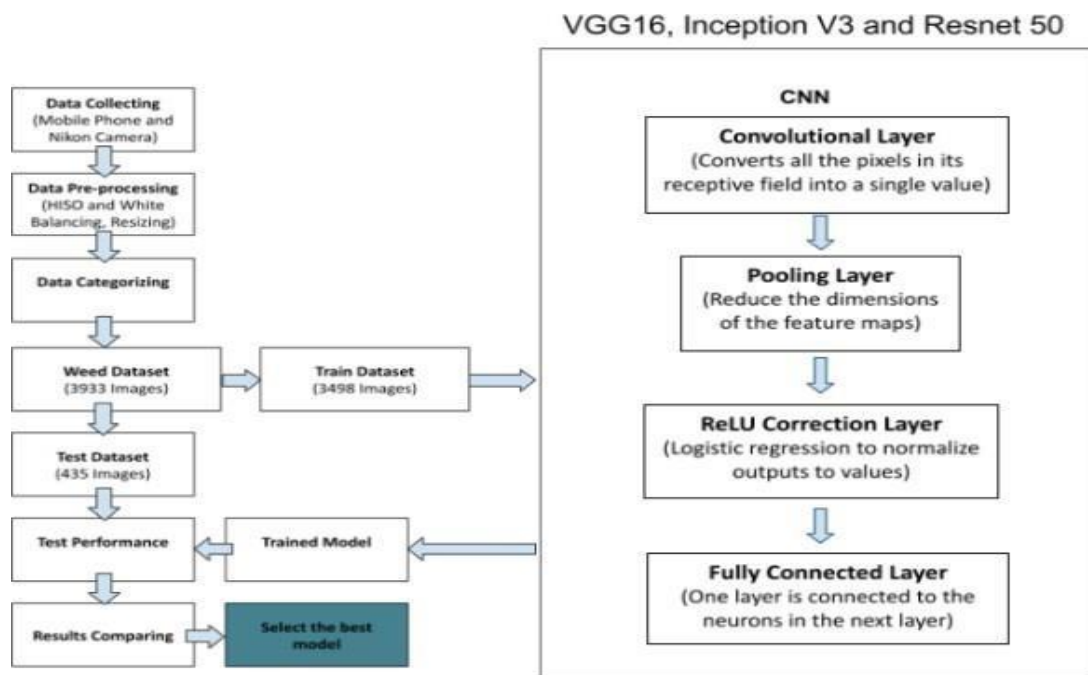


Figure 02: Model Training Flow Diagram

C. Data Set Training with Transfer Learning VGG16

The selected data set was pre-processed before using this model as $(224 * 224 * 3)$. That model has 1,383 million trainable weights and 16 layers (convolution layers 3×3 filter, max pool layers 2×2 filter, 2 fully connected layers, and Softmax layer). The basic Architecture of a VGG16 is represented in Fig. 3.

Inception V3

Inception v3 is a pre-trained CNN model that Consists 48 layers deep{(Convolutional layers 4 $[1 * 1]$, $[3 * 3]$, $[5 * 5]$), MaxPooling layers $[3 * 3]$, Fully connected layer 1} and 23,8 million trainable weights. It was trained on more than a million images in the ImageNet database. All images in the ImageNet database have a fixed size of $224 * 224$ and have

RGB channels (3) therefore we had to pre-process the images as we did in VGG16 ($224 * 224$).

ResNet 50

Microsoft introduced a deep residual learning framework (Resnet 50) to overcome the problem that occurs when adding more layers to a deep network may cause a higher training error rate. The $[F(x)+x]$ formula makes shortcut connections to skip one or more layers. Then ResNets can get high accuracy when increased depth without training error.

IV. RESULTS

The image dataset was trained using the three models (VGG16, Inception v3, and ResNet50) two times. On the first try, only the VGG16 performed well and the other two models did not perform as expected. On the second try, both VGG16 and inception V3 performed well (accuracy of 98% and 99%), but the accuracy of ResNet 50 was not enough (75%). VGG16 was selected as the best model from VGG16 and Inception v3 by considering the accuracy and loss. Both models gave the best results but the validation loss of Inception V3 is higher than VGG16. And also the accuracy of the VGG 16

model was 100%. So VGG16 was selected as the best model for the weed detection system.

Data Training Results

Data training Results with VGG16

- Validation accuracy - 75%
- Validation loss - 0.69
- Model loss - 0.63
- Model accuracy - 77%

Data training Results with Resnet50

- Validation accuracy - 98%
- Validation loss - 0.0464
- Model accuracy - 100%

- Model loss - 0.0025
- Epoch - 15

Data training Results with Inception V3

- Validation accuracy - 99%
- Validation loss - 0.1076
- Model accuracy - 99%
- Model loss - 0.0053
- Epoch - 15

B. Primary Detection Results

After considering the data training results, VGG16 model was selected as the image dataset training model. Results are shown in the TABLE.I

Table 01: Data Training Results

Model Name	Model Accuracy	Model Loss	Validation Accuracy	Validation Loss
Resnet 50	77%	0.63	75%	0.69
VGG16	100%	0.0025	98%	0.0464
Inception V3	98%	0.0053	99%	0.1076

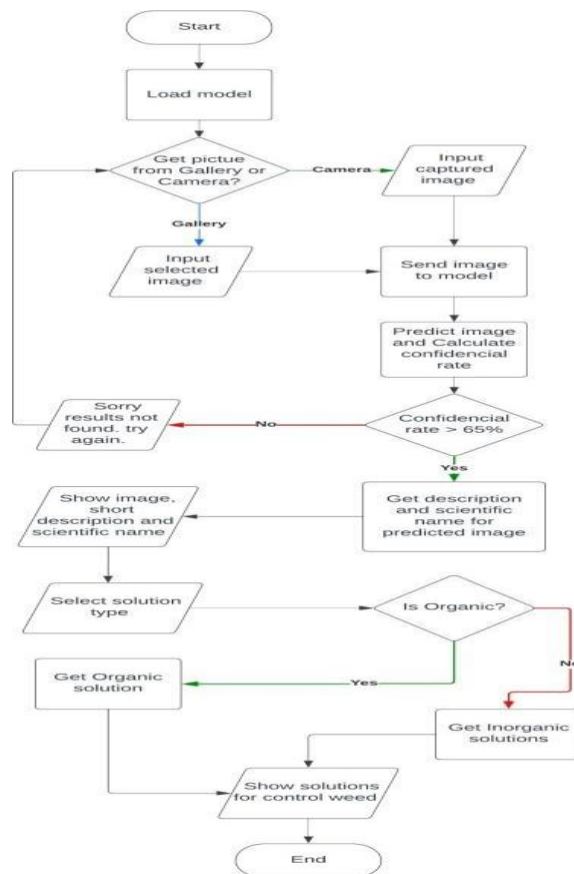


Figure 03: Mobile Application Data Flow Diagram

C. Mobile Application

In this research, we propose a mobile application designed to assist farmers in identifying weeds in agricultural fields. The app allows users to upload or capture images of weeds using their mobile devices. It employs image recognition technology to identify the weed type and provides detailed information on how it impacts the harvest, including potential yield loss. Furthermore, it offers step-by-step guidance on both

chemical and organic methods to manage the weeds effectively. As paddy fields often lack internet connectivity, the dataset is stored locally within the application's database, ensuring farmers can access and use the app even in remote areas without a network connection. This solution aims to enhance farming efficiency by delivering real-time, actionable insights directly to farmers.

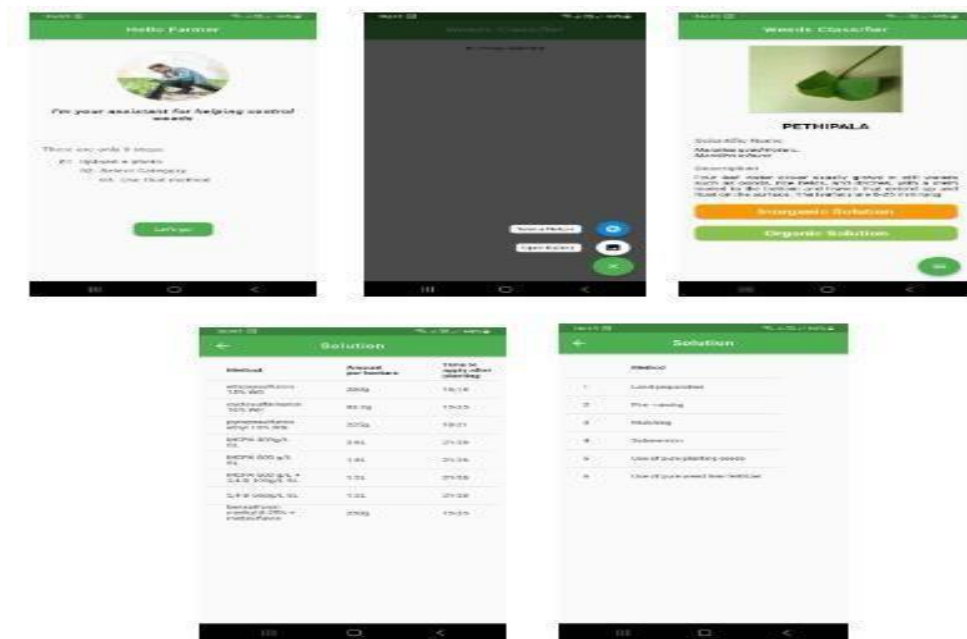


Figure 04: Mobile Application Interface

V. DISCUSSION

The primary objective of this project was to achieve accurate weed species detection in paddy fields. The developed mobile application, powered by the selected Convolutional Neural Network (CNN) model, successfully identifies six distinct weed types and offers tailored weed control methods for each. The literature review highlighted the longstanding challenges Sri Lankan farmers face in combatting weeds, with existing solutions relying heavily on traditional knowledge and experiences. The system created through this research serves as a significant support mechanism for farmers, the younger generation, students, and researchers engaged in weed detection and control methods. By leveraging advanced technology, the developed application not only addresses the immediate concerns of weed infestation in paddy fields but also contributes to the knowledge base within the agricultural community. This initiative represents a crucial step towards providing practical and efficient solutions to a persistent problem in Sri Lanka's agricultural landscape.

VI. CONCLUSION

In the course of this research, we proposed a machine learning-based system designed for the detection of various weed types in paddy fields, coupled with the provision of effective weed control methods. Drawing from pertinent literature, it became evident that Convolutional

Neural Networks (CNNs) have proven success in object detection. In our study, the specifically employed VGG16 model demonstrated notable efficacy in identifying weeds in rice fields, achieving an impressive 97% validation accuracy, 0.0464 validation loss, 100% model accuracy, and a minimal 0.0025 model loss.

Through the developed application, we successfully implemented the VGG16 model to detect six selected weed types with high accuracy. Notably, the system can be further enhanced by expanding the range of detectable weed types. This involves refining the image dataset with higher-quality and more informative images. Subsequently, the dataset should undergo training with the VGG16 model, exploring optimal epoch sizes to maximize accuracy. This iterative process allows for the continuous improvement of the system's weed detection capabilities, making it adaptable to a broader spectrum of weed types in paddy fields.

VII. FUTURE WORKS

Here at this stage, the mobile application only shows 06 weed types from the selected weed types. The descriptions of those 06 weed types and the weed control solutions for them can be added within the same source code. For our future work, the application will be broadened by adding more weed types and we can create or link the application with external resources. But if the application was connected with such an external

server, the ability to work offline would be lost. In this stage, the developed mobile application can provide offline services. Further, trying to enhance the user experience by making some changes to the designed user interfaces. For now, the application is only compatible with the English language, but it will be developed as compatible with the other languages.

REFERENCES

Du, J. (2018) 'Understanding of Object Detection Based on CNN Family and YOLO', *Journal of Physics: Conference Series*, 1004(1). doi:10.1088/1742-6596/1004/1/012029.

Jaiganesh, M. *et al.* (2020) 'Identification of Plant Species using CNN- Classifier', *Journal Of Critical Reviews*, 7(3), pp. 923–931.

Ratnasekera, D. (2015) 'Weedy rice: A threat to rice production in Sri Lanka', *Journal of the University of Ruhuna*, 3(1), p. 2. doi:10.4038/jur.v3i1.7859.

Rao, A. *et al.* (2017) *An overview of weeds and weed management in rice of South Asia*. Available at: <http://oar.icrisat.org/10211/>.

Ma, X. *et al.* (2019) 'Fully convolutional network for rice seedling and weed image segmentation at the seedling stage in paddy fields', *PLoS ONE*, 14(4). doi:10.1371/journal.pone.0215676.

Szegedy, C., Toshev, A. and Erhan, D. (2013) 'Deep Neural Networks for object detection', *Advances in Neural Information Processing Systems*, pp. 1–9.

Richey, B. *et al.* (2020) 'Real-time detection of maize crop disease via a deep learning-based smartphone app', (April 2020), p. 10. doi:10.1117/12.2557317.

Kamath, R., Balachandra, M. and Prabhu, S. (2020) 'Paddy Crop and Weed Discrimination: A Multiple Classifier System Approach', *International Journal of Agronomy*, 2020. doi:10.1155/2020/6474536.

Rajapakse, R. *et al.* (2012) 'Planning for effective weed management: lessons from Sri Lanka', *Pakistan Journal of Weed Science Research*, 18(Special Issue), pp. 843–853.

Zhang, X. *et al.* (2021) 'A Gans-Based Deep Learning Framework for Automatic Subsurface Object Recognition from Ground Penetrating Radar Data', *IEEE Access*, 9, pp.

39009–39018.

doi:10.1109/ACCESS.2021.3064205.

Sambolek, S. and Ivasic-Kos, M. (2021) 'Automatic person detection in search and rescue operations using deep CNN detectors', *IEEE Access*, 9, pp. 37905–37922. doi:10.1109/ACCESS.2021.3063681.

Bai, Q. *et al.* (2020) 'Object detection recognition and robot grasping based on machine learning: A survey', *IEEE Access*, 8, pp. 181855–181879. doi:10.1109/ACCESS.2020.3028740.

Lin, Y. *et al.* (2020) 'Semantic Segmentation with Oblique Convolution for Object Detection', *IEEE Access*, 8, pp. 25326–25334. doi:10.1109/ACCESS.2020.2971058.

Kristo, M., Ivasic-Kos, M. and Pobar, M. (2020) 'Thermal Object Detection in Difficult Weather Conditions Using YOLO', *IEEE Access*, 8, pp. 125459–125476. doi:10.1109/ACCESS.2020.3007481.

Wu, Z. *et al.* (2021) 'Review of weed detection methods based on computer vision', *Sensors*, 21(11), pp. 1–23. <https://doi.org/10.3390/s21113647>

Ofori, M. and El-Gayar, O. (2021) 'An approach for weed detection using CNNs and transfer learning', *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2020-January, pp. 888–895. doi:10.24251/hicss.2021.109.

Perera, P.C.D. and Dahanayake, N. (2015) 'Review of major abundant weeds of cultivation in Sri Lanka', *International journal of scientific and research publications*, 5(5), pp. 1–9.

Islam, N. *et al.* (2021) 'Early weed detection using image processing and machine learning techniques in an australian chilli farm', *Agriculture (Switzerland)*, 11(5). doi:10.3390/agriculture11050387.

S.K. and . S.. A. (2019) 'Iot Based Weed Detection Using Image Processing and Cnn', *International Journal of Engineering Applied Sciences and Technology*, 4(3), pp. 606–609. doi:10.33564/ijeast.2019.v04i03.089.

Liu, T. *et al.* (2015) 'Implementation of Training Convolutional Neural Networks'. Available at: <http://arxiv.org/abs/1506.01195>.

Gunasena,H,P,M(1992 Weed Research in Sri Lanka and Annotated Bibliography, Department Of Agriculture Peradeniya, Sri Lanka,p.143

Naglot, D., Kasliwal, P.S., Gaikwad, S.J. and Agrawal, N.D. (2019). Indian Plant Recognition System Using

Convolutional Neural Network. *International Journal of Computer Sciences and Engineering*, 7(6), pp.276–280.

Jaderberg, M., Simonyan, K., Vedaldi, A. and Zisserman, A. (2015). Reading Text in the Wild with Convolutional Neural Networks. *International Journal of Computer Vision*, 116(1), pp.1–20.

Pioneering Disease Prediction in Cinnamon Leaves using Machine Learning: A Systematic Literature Review

D.A.S. Dilhari¹ and A. Mohamed Aslam Sujah²

^{1,2}Department of Information and Communication Technology, South Eastern University of Sri Lanka, Sri Lanka

¹sisaradilhari.1998@gmail.com, ²ameersujah@seu.ac.lk

Abstract

The integration of Machine Learning (ML) in agricultural disease prediction has become increasingly prominent. This review paper explores the evolution of techniques used for predicting diseases in cinnamon leaves and analyzes common cinnamon leaf diseases, drawing on research conducted up to 2023. The paper highlights the evolution of ML methodologies, particularly in the areas of image processing, feature extraction, and classification algorithms. It provides an in-depth analysis of various approaches, such as Convolutional Neural Networks (CNNs), Support Vector Machines (SVMs), and Random Forests, evaluating their effectiveness in disease prediction. From an initial set of 100 studies, 22 were selected for detailed analysis based on their relevance and contribution to the field. Additionally, the review addresses the challenges associated with developing reliable ML models. Through the synthesis of findings from multiple studies, this paper offers a comprehensive overview of current research in cinnamon leaf disease and prediction, identifying existing gaps and proposing directions for future investigations to improve the precision and applicability of ML-driven solutions in agriculture.

Keywords: Cinnamon Leaf Diseases, Machine Learning, Agricultural Disease Prediction, Classification Algorithms

I. INTRODUCTION

In Cinnamon, known locally as "Kurundu" in Sri Lanka, has been a valuable commodity since antiquity, cherished for its unique aroma, flavor, and medicinal properties. Derived from the inner bark of various species within the Cinnamomum genus, cinnamon has a rich history intertwined with trade, culture, and culinary traditions. Among the many species, Cinnamomum verum, commonly referred to as Ceylon cinnamon or "true" cinnamon, holds a place of particular

significance. Native to Sri Lanka, Ceylon cinnamon is distinguished by its superior quality, which has earned it a prominent position in both local and global markets (Wickramasinghe *et al.*, 2018). The cultivation of cinnamon in Sri Lanka dates back several centuries, with the Dutch colonial rulers playing a crucial role in establishing systematic cultivation practices in the 18th century. The Dutch Governor Falk was instrumental in promoting cinnamon cultivation, which soon became one of the island's most lucrative exports. By 1841, the demand for cinnamon had surged, leading to the commercial production of cinnamon leaf oil, by product that further expanded the industry's economic footprint (Wijesekera and Chichester, 1978).

Today, cinnamon is cultivated across approximately 14,000 to 16,000 hectares in Sri Lanka, with the Ambalangoda region being a key production area. This region alone accounts for nearly 50% of the country's total cinnamon output, underscoring its importance in the national economy. Ceylon cinnamon's economic value extends beyond its use as a spice. The chemical composition of Ceylon cinnamon, particularly its low levels of coumarin, makes it highly sought after in the global market. Coumarin, a naturally occurring compound found in higher concentrations in other cinnamon varieties like Cassia, can be harmful in large doses, which enhances the appeal of Ceylon cinnamon for health-conscious consumers. Additionally, the extracts from cinnamon leaves and bark are used in the food industry as natural preservatives due to their antimicrobial properties, as well as in the pharmaceutical and cosmetic industries for their therapeutic benefits (Suriyagoda *et al.*, 2021).

Despite its global recognition and economic importance, the cinnamon industry in Sri Lanka faces significant challenges, particularly from pests and diseases that threaten both the yield and quality of the crop. Among the most common

foliar diseases are leaf blight, caused by *Colletotrichum gloeosporioides*, and algal leaf spot, caused by *Cephaleuros virescens*. These diseases, along with infestations by insect pests like jumping plant louse/ leaf galls, thrips attack, can lead to severe reductions in cinnamon yield, causing substantial economic losses for farmers (Rajapakse and Kumara, 2007). Traditional disease management strategies, including the use of fungicides and pest control measures, have proven to be only partially effective. These methods are often labor-intensive, environmentally harmful, and unsustainable in the long term, necessitating the exploration of more advanced and targeted approaches (Jayasinghe *et al.*, 2020).

Beyond its challenges in cultivation, cinnamon's importance extends far beyond its role as a spice. Owing to its special properties, cinnamon is a multipurpose ingredient widely used not only in kitchens as a tasty addition to various dishes but also in medicine (Pathirana and Senaratne, 2020). The essential oil derived from cinnamon leaf, which contains a high concentration of trans-cinnamaldehyde, possesses strong antibacterial properties. These properties are effective against infections in plants and animals, as well as bacteria and fungi associated with food spoilage and food poisoning. In addition to its culinary applications, cinnamon offers numerous health advantages, including anti-inflammatory properties, antimicrobial activity, a reduced risk of cardiovascular disease, improved cognitive function, and a decreased chance of colon cancer. The various parts of the cinnamon plant, including the outer bark, inner bark, and leaves, are used for medicinal purposes.

In this context, the advent of Machine Learning (ML) presents a transformative opportunity for the cinnamon industry. ML, a subset of artificial intelligence (AI), involves the use of algorithms and statistical models to analyze large datasets, identify patterns, and predict outcomes with a high degree of accuracy. The application of machine learning (ML) in cinnamon disease management is still in its early stages. Some studies have made strides in related areas, such as using image processing techniques and algorithms like Speeded up Robust Features (SURF) for data extraction from cinnamon (Chandima and Kartheeswaran, 2016; Sunitha *et al.*, 2022). These techniques have been employed to predict the

maturity levels of cinnamon trees using classifiers such as Support Vector Machines (SVMs). While these initial findings are promising, there remains a significant gap in research specifically targeting the detection and management of diseases in cinnamon leaves using ML techniques.

In recent years, the application of Machine Learning (ML) has emerged as a transformative opportunity for addressing these challenges. ML, a subset of Artificial Intelligence (AI), allows for the analysis of large datasets, enabling accurate predictions and early detection of diseases through pattern recognition (Gunasekara *et al.*, 2021; Shandilya *et al.*, 2024). Techniques like deep learning and Convolutional Neural Networks (CNNs) have shown significant promise in detecting diseases in cinnamon crops by leveraging image processing technologies. For instance, the potential of CNNs for characterizing cinnamon diseases such as rough bark and stripe canker, providing a model for future applications in this field (Jayasena *et al.*, 2023). Recent reviews have emphasized the growing interest in the application of ML in the cinnamon industry. These studies highlight the potential of deep learning models, particularly Convolutional Neural Networks (CNNs), in improving disease detection and management for cinnamon crops (Giraddi, Desai and Deshpande, 2020; Feltes *et al.*, 2023). Other research has discussed the integration of remote sensing technologies with ML, which offers greater precision in monitoring crop health and managing diseases (T* *et al.*, 2020; Tusher *et al.*, 2022). Additionally, advanced ML techniques such as Transfer Learning have shown potential in addressing key challenges, particularly in disease detection and quality control within the cinnamon industry (Fatima *et al.*, 2021).

This systematic literature review aims to explore the current state of research on cinnamon leaf diseases, their management strategies, and the use of ML in disease detection and prediction. By examining studies conducted up to 2023, the review seeks to evaluate the effectiveness of ML techniques in this domain, assess the impact of these technological advances on the cinnamon industry, and identify areas that require further research. The findings of this review are expected to provide valuable insights for researchers, practitioners, and policymakers interested in improving the sustainability and efficiency of cinnamon production through the integration of

ML technologies. By bridging the gap between traditional practices and technological advancements, this review aims to contribute to the ongoing efforts to enhance the resilience and productivity of the cinnamon industry in Sri Lanka and beyond.

II. METHODOLOGY

A. Systematic Literature Review

The This study adopts a systematic literature review (SLR) methodology to explore the application of machine learning (ML) in the prediction and management of diseases in cinnamon leaves. The review process was structured into three key phases: planning, conducting, and reporting.

In the planning phase, we identified relevant electronic databases, including IEEE Xplore, Springer, and ACM Digital Library, Google Scholar to source research papers related to our study. Key research questions were defined to guide the review, and specific search strings were developed based on keywords pertinent to the intersection of cinnamon leaf diseases and machine learning. During the conducting phase, we systematically searched the selected databases using the predefined search strings. The search results were carefully reviewed to select studies that addressed our research focus. Additionally, references from these studies were used to perform snowballing, ensuring a comprehensive inclusion of relevant literature. The studies were analyzed to extract essential information such as abstracts, keywords, methodologies, and findings, which were then categorized for further evaluation. In the reporting phase, the selected studies were synthesized and organized into a detailed analysis. The studies were documented with a focus on their contributions to the understanding of ML applications in cinnamon leaf disease prediction. The summarized research was then compiled into structured documents that provide a comprehensive overview of the literature, highlighting key trends, methodologies and future directions in the field. This systematic review process, guided by established SLR protocols, ensures a thorough and objective evaluation of the existing literature, offering valuable insights into the potential of machine learning in enhancing disease management strategies for the cinnamon industry.

B. Research Questions

Research questions are central to guiding a systematic literature review. Table 01 presents the research questions that this study aims to address. By examining these questions, we can identify gaps in the current literature and better understand the state of research in the application of machine learning for disease prediction in cinnamon leaves.

Table 01: Research questions

No	Research Question
RQ1	What are the common cinnamon leaf diseases identified in existing studies?
RQ2	How can various machine learning techniques be effectively utilized to create a predictive model for accurately identify cinnamon diseases?
RQ3	What challenges and limitations have been identified in past studies on machine learning applications for leaf disease prediction?

C. Study Selection

The study selection process involved several key steps:

1) Terms and Search Strings:

The search terms were applied across two main segments: cinnamon leaf disease and machine learning. The search string was applied to three metadata fields: title, abstract, and keywords. Table 02 represents the search strings applied in the databases.

Table 02: Search terms of the mapping study on pioneering disease prediction in cinnamon leaves using machine learning

Area	Search Terms
Cinnamon Disease	"Cinnamon leaf disease", "cinnamon diseases "
Disease Identification	"Leaf disease identification using machine learning", "plant disease detection", "plant disease classification"
Machine Learning	"Machine learning", "ML approaches", "artificial intelligence in agriculture"
Cinnamon disease management	"Cinnamon leaf disease management", "cinnamon leaf disease treatment"
Search String	("cinnamon leaf disease" OR "cinnamon disease") AND ("leaf disease identification using machine learning" OR "plant disease detection" OR "plant disease classification") AND ("machine learning" OR "ML approaches" OR "artificial intelligence in agriculture") AND ("cinnamon leaf disease management", "cinnamon leaf disease treatment")

3) Sources:

This Systematic Literature Review was performed using the following electronic databases and considered the most relevant studies.

- i. IEEE Xplore <<http://ieeexplore.ieee.org>>
- ii. Springer Link <<https://link.springer.com>>
- iii. Science Direct<<https://www.sciencedirect.com>>
- iv. Google Scholar<<https://www.sciencedirect.com>>
- v. ACM Digital Library<<https://dl.acm.org>>

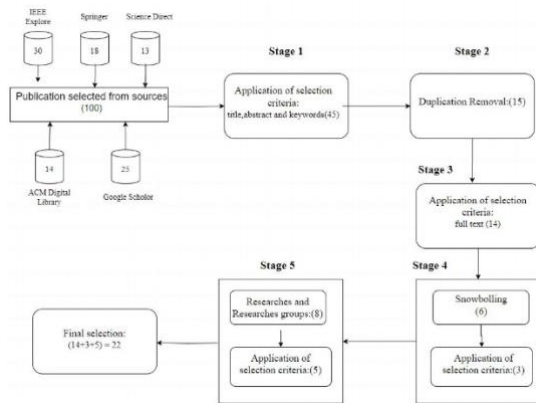


Figure 01: Selection Process Flow

4) Inclusion and exclusion criteria:

The selection process for this study was organized based on two inclusion criterion and five exclusion criteria. Table 03 and Table 04 depict the inclusion and exclusion criteria used in the filtering process respectively.

Table 03: Inclusion criteria of the selection process

No	Inclusion criteria (IC)
IC1	Focuses on common cinnamon leaf diseases.
IC2	Describes the application of machine learning techniques for the prediction of cinnamon plant diseases.

Table 04: Exclusion criteria of the selection process

No	Exclusion criteria (EC)
EC1	The paper does not contain an abstract
EC2	The paper is published only as an abstract
EC3	The paper is not written in English
EC4	The paper is an earlier version of a study that has already been selected
EC5	The paper is not a primary study. It is either editorial or summaries of keynotes and tutorials

5) Data extraction and synthesis:

In this study, a thorough examination of publications to 2023 was conducted to evaluate the application of machine learning techniques in predicting diseases in cinnamon leaves. The initial retrieval process gathered 100 publications from several reputable digital libraries, including IEEE Xplore, Springer, ScienceDirect, ACM Digital Library, and Google Scholar.

The selection process for these studies was organized into five stages. The first stage involved applying predefined selection criteria to the title, abstract, and keywords of each study. In the first stage, 45 of these publications were selected based on their titles, abstracts, and keywords. This filtration aimed to distill the most relevant studies pertinent to research topics. In second stage, Identified and removed 15 duplicate studies from the remaining 45, leaving 30 publications for further scrutiny.

In the third stage, an in-depth review of the full texts of these 30 publications was conducted. Employing stringent inclusion and exclusion criteria ensured the relevance and completeness of the studies, which led to 14 publications proceeding to the next stage of the selection process. The fourth stage expanded search to include studies referenced in the initial set, a method known as snowballing. This yielded 6 additional publications, out of which 3 met rigorous selection criteria and added to study pool.

The final stage focused on significant contributions from key researchers and research groups. Identified 8 additional relevant publications through this method. After applying final selection criteria, 5 of these were deemed highly pertinent to research aims. This comprehensive process resulted in a final selection of 22 publications, which were considered for systematic review, ensuring a broad and thorough coverage of the final synthesis.

III. RESULT AND DISCUSSION

This Cinnamon, particularly Ceylon cinnamon, is a vital economic asset for Sri Lanka, valued not only for its culinary and medicinal applications but also for its significant contribution to the agricultural sector (Suriyagoda *et al.*, 2021). However, the industry faces persistent challenges from diseases and pests that impact both yield and quality.

Among the most common cinnamon leaf diseases, leaf blight and leaf spot diseases, primarily caused by the fungus *Colletotrichum gloeosporioides*. This disease manifests as brown or black lesions on leaves, which can spread rapidly, leading to significant foliage loss and a reduction in the plant's photosynthetic ability. Another common issue is algal leaf spot, caused by *Cephaleuros virescens*, which results in orange or reddish spots on the leaves, eventually causing leaf deterioration (Wickramasinghe *et al.*, 2018; Jayasinghe *et al.*, 2020). In addition to fungal infections, cinnamon plants are also affected by various pests. Jumping plant louse and thrips are frequent insect pests that cause leaf galls and deformation, further impacting the plant's health. These diseases and pests represent major challenges to cinnamon cultivation, addressing these challenges effectively is crucial for maintaining the industry's economic viability (Pathirana and Senaratne, 2020).

Recent advancements in machine learning (ML) have demonstrated great potential for revolutionizing agricultural practices, particularly in disease detection and management. However, research specifically applying these methods to predict diseases in cinnamon leaves remains largely unexplored. Most of the existing literature focuses on general applications of ML in agriculture or on detecting diseases in other parts of the cinnamon plant. One of the key challenges in cinnamon leaf disease prediction using ML techniques is the lack of annotated datasets for training robust models. Image processing techniques have been utilized in related applications, such as recognizing mature cinnamon trees, suggesting the possibility of adapting similar methods for identifying diseases in cinnamon leaves (Chandima and Kartheeswaran, 2016). However, without sufficient data, building an accurate ML model remains a challenge.

Transfer learning has been identified as a promising method to address this data scarcity issue. It has been applied to trace adulteration in spices like cinnamon, which suggests the possibility of applying this technique to cinnamon leaf disease detection (Fatima *et al.*, 2021). By leveraging pre-trained models on other plant datasets, researchers can reduce the need for large, specialized datasets specific to cinnamon.

Convolutional Neural Networks (CNNs) have been a popular choice for image-based disease detection. (Sardogan, Tuncer and Ozen, 2018; Singla, Kalavakonda and Senthil, 2024) Techniques such as CNNs with Learning Vector Quantization (LVQ) have been used in plant disease classification, and similar methodologies that could be adapted for cinnamon disease prediction. The potential of CNNs in identifying specific cinnamon plant diseases has been explored, particularly in deep learning models used to identify conditions like rough bark and stripe canker (Pratondo *et al.*, 2024; Shandilya *et al.*, 2024).

The integration of remote sensing with ML models has also been proposed as a solution for large-scale monitoring of plant diseases. By combining remote sensing data with deep learning, precision agriculture in cinnamon farming could be enhanced, enabling more efficient disease monitoring and management (Lakshan S *et al.*, 2023). This integration could address the environmental variability that impacts disease development in cinnamon plants, ensuring that models generalize well across different regions and climates.

Another emerging technology in cinnamon disease prediction is the use of mobile applications for real-time disease detection (Ekanayaka and Kumara, 2022; Jayasena *et al.*, 2023). Mobile-based tools that utilize image processing have been developed to enhance cinnamon quality and health. This tool demonstrates the feasibility of creating mobile-based solutions for disease detection, providing farmers with immediate insights into the health of their crops.

A significant barrier to the widespread adoption of ML in cinnamon disease prediction is the need for interdisciplinary collaboration. Integrating agricultural knowledge with technological innovations is essential, especially considering the unique characteristics of the cinnamon industry in regions like Sri Lanka (Prof. Koliya Pulasinghe, Dr. Dharshana Kasthurirathna and S.A.A. Ravishan, 2023). Highlight that the unique characteristics of the cinnamon industry in Sri Lanka, including its environmental and climatic factors, must be considered when developing ML-based disease prediction systems. Also, there is significant potential for using ML to predict diseases in cinnamon leaves, the field is still in its early stages.

Future research should focus on developing Additionally, Interdisciplinary collaboration datasets specific to cinnamon leaf diseases, between agricultural experts and technologists is exploring the use of transfer learning, and crucial for ensuring that these models are accurate integrating ML with remote sensing technologies. and applicable in real-world farming scenarios.

Table 05: Feature extraction and system results across studies

Paper ID	Title	System Type	Key Features	Results	Limitations
1	Historical overview of the cinnamon industry.	Historical overview Economic impact	Not applicable	Provides historical context	Does not address current disease management practices or specific disease issues.
2	Ceylon cinnamon?: Much more than just a spice	Economic value culinary and medicinal uses	Not applicable	Highlights the multifaceted value of Ceylon cinnamon beyond its culinary uses.	Does not specify common Sri Lankan cinnamon leaf diseases or management techniques.
3	An Introduction to Sri Lanka and Its Cinnamon Industry	Economic impact, industry overview	Not applicable	Provides historical context, cinnamon leaf and economic significance of cinnamon in Sri Lanka	Does not focus on disease management or prediction using ML.
4	Chemical and biological studies of value-added cinnamon products	Cinnamon leaf disease Disease management strategies	Not applicable	Focuses on disease management strategies for cinnamon diseases.	Limited to traditional methods. Lacks advanced machine learning techniques for disease prediction.
5	A Review of Identification and Management Pests and Diseases of Cinnamon (Cinnamomum zeylanicum Blume)	Pests and diseases of cinnamon Environmental factors	Not applicable	Detailed description of pests and diseases in cinnamon	Lacks the application of machine learning for disease prediction.
6	Identification and management of pests and diseases of cinnamon.	Pests and disease identification Cinnamon leaf diseases	Not applicable	Focuses on disease management	Doesn't suggest comprehensive management strategies
7	AgroX: Uplift Ceylon Cinnamon Industry	Technological interventions, Economic development,	Not applicable	Highlights technology's role in advancing the cinnamon industry	No specific mention of disease prediction technologies.
8	Exploring Deep Learning Models for Cinnamon Plant Disease Characterization	Disease types, Image features	Convolutional Neural Networks (CNN)	Achieves high accuracy in identifying specific cinnamon diseases through image analysis	Focuses on specific diseases. Not generalize to all cinnamon diseases and cinnamon leaf diseases.
9	Classification of Cassia Cinnamon and	Visual and chemical	Deep Learning	High accuracy in distinguishing	Focuses on classification, not disease prediction.

	Ceylon Cinnamon using Deep Learning	features of cinnamon	CNN	Cassia and Ceylon cinnamon	
10	Modeling CNN for Detection of Plant Leaf Spot Diseases.	Leaf spot diseases, Image classification	Convolutional Neural Network (CNN)	Classification accuracy of 90.6% for plant leaf diseases.	The model focuses only on leaf spot disease.
11	Recognizing matured cinnamon tree using image processing techniques	Maturity level of cinnamon trees Image processing	Support Vector Machine (SVM)	Identification of maturity level.	Accuracy might decrease in varying environmental conditions or with different cinnamon tree varieties
12	Plant Leaf Disease Detection and Classification Based on CNN with LVQ.	Plant leaf diseases.	CNN with Learning Vector Quantization (LVQ)	Accurate identification of leaf diseases.	LVQ algorithm may not perform well with large and diverse datasets of plant leaf diseases.
13	CinnaSense: Enhancing Cinnamon Quality and Health with Image Processing	Image-based features for cinnamon quality	Image processing	Focuses on improving cinnamon quality through advanced image processing techniques	Limited application to disease prediction and broader health assessments
14	Automatic Recognition of Plant Leaf Diseases Using Deep Learning (Multilayer CNN) and Image Processing	Leaf images Disease features	Multilayer CNN	Accurate leaf disease identification	Limited testing in real-world agricultural settings Results may not fully generalize to outdoor conditions.
15	Differentiating True and False Cinnamon: Exploring Multiple Approaches for Discrimination	Chemical and structural properties of cinnamon	Machine Learning (varied approaches)	Effectively differentiates true vs. false cinnamon	Does not address cinnamon leaf disease detection or prediction.
16	Machine Learning-Based Nutrient Deficiency Detection in Crops	Nutrient deficiencies, Leaf images	Convolutional Neural Network (CNN)	High accuracy in detecting nutrient deficiencies and recommending fertilizers.	System might not account for variations in nutrient deficiency symptoms across different crop types
17	Machine Learning Approach for New Crop Disease Predict and Alert System	Various crop features for disease prediction	Machine Learning, CNN, Random Forest	Efficient in predicting multiple crop diseases	May be too complex for practical deployment in regions with limited technological infrastructure, such as smallholder cinnamon farms.
18	Deep Learning for Agricultural Plant Disease Detection	Leaf images, Disease symptoms	Deep Learning, CNN	High accuracy in disease detection in agriculture	Model's performance may vary across different crop types.
19	Tracing Adulteration in Cumin, Cinnamon, and Coffee using	Chemical composition and adulteration detection	Transfer Learning	High accuracy in detecting adulteration in cinnamon,	Focuses on quality control rather than disease detection

	Transfer Learning			cumin, and coffee	
20	Detection of plant leaf diseases using deep convolutional neural network models	Leaf texture, color, shape	Deep Convolutional Neural Network	High accuracy in disease detection	Limited dataset size
21	Integrating Remote Sensing and Deep Learning for Precision Agriculture in Cinnamon Farming	Remote sensing data and cinnamon farming characteristics	Deep learning	Effective integration of remote sensing data for precision agriculture in cinnamon farming	Lack of real-world validation for disease-specific applications
22	Expert Prediction System for Spice Plants Grown in Sri Lanka	Environmental parameters Historical data	Various ML algorithms	Improves early detection of disease outbreaks in cinnamon crops, allowing for timely intervention	Relies on comprehensive data input

IV. CONCLUSION

This systematic review has explored the application of machine learning (ML) techniques in the prediction and management of cinnamon leaf diseases, with a focus on advancements in image processing and classification algorithms like Convolutional Neural Networks (CNNs), Support Vector Machines (SVMs), and Random Forests. While significant progress has been made in applying ML for disease detection across various crops, research specifically targeting the most common cinnamon leaf diseases in Sri Lanka remains limited. To date, no research has successfully identified a machine-learning model capable of detecting cinnamon leaf prevalent diseases comprehensively.

This review, therefore, aimed to assess and recommend the most suitable ML techniques that could be adapted for cinnamon disease detection. CNNs, with their ability to analyze image data and detect intricate patterns, have emerged as the most promising approach for this purpose. However, there are challenges to overcome, such as the availability of comprehensive datasets that accurately reflect real-world conditions and the development of models that are scalable and robust across different disease types and environmental variations. The integration of ML offers a transformative approach to disease detection and management, providing a more efficient and precise alternative to traditional

methods. Future research should focus on bridging the current gaps by developing more adaptable and scalable ML models, leveraging real-world field data, and addressing environmental variability. These advancements are crucial to enhancing disease management strategies and ensuring the sustainability and productivity of the cinnamon industry in Sri Lanka and beyond.

REFERENCES

- Chandima, T. D. K. D. and Kartheeswaran, T. (2016) 'Recognizing matured cinnamon tree using image processing techniques', in *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*. IEEE, pp. 1–5. doi: 10.1109/ICCIC.2016.7919665.
- Ekanayaka, E. M. H. I. and Kumara, P. P. N. V. (2022) 'Machine Learning Approach for New Crop Disease Predict and Alert System: A Review', in *2022 International Conference on Emerging Techniques in Computational Intelligence (ICETCI)*. IEEE, pp. 45–49. doi: 10.1109/ICETCI55171.2022.9921364.
- Fatima, N. *et al.* (2021) 'Tracing Adulteration in Cumin, Cinnamon, and Coffee using Transfer Learning', in *2021 International Conference on Data Analytics for Business and Industry (ICDABI)*. IEEE, pp. 562–566. doi: 10.1109/ICDABI53623.2021.9655951.
- Feltes, G. *et al.* (2023) 'Differentiating True and False Cinnamon: Exploring Multiple Approaches for Discrimination', *Micromachines*, 14(10), p. 1819. doi: 10.3390/mi14101819.

- Giraddi, S., Desai, S. and Deshpande, A. (2020) 'Deep Learning for Agricultural Plant Disease Detection', in, pp. 864–871. doi: 10.1007/978-981-15-1420-3_93.
- Gunasekara, R. *et al.* (2021) 'Expert Prediction System for Spice Plants Grown in Sri Lanka: An Incentive for Planters', in *2021 IEEE 16th International Conference on Industrial and Information Systems (ICIIS)*. IEEE, pp. 86–91. doi: 10.1109/ICIIS53135.2021.9660721.
- Jayasena, C. *et al.* (2023) 'CinnaSense: Enhancing Cinnamon Quality and Health with Image Processing', in *2023 5th International Conference on Advancements in Computing (ICAC)*. IEEE, pp. 173–178. doi: 10.1109/ICAC60630.2023.10417405.
- Jayasinghe, G. G. *et al.* (2020) 'Pests and Diseases of Cinnamon (*Cinnamomum zeylanicum* Blume)', in *Cinnamon*. Cham: Springer International Publishing, pp. 201–232. doi: 10.1007/978-3-030-54426-3_8.
- Lakshan S *et al.* (2023) 'Integrating Remote Sensing and Deep Learning for Precision Agriculture in Cinnamon Farming', *International Research Journal of Innovations in Engineering and Technology*, 07(08), pp. 65–71. doi: 10.47001/irjiet/2023.708009.
- Pathirana, R. and Senaratne, R. (2020) 'An Introduction to Sri Lanka and Its Cinnamon Industry', in *Cinnamon*. Cham: Springer International Publishing, pp. 1–38. doi: 10.1007/978-3-030-54426-3_1.
- Pratondo, A. *et al.* (2024) 'Classification of Cassia Cinnamon and Ceylon Cinnamon using Deep Learning', in *2024 10th International Conference on Wireless and Telematics (ICWT)*. IEEE, pp. 1–5. doi: 10.1109/ICWT62080.2024.10674700.
- Prof. Koliya Pulasinghe, Dr. Dharshana Kasthurirathna and S.A.A. Ravishan (2023) 'AgroX: Uplift Ceylon Cinnamon Industry', 7(11), pp. 27–34.
- Rajapakse, R. H. S. and Kumara, K. L. W. (2007) *A Review of Identification and Management of Pests and Diseases of Cinnamon (Cinnamomum zeylanicum Blume)*, *Tropical Agricultural Research & Extension*.
- Sardogan, M., Tuncer, A. and Ozen, Y. (2018) *Plant Leaf Disease Detection and Classification Based on CNN with LVQ Algorithm*, *2018 3rd International Conference on Computer Science and Engineering (UBMK)*.
- Shandilya, G. *et al.* (2024) 'Exploring Deep Learning Models for Cinnamon Plant Disease Characterization: Discriminating Rough Bark from Stripe Canker', in *2024 Asia Pacific Conference on Innovation in Technology (APCIT)*. IEEE, pp. 1–6. doi: 10.1109/APCIT62007.2024.10673493.
- Singla, P., Kalavakonda, V. and Senthil, R. (2024) 'Detection of plant leaf diseases using deep convolutional neural network models', *Multimedia Tools and Applications*, 83(24), pp. 64533–64549. doi: 10.1007/s11042-023-18099-3.
- Sunitha, G. *et al.* (2022) 'Modeling Convolutional Neural Network for Detection of Plant Leaf Spot Diseases', in *2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC)*. IEEE, pp. 1187–1192. doi: 10.1109/ICESC54411.2022.9885593.
- Suriyagoda, L. *et al.* (2021) "'Ceylon cinnamon": Much more than just a spice', *Plants People Planet*. Blackwell Publishing Ltd, pp. 319–336. doi: 10.1002/ppp3.10192.
- T*, A. *et al.* (2020) 'Machine Learning Based Nutrient Deficiency Detection in Crops', *International Journal of Recent Technology and Engineering (IJRTE)*, 8(6), pp. 5330–5333. doi: 10.35940/ijrte.F9322.038620.
- Tusher, A. N. *et al.* (2022) 'Automatic Recognition of Plant Leaf Diseases Using Deep Learning (Multilayer CNN) and Image Processing', in, pp. 130–142. doi: 10.1007/978-3-031-12413-6_11.
- Wickramasinghe, H. C. *et al.* (2018) 'CHEMICAL AND BIOLOGICAL STUDIES OF VALUE-ADDED CINNAMON PRODUCTS IN THE SRI LANKAN MARKET', *International Journal of Pharmaceutical Sciences and Research*, 9(11), p. 4674. doi: 10.13140/RG.2.2.26098.48324.
- Wijesekera, R. O. B. and Chichester, C. O. (1978) 'The chemistry and technology of cinnamon', *CRC Critical Reviews in Food Science and Nutrition*, 10(1), pp. 1–30. doi: 10.1080/10408397809527243.

Recognition of Sri Lankan Traffic Signs using Machine Learning Techniques

A. Priscilah Nivetha¹ and M.S. Suhail Razeeth²

¹Department of Information Technology, Trincomalee Campus, Eastern University, Sri Lanka

²Department of Information and Communication Technology, South Eastern University, Sri Lanka

¹nivethaa@esn.ac.lk, ²razeethsuhail@seu.ac.lk

Abstract

The recognition of traffic signs is a crucial component of driver assistance systems that have been extensively researched worldwide. However, it remains a challenging issue due to the increasing number of vehicles, road signs, and the lack of awareness among drivers and other road users. A Traffic Sign Recognition (TSR) system is an advanced autonomous technology designed to assist drivers by accurately identifying and interpreting traffic signs. This system plays a crucial role in enhancing driver awareness and ensuring appropriate responses to various traffic conditions. The precise recognition of traffic signs is essential for maintaining road safety and improving the overall driving experience. This study focuses on the recognition of Sri Lankan traffic signs and examines the combination of classifiers with a specific feature extractor. A dataset of 300 images of road signs was utilized for this study by capturing the images. The Scale-Invariant Feature Transform (SIFT) was used as a feature descriptor in this process. The classifiers employed were Support Vector Machine (SVM) and *k*-Nearest Neighbor (*k*-NN). Different combinations of SVM and *k*-NN were applied to the dataset, and the study achieved 100% accuracy with various combinations of *k*-NN. The study found that the combination of SIFT and SVM is the most effective method for the proposed recognition of traffic signs.

Keywords: Sri Lankan Traffic signs, Traffic signs recognition, SIFT, SVM, *k*-NN, machine learning

I. INTRODUCTION

With the rise in traffic density and the push towards autonomous vehicles, accurately identifying and responding to traffic signs is essential for ensuring road safety and compliance with traffic laws. Human drivers can easily miss or misinterpret signs due to distractions or poor visibility, leading to accidents and traffic

violations. Traffic sign recognition systems address these issues by providing real-time, reliable detection and interpretation of road signs, aiding drivers in making safer decisions and enabling autonomous vehicles to navigate more effectively. This technology is vital for reducing accidents, enhancing driver assistance systems, and paving the way for fully autonomous driving solutions.

Traffic sign recognition is a crucial technology in modern transportation systems, playing a vital role in enhancing road safety and enabling autonomous driving. By using advanced image processing and machine learning algorithms, traffic sign recognition systems can accurately identify and interpret various road signs. This technology helps drivers make informed decisions in real-time and assists autonomous vehicles in navigating complex road environments. As the development of intelligent transportation systems continues, traffic sign recognition stands out as a key component in reducing accidents and improving the overall efficiency of road networks. Traffic sign recognition using machine learning involves training algorithms to automatically detect and classify traffic signs from images. Machine learning (ML) is an umbrella term that refers to a broad range of algorithms that perform intelligent predictions based on a dataset (Nichols, et al., 2019). The traffic sign datasets are often large, perhaps consisting of millions of unique data points.

Machine learning models are powerful tools that enable systems to learn from data and make predictions or decisions without being explicitly programmed. These models analyze patterns and relationships within large datasets, allowing them to identify trends, classify information, and make informed predictions. In traffic sign recognition machine learning models are used to process and interpret visual data, identifying and categorizing various traffic signs with high accuracy.

In this study, systematic experiments were conducted to evaluate the performance of classification using SIFT feature representation combined with k-Nearest Neighbor (k-NN) and Support Vector Machines (SVM). The evaluation was performed on a newly created dataset of Sri Lankan traffic signs, focusing on six randomly selected sign types.

The rest of the paper is organized as follows. The literature is reviewed in Section II. The experimental methods are presented in Section III. The results of the experiment and the discussion are presented in Section IV and lastly, the conclusion and future works of the paper are presented in Section V.

II. LITERATURE REVIEW

Table 01: Comparability Study with Different Datasets and Models

Study	Model	Dataset	Accuracy
SVM Based Method			
(Mahesh, 2018)	SIFT and SVM	No Details are provided	90%
(Møgelmoose, et al., 2012)	SVM, KNN, Random Forest, and Naïve Bayes	Real-time Indonesian Traffic Signs	86%
(Ali, et al., 2023)	SVM and HOG	GTSDDB, GTSRB, Linköping University, Real-Time Taiwanese Traffic Signs	94.9%
(Rahmad, et al., 2018)	SVM and KNN	Real-time Indonesian Traffic Signs	82.01%
Other Methods			
(Ardianto, et al., 2017)	CNN	GTSDDB, GTSRB	96%
(Sugiharto & Harjoko, 2016)	CNN	Real-Time Russian Traffic Signs	87%
(Chakraborty & Deb, 2015)	SVM, CNN	Own collection of 12 signs of Sri Lankan traffic signs	SVM – 98.33 % CNN – 96.40%
(Roxas, et al., 2018)	CNN	GTSRB	95%
(Wang, 2018)	SVM and SIFT	Sri Lankan Traffic sign Dataset	87%
(Filatov, et al., 2017)	SVM, KNN, MPC, DT, AdaBoost	Sri Lankan Traffic sign Dataset	90%
(Kiridana, et al., 2022)	CNN, SVM	Google Street View	98.5%
Our Method			
*	SVM, KNN, and SIFT	300 images of Sri Lankan traffic signs	100%

(Adam & Ioannidis, 2014) presented a comprehensive methodology for road sign detection and recognition, addressing various challenges. It emphasized the effectiveness of using HOG descriptors to represent Regions of Interest (ROIs) and employed HIS color space for thresholding in the detection stage. For recognition, a Histogram of Oriented Gradients (HOG) is used, with Support Vector Machines (SVMs) handling classification. The system also included a successful step for separating

In this section a summarization of some studies related to road sign recognition is presented.

There are varieties of models and algorithms available for road sign detection. (Nikam & Dhaigude, 2017), proposed a novel system for automatic road sign detection and recognition. The system segments input images in YCbCr color space and detects road signs using shape filtering. Recognition of the road sign symbols is achieved through Principal Component Analysis (PCA). The study also discussed many roads sign detection and recognition techniques. MESR, HSV, SVM, OCR, HIS, and HOG. The system is designed to be both efficient and robust in detecting and recognizing road sign symbols. It is obvious from the study that machine learning and deep learning techniques are used for road sign detection algorithms.

overlapping signs and demonstrated high competence, showing robustness to changes in illumination, scale, and partial occlusions. The study shows 94.34% of accuracy the road sign signal.

In (HU, et al., 2010) they proposed a Traffic Sign Recognition (TSR) method that effectively addresses challenges such as weather conditions, shooting angle, and distance variation. Experiments were conducted on training image sets classified by these factors. The method

utilizes the SIFT technique to extract sign features, forms a codebook using K-means clustering, and classifies the signs with an SVM based on feature distribution. The results demonstrate that our method is robust to variations in weather, distance, and shooting angle, achieving a high accuracy rate of 93% with a low computation time of 0.098 seconds per image.

III. METHODOLOGY

Here, we discuss the compositional parts, and the process engaged with building and fostering our model. The proposed work is based on the Bag of Features approach. The bag of features approach includes the following phases: feature extraction, codebook creation, histogram representation and learning and classification. The main idea is to generate histograms of images for the classification process. MATLAB R2021a and Windows 10 with 8GB RAM were used for all the implementations. Figure 01 illustrates the steps involved in the implementation process.

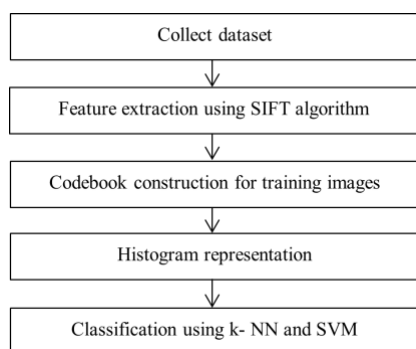


Figure 01: Methodology







Dataset

The dataset for this study comprises 300 images of Sri Lankan traffic signs: two informational signs and four warning signs. The dataset was captured from mobile with a good lighting condition. Initially, the road signs were cropped and normalized to a size of 200×200 pixels from the original image. The Images in each class were divided into two parts 70% for the training dataset and 30% for the testing dataset.



Figure 02: Class images of the dataset

Table 02: Road Sign Dataset

Sign	Train	Test
	35 Images	15 Images
	35 Images	15 Images
	35 Images	15 Images
	35 Images	15 Images
	35 Images	15 Images
	35 Images	15 Images

Feature extraction using the SIFT algorithm

SIFT was employed to extract features from the images. SIFT transforms data images into scale-invariant coordinates relative to local features. This process involves extracting features from the image, representing them as distinct patches, and then converting these patches into a collection of 128-dimensional vectors.

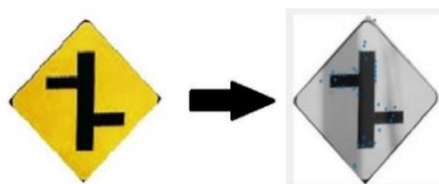


Figure 03: SIFT feature extraction Codebook construction for training images.

Codebook construction involves generating visual words or codewords through clustering techniques applied to the original feature space. The K-means algorithm was used for clustering on the training dataset. It is because of the simplicity, efficiency, scalability and speed of the algorithm. Generally, K-means clustering algorithms are straightforward to understand, works very fast, handle large amount of data, therefore this study is utilized k-means cluster as a one of the algorithms over SVM to train the dataset and check the validity. Besides, K-means clustering partitions a set of N features into k clusters, with each feature assigned to the nearest cluster based on the mean of the cluster's members. The centers of these clusters are then used to create codewords, which together form the codebook.

Histogram Representation

The images were represented by histograms using the constructed codebook. Separate histograms were generated for both the training and testing images.

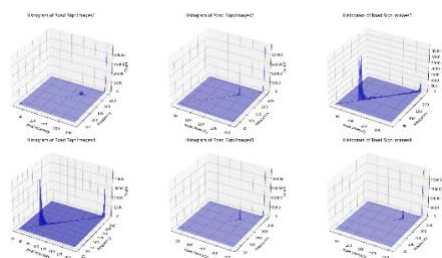


Figure 04: Histogram of road sign images

Classification using k- NN and SVM

The k-Nearest Neighbor (k-NN) approach was used for classification. It calculates pairwise Euclidean distances between key point representations of a test image and all labelled training images in the dataset. The Euclidean norm distance was employed to measure the distance between key points, with k values set to 1, 3, 5, 7, and 9.

Support Vector Machine (SVM) was used as an additional classifier. For multiclass classification, a linear SVM was trained using the one-versus-all (OVA) approach. The OVA rule separates each class from the others and assigns the test image to the class with the highest classifier response. The SVM^{light} package was utilized for the experiments. The accuracy rate for each classification was calculated, and this data was used to analyze the performance of the classifiers.

IV. RESULTS AND DISCUSSION

In this section, the outcomes of the execution of the methods referenced in the above-proposed works will be specified.

Table 03 presents the classification rates for different k values in the k-NN algorithm (k = 1, 3, 5, 7, and 9). All results shown in Table 01 were obtained without dimensionality reduction of the features extracted by the Bag of Features (BoF) method. This approach was chosen to preserve the full feature set and ensure that no potentially important information was lost during the dimensionality reduction process.

Table 03: Classification using different k values

K values	Accuracy
k=1	100.00%
k=3	97.78%
k=5	97.78%
k=7	96.67%
k=9	94.44%

Figure 05 illustrates the changes in classification rates when using the combination of SIFT and SVMs with parameter tuning. The graph depicts the impact of varying C values between 2^{-14} and 2^{10} on classification performance. The C value is the regularization parameter to avoid overfitting values. Normally high C value provides overfitting and low C value provides underfitting, the value between 2^{-14} and 2^{10} provide optimal performance to provide best classification performance without overfitting and underfitting.

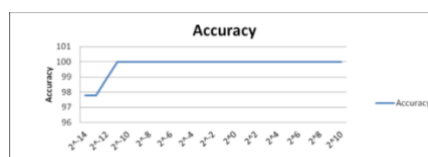


Figure 05: Graph of classification accuracy rates for SIFT + SVM with parameter tuning of C values ranging from 2^{-14} to 2^{10}

According to Table 03, as the number of nearest neighbors k in the k-NN algorithm increases, the accuracy rate decreases. Parameter k determines how many neighboring examples are considered when making a classification. Although considering more neighbors can generally improve classification accuracy, an increase in k may lead to a decline in accuracy if it introduces noise or less

relevant information into the decision-making process.

In Figure 05, which illustrates the combination of SIFT and SVM, the accuracy remains stable until a notable increase is observed when the parameter C is set from 2^{-12} to 2^{-10} . The parameter C in SVM controls the cost of classification errors. A larger C value focuses on minimizing classification errors by finding a more precise margin, while a smaller C emphasizes maximizing the margin, potentially leading to some classification errors. As C increases, the model becomes better at classifying points accurately, leading to higher accuracy rates. At certain points, the accuracy reaches 100% and remains consistent. Therefore, combining SIFT with an SVM classifier, particularly with optimal C values, yields the highest accuracy in our proposed method.

Table 01 above shows the comparative studies of SVM, CNN with different datasets. Different Dataset with different SVM, and other techniques shows different accuracies for relevant countries. For the Sri Lankan dataset k-NN with SIFT feature extraction technique is better suite, and it provides highest accuracy. Besides, In the available road sign dataset of Sri Lanka, this study got more accuracy with k-NN and SIFT.

On the other hand, if the other studies are changing parameters of the model and consider feature engineering techniques it also provides higher accuracy without any suspects.

V. CONCLUSION

Traffic sign recognition is a critical aspect of driver assistance systems and has been extensively studied globally. Intelligent autonomous systems for traffic sign recognition are essential for helping drivers understand and respond accurately to road signs. This study focuses on the recognition of Sri Lankan traffic signs, utilizing machine learning techniques to enhance the system's effectiveness. Specifically, Scale-Invariant Feature Transform (SIFT) is used as the feature descriptor, with Support Vector Machine (SVM) and k-Nearest Neighbor (k-NN) classifiers employed for the recognition process. The results indicated that combining SIFT with SVM is the most effective method for traffic sign recognition, offering significant improvements in classification accuracy and efficiency.

In the future, this study can be enhanced by extending the recognition to include a broader range of Sri Lankan traffic signs and enabling real-time detection, as the current approach is limited to still images. Additionally, expanding the dataset to include more diverse examples would improve the model's robustness. Further research could also involve comparing the proposed model with other existing models to evaluate its relative performance and effectiveness.

REFERENCES

- Nichols, J. A., Chan, H. W. H. & Baker, M. A. B., 2019. Machine learning: applications of artificial intelligence to imaging and diagnosis. *Biophysical Reviews*, 11(1), p. 111–118.
- Mahesh, B., 2018. Machine Learning Algorithms - A Review. *International Journal of Science and Research (IJSR)*, pp. 381-386.
- Møgelmoose, A., Trivedi, M. M. & Moeslund, T. B., 2012. *Traffic Sign Detection and Analysis: Recent Studies and Emerging Trends*. Anchorage, Alaska, s.n.
- Ali, G., Sadıkoğlu, E. & Abdelhak, H., 2023. Design a Hybrid Approach for the Classification and Recognition of Traffic Signs Using Machine Learning. *Wasit Journal of Computer and Mathematics Science*, 2(2), pp. 18-25.
- Rahmad, C., Rahmah, I. F., Asmara, R. A. & Adhisuwignjo, S., 2018. *Indonesian Traffic Sign Detection and Recognition Using Color and Texture Feature Extraction and SVM Classifier*. Yogyakarta, Indonesia, IEEE.
- Ardianto, S., Chen, C.-J. & Hang, H.-M., 2017. *Real-Time Traffic Sign Recognition using Color Segmentation and SVM*. Poznan, Poland, IEEE.
- Sugiharto, A. & Harjoko, A., 2016. *Traffic sign detection based on HOG and PHOG using binary SVM and k-NN*. Semarang, Indonesia, IEEE, pp. 317-321.
- Chakraborty, S. & Deb, K., 2015. *Bangladeshi Road Sign Detection Based on YCbCr color model and DtBs Vector*. Rajshahi, Bangladesh, IEEE, pp. 158-161.
- Roxas, E. A. et al., 2018. *2018 IEEE International Conference on Applied System Invention (ICASI) Convolutional Neural Network*. Chiba, Japan, IEEE, pp. 120-123.
- Wang, C., 2018. *Research and Application of Traffic Sign Detection and Recognition Based on Deep Learning*. Changsha, China, IEEE, pp. 150-152.

- Filatov, D. M., Ignatiev, K. V. & Serykh, E. V., 2017. *Neural Network System of Traffic Signs Recognition*. St. Petersburg, Russia, IEEE, pp. 422-423.
- Kiridana, Y. M. W. H. M. R. P. J. R. B. et al., 2022. *Intelligent Detection of Sri Lankan Road Signs by Using Google Street View Images*, s.l.: s.n.
- Nikam, P. A. & Dhaigude, N. B., 2017. Road Sign Symbol Detection and Recognition. *International Journal of Electrical and Electronics Engineers*, 09(01), pp. 1028-1032.
- Adam, A. & Ioannidis, C., 2014. Automatic Road-sign Detection and Classification Based on Support Vector Machines and HOG Descriptors. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, II(5).
- HU, X., ZHU, X., LI, D. & LI, H., 2010. *Traffic Sign Recognition Using Scale Invariant Feature Transform and SVM*. Florida, s.n.