ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr893

Pathogen Identification using Linear Regression and Convolutional Neural Networks

Harsith Adhithya Senthil Kumaran[1]; Aakaash Suman Suresh[2]; Prakash. J[3]

[1;2]PSG Institute of Advanced Studies
[3]PSG College of Technology

Publication Date: 2025/04/28

Abstract: With the increase in awareness regarding conservation of forests, we must be wary to preserve them sustainably from potential pathogens. Statistics tells us that the number of trees that we lose every year due to pathogen attacks is huge and thus requires a machine learning model to identify the presence of pathogens to significantly reduce the number of deaths per year. TIn this paper we have done a cumulative study about the efficiency of two different models namely Linear Regression and CNN(Convolutional Neural Networks) and have achieved the following accuracies with respect to the actual data. For Linear Regression we have achieved an accuracy of 65.71% and an accuracy of 80.85% for CNN. Further analysis of various metrics like RMS(Root Mean Square) value, MAE(Mean Absolute Error) and MSE(Mean Squared Value) is done for both the models.

Keywords: *Machine Learning, Linear Regression, Convolutional Neural Networks, Deep Learning, Pathogen.*

How to Cite: Harsith Adhithya Senthil Kumaran; Aakaash Suman Suresh; Prakash. J (2025). Pathogen Identification using Linear Regression and Convolutional Neural Networks. *International Journal of Innovative Science and Research Technology*, 10(4), 1593-1598. https://doi.org/10.38124/ijisrt/25apr893

I. INTRODUCTION

Pathogen refers to a microscopic organism that causes harm to it's host (the organism it resides on) in the form of diseases or illness. They might be bacteria, fungi, nematodes or parasitic plants. They cause harm by secreting toxins, invasion damaging the host cells or secretion of growth regulators. To achieve this, we are leveraging machine learning—a technology renowned for its ability to handle complex tasks and process large volumes of data with impressive speed and precision. Machine learning is especially effective for this purpose because it can quickly shift through vast datasets and identify patterns that might indicate the presence of harmful pathogens. Beyond just pathogen detection, machine learning is a versatile tool with applications in several areas. It aids in prediction, which forecasts future outcomes based on historical data; and image recognition, which helps in identifying and classifying visual information[1-6].

In summary, machine learning equips us with the advanced capabilities needed to monitor and protect plant ecosystems effectively. By using these tools, we aim to enhance the health and preservation of both cultivated and wild plant life.

The article is organized as follows Section 2 discusses the Related Work which explains various research related to this and future enhancement for the same. As in Section 3 the paper discusses the various methodologies used to incorporate the following models and their implementation. Section 3 talks about Linear Regression and 3.1 explains the step by step algorithm for the implementation of the model Linear Regression. Section 3.2 talks about the step by step implementation of CNN(Convolutional Neural Networks) in an algorithmic step by step implementation. Section 4 talks about Result Analysis which compares two models with respect to various metrics to evaluate the efficiency of both these models. Section 4.1 talks about Mean Squared Error and 4.2 talks about Root Mean Squared Value Analysis. A table is made to compare both these models with various metrics and a visual representation in the form of a bar graph. Section 5 gives the conclusion of the article and gives the overall overview.

II. RELATED WORK

Emerging infectious diseases (EIDs) represent a growing threat to both plant conservation and public health. By extending the concept of EIDs, typically used in medicine and veterinary science, to the world of botany, this review highlights several new and concerning plant diseases. These include issues affecting both cultivated crops and wild plants, some of which have significant conservation implications. The main drivers behind most of these plant EIDs are human activities that introduce new parasites, though extreme weather events also play a role. While much is known about EIDs in crops, there is less information about wild plant EIDs, indicating that their impact on conservation may be underestimated. The review concludes with suggestions for

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr893

improving surveillance and control strategies for these plant diseases (Pamela K et.al, 2004))[7].

Ilaria Buja et al., in an article reviews the latest non-invasive techniques for detecting plant diseases. It segregates these methods two two different types such as spectroscopic imaging techniques and as techniques based on profiling volatile organic compounds. The first method namely includes methods/techniques like fluorescence spectroscopy visible IR , fluroscence imaging and hyper spectral imaging. This helps us to observe plant health without contact and to keep track of plant health and longitivtiy [8].

Jagadeesh D et al., in a study explores how image processing techniques can be applied to identify and classify symptoms of fungal diseases in various crops. Computers are increasingly used for automating processes and supporting decision-making in agriculture. Since many early symptoms of plant diseases are microscopic and difficult to detect with the naked eye, image processing technologies are crucial for accurate diagnosis and management of plant health [9].

The Concept of Machine learning is usually divided in supervised and unsupervised learning based on Data Availability. For smaller datasets Supervised, the first category seems viable. The data should be less ambiguous with clearly labeled examples, while unsupervised is for larger datasets. For these kind of datasets with huge data deep learning is preferred at most cases. This paper also talks about the applications and limitations of neural networks giving an overview about various machine learning models and how it affects various websites from algorithms to client relationship management [10]

III. METHODOLOGY

In this article we have used two popular algorithms with respect to Machine Learning so as to have a contrast between fundamental statistical model and a model specialized in image recognition and processing. Linear Regression is a model that predicts outcomes based on two or more dependent variables by means like plotting intercepts and estimating values whereas CNN(Convolutional Neural Networks) can handle high-dimensional data through hierarchical feature extraction and provide higher accuracy and precision.

- *Linear Regression:*

This machine learning model uses existing datasets to predict new values or intermediate missing values. This uses the intercept method by drawing a plot between the values (A line close to the dataset) and this model will be able to predict missing values, intermediate or new values by using the intercept formula y = mx. This has a wide range of applications with respect to a linear nature or linear growth like amount of bacteria in an environment, land prices across the years or the average speed of vehicles.

Table 1 Linear Regression

Algorithm-1: Linear Regression
1. Importing Required Modules
2. Inputting the Datasets
2.1.Loading Data
2.2.Inspecting Data
2.3.Preprocessing.
3. Creating the Linear Regression Model
3.1.Splitting the Data.
3.2.Initializing the Model.
3.3.Training the Model.
3.4.Evaluating the Model.
3.5.Plot the model.
4. Computing the Results
4.1.Making Predictions.
4.2.Evaluating Performance.
4.3.Analyzing Residuals.
5. Refining the Model
5.1.Feature Engineering.
5.2.Hyperparameter Tuning.
5.3.Algorithm Enhancement.
6. Visualizing the Data
6.1.Creating Visualizations
6.2.Analyzing Trends.
6.3.Communicating Findings.

Fig.1 outlines the workflow of a linear regression model. It starts with data collection, followed by converting the input into an appropriate format. The model is then built, and parameters are estimated using graphs. The linear model is examined to ensure it aligns with the existing dataset. If the model performs well, it is executed; otherwise, adjustments are made to improve its accuracy.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr893

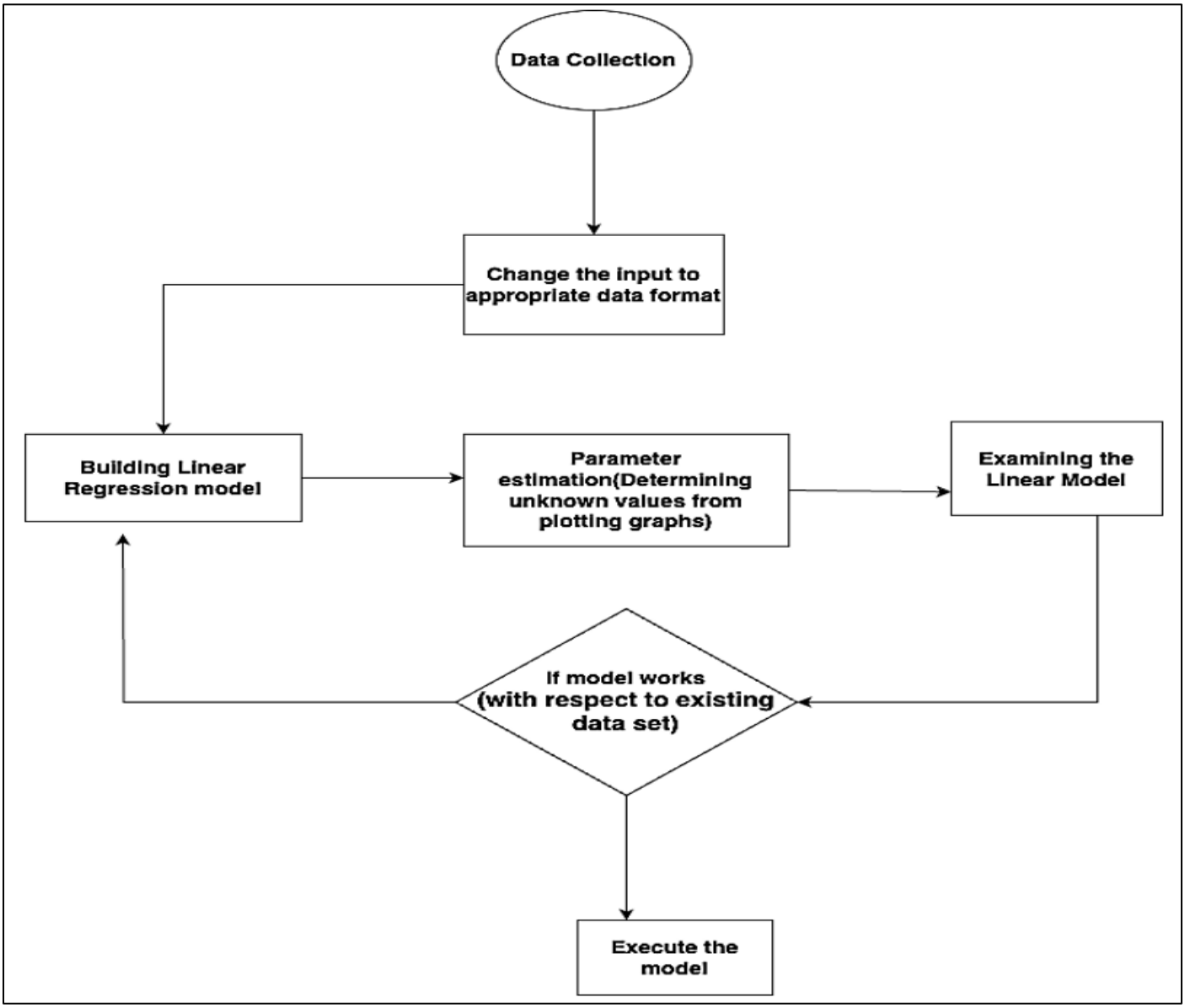

Fig 1 Working of Linear Regression Model.

➢ *Convolutional Neural Networks:*

This is a type of Machine learning model (Deep Learning model) used for larger datasets used specifically in image processing as it is highly efficient in image processing and and interpretation. In this model we would have imported the image and converted the image to 64x64 pixels for better comparison and image decoding.

Table 2 Convolutional Neural Networks

Algorithm-2: Convolutional Neural Networks
1.Import Appropriate Modules
2.Load the Pathogen Dataset
3.Create Visualizations
3.1 Plot Distributions:
3.2 Explore Relationships:
4.Extract Target Values into a DataFrame
4.1 Select Target Column:
4.2 Create a New DataFrame:
5.Convert Categorical Data to Numerical Data
5.1 Encode Categorical Variables:
6.Use a Single Column for the Target Value
6.1 Identify the Target Column:
6.2 Configure the Model:

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr893

7.Compute Correlation Matrix
7.1 Calculate Correlations:
8.Reduce Attributes in DataFrame X
8.1 Select Important Features:
9.Apply Convolutional Neural Network (CNN) for the Pathogen data.
9.1 Build the CNN Model:
9.2 Train the Model:
10.Compare Model Fitting and Predictions
10.1 Assess Accuracy:
10.2 Analyze Performance Metrics:
11.Visualize and identify, true positives, true negatives and so on.
11.1 Generate the Matrix:
12. Print Classification Report

The Fig.2 illustrates an image classification process that begins with collecting and preprocessing image data by converting it to JPEG format and resizing it to 64×64 pixels. The data is then split into training and testing sets, and features are extracted using a pre-trained CNN. These features are classified using an SVM, and the CNN parameters are tuned to improve performance. If the target accuracy is achieved, the final model is executed and validated against correct images.

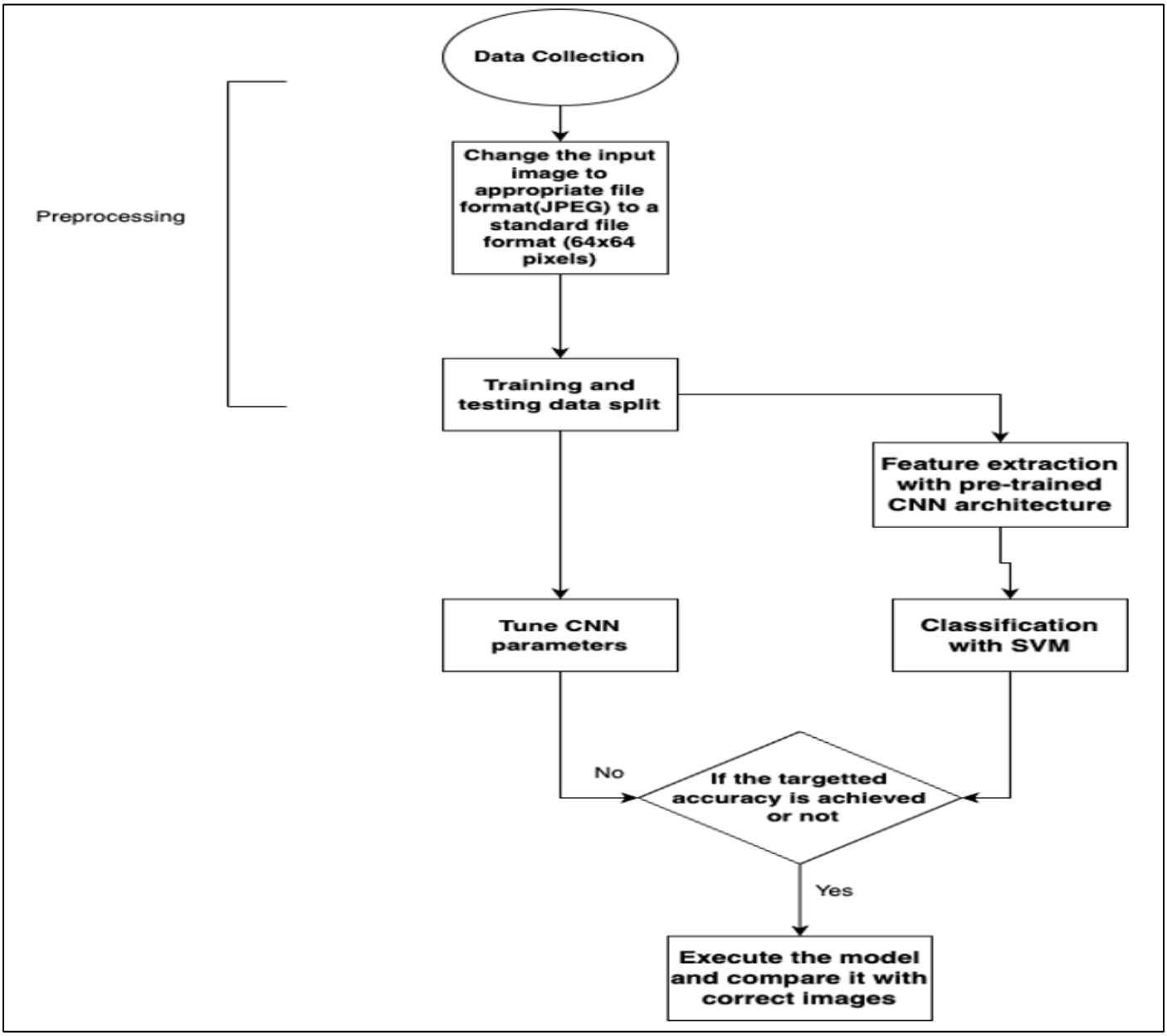

Fig 2 Working of CNN Model.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr893

IV. PERFORMANCE METRICS

➢ *Mean Absolute Error:*

In the Brach of statistics, Mean Absolute Error is the count/measure of errors between concurrent/ consecutive values that is experiencing the same phenomenon. Mathematically it is defined as the sum of absolute errors divided by the whole size. A good Mean Absolute Error value is usually under 10%. Hence it finds the mean value or average magnitude of existing errors in a dataset.

$$MAE(y,y) = \frac{\sum_{i=0}^{N-1}|y_i - y_i|}{N}$$

➢ *Mean Squared Value:*

Mean squared value (MSE) refers to arithmetic mean of the set of square values in the dataset of a random variable. In other words it is the arithmetic mean of squares of deviations. With respect to the concept of machine learning it is defined as a loss function holding the value of the average of the squared differences.

$$MSE(y,y) = \frac{\sum_{i=0}^{N-1}(y_i - y_i)^2}{N}$$

➢ *Root Mean Square:*

This is a metric used in machine learning models to find out the quality of the existing test dataset and the predictions made by the model itself as both tend to be mutually dependent. This helps us to visualize how the values predicted by our model is close to the actual values, helping us estimate the accuracy of a model.

$$RMSE(y,y) = \sqrt{\frac{\sum_{i=0}^{N-1}(y_i - y_i)^2}{N}}$$

V. RESULT ANALYSIS

The comparative analysis between Linear Regression and Convolutional Neural Networks (CNN) demonstrates significant performance differences across key evaluation metrics. As illustrated in the Fig.3, CNN consistently outperforms Linear Regression in terms of precision and accuracy, recording values of 58.04% and 80.85%, respectively, compared to 51.33% and 65.71% achieved by Linear Regression. These results indicate that CNN is more effective at correctly identifying relevant patterns and making accurate predictions.

Furthermore, CNN exhibits lower error values in both Mean Absolute Error (MAE) and Mean Squared Error (MSE), with percentages of 58.3% and 33.99%, respectively. In contrast, Linear Regression yields higher error rates at 69.44% (MAE) and 48.22% (MSE), suggesting less reliable predictions. This demonstrates that CNN generalizes better to unseen data and produces more stable outputs. However, it is noteworthy that the Root Mean Squared Error (RMSE) for CNN is higher at 82.44% compared to 48.22% for Linear Regression. This deviation suggests that while CNN reduces the average prediction error, it may occasionally produce larger individual errors. Such behavior could be attributed to the model's complexity and sensitivity to outliers.

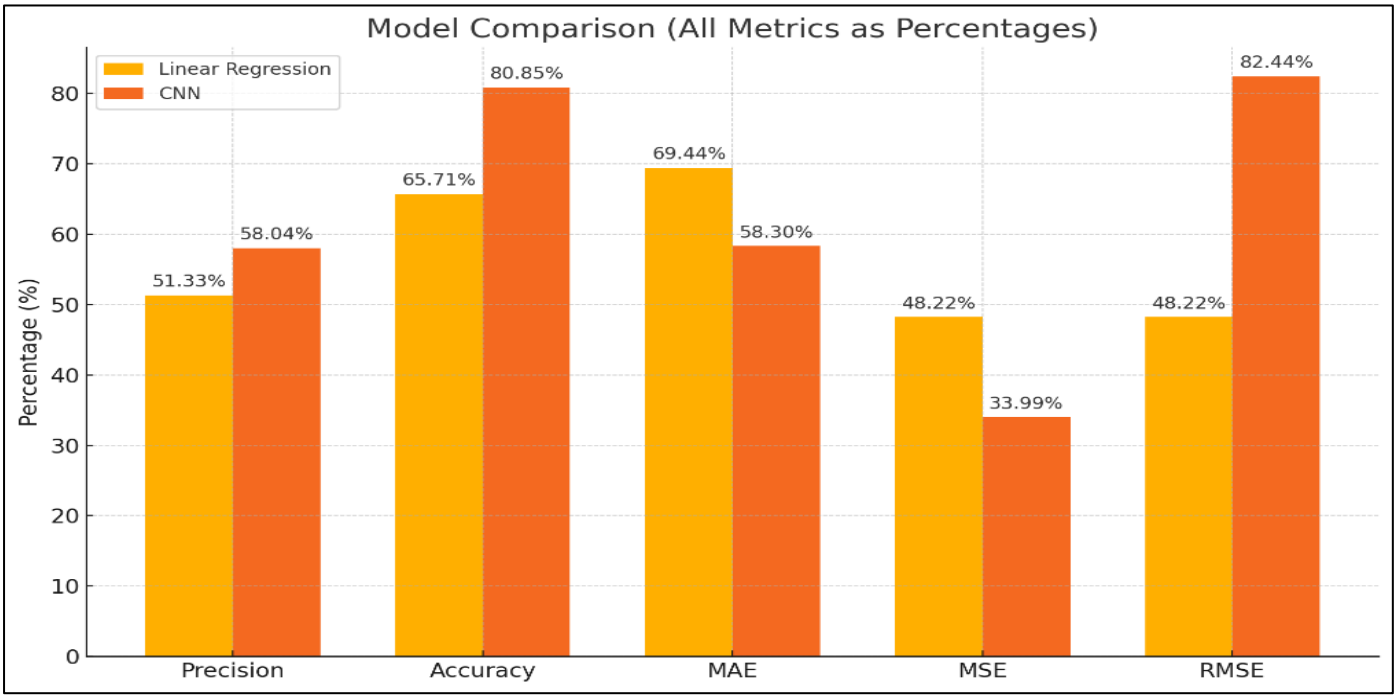

Fig 3 Performance of Linear Regression & CNN

Overall, the results highlight the superiority of CNN over Linear Regression in the given context, especially in classification accuracy and error minimization, making it a more robust model for the task under study.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25apr893

Table 3 Performance of Linear Regression & CNN

	Linear Regression	Convolutional Neural Networks
Precision	51.33%	58.04%
Accuracy	65.71%	80.85%
MAE	69.44%	58.3%
MSE	48.22%	33.99%
Root Mean Squared	48.22%	82.44%

VI. CONCLUSION

The comparison between Convolutional Neural Networks (CNNs) and Linear Regression underscores their distinct capabilities, advantages, and appropriate applications. Linear Regression offers simplicity and interpretability, making it suitable for problems with clear, linear relationships between variables. However, CNNs are far more effective in handling complex tasks, especially those involving image analysis, due to their ability to learn hierarchical features directly from raw data. In the context of plant pathogen detection, CNNs demonstrate superior performance. Their ability to autonomously extract relevant visual patterns enables them to operate effectively even with noisy inputs or variations across different plant species and environmental conditions. This robustness makes CNNs a dependable choice for real-world agricultural applications.

Future improvements could involve implementing more sophisticated CNN architectures such as ResNet or EfficientNet, which offer deeper networks and improved computational efficiency. Integrating CNNs with other machine learning approaches may also boost accuracy and generalizability. Expanding the dataset to cover a broader spectrum of plant diseases and growing conditions would further enhance the model's applicability.

REFERENCES

[1]. Dolatabadian, A., & Neik, T. X. (2023). Image-based crop disease detection using machine learning: A review. *Plant Pathology*, *72*(4), 587–604. https://doi.org/10.1111/ppa.14006

[2]. Zhang, S., Huang, W., Zhang, C., & He, Y. (2021). Plant diseases and pests detection based on deep learning: A review. Plant Methods, 17, Article 22. https://doi.org/10.1186/s13007-021-00722-9

[3]. Sivanandhini, P., & Prakash, J. (2020). Crop yield prediction analysis using feed forward and recurrent neural network. *International Journal of Innovative Science and Research Technology*, *5*(5), 1092–1096

[4]. Yadav, A. K. (2021). Image captioning using R-CNN & LSTM deep learning model. *image*, *5*, 8.

[5]. Prakash, J., Vinoth Kumar, B., & Shyam Ganesh, C. R. (2020). A comparative analysis of deep learning models to predict dermatological disorder. *J Xi'an Univ Archit Technol*, *12*(11), 11.

[6]. Sivanandhini, P., & Prakash, J. (2020). Comparative Analysis of Machine Learning Techniques for Crop Yield Prediction. *International Journal of Advanced Research in Computer and Communication Engineering*, *289*.

[7]. Anderson, P. K., Cunningham, A. A., Patel, N. G., Morales, F. J., Epstein, P. R., & Daszak, P. (2004). Emerging infectious diseases of plants: pathogen pollution, climate change and agrotechnology drivers. Trends in Ecology & Evolution, 19(10), 535–544. https://doi.org/10.1016/j.tree.2004.07.021

[8]. Buja, I., Sabella, E., Monteduro, A. G., Chiriacò, M. S., De Bellis, L., Luvisi, A., & Maruccio, G. (2021). Advances in plant disease detection and monitoring: From traditional assays to in-field diagnostics. Sensors, 21(6), 2129. https://doi.org/10.3390/s21062129

[9]. Pujari, J. D., Yakkundimath, R., & Byadgi, A. S. (2014). Identification and classification of fungal disease affected on agriculture/horticulture crops using image processing techniques. 2014 IEEE International Conference on Computational Intelligence and Computing Research, 1-4. https://doi.org/10.1109/iccic.2014.7238283

[10]. Mahesh, B. (2020). Machine learning algorithms - A review. International Journal of Science and Research (IJSR), 9(1), 381-386. https://doi.org/10.21275/art20203995