# MangoFruitDDS Statistics, Analysis, Prediction

**Mohd Ayan Khan & Yash Sharma**

*B.Tech Information Technology*

*Manipal University Jaipur*

**Jaipur, India**

[mohd.229302398@muj.manipal.edu](mailto:mohd.229302398@muj.manipal.edu) & [yash.229303179@muj.manipal.edu](mailto:yash.229303179@muj.manipal.edu)

## ABSTRACT

This project explores MangoFruitDDS, a comprehensive dataset and framework designed for the detection, analysis, and prediction of mango fruit diseases.1 Leveraging cutting-edge computer vision and machine learning models, the study aims to address critical challenges in agricultural productivity, such as accurate disease diagnosis, yield estimation, and quality assessment.1 The dataset includes high-resolution images of mangoes, both healthy and diseased, annotated with labels for four common diseases: Alternaria, Anthracnose, Black Mould Rot, and Stem-end Rot.1 Advanced deep learning techniques, including YOLOv5 and other object detection algorithms, are utilized to train models for identifying mango conditions with high accuracy.1 Data preprocessing, such as resizing, augmentation, and background removal, ensures robust model training and real-world adaptability.1 Applications extend to automated fruit counting, ripeness prediction, and robotic harvesting, demonstrating the potential to revolutionize agricultural practices.1 The results underline the effectiveness of MangoFruitDDS in improving detection precision, optimizing resource utilization, and enabling scalable deployment in orchard management systems.1 This study paves the way for innovative solutions in precision agriculture through data-driven methodologies.1

## I. INTRODUCTION

### Prevalence and Impact of Mango Cultivation

Agriculture forms the backbone of many economies, particularly in developing countries, where it is a primary source of income and sustenance for millions of people.1 Mangoes, often referred to as the "king of fruits," are one of the most widely cultivated and consumed fruits in the world.1 They hold immense economic value in tropical and subtropical regions, contributing significantly to international trade and local livelihoods.1 However, mango production is fraught with challenges, one of the most critical being the onset of diseases that compromise both fruit quality and yield.1

### Issues with Conventional Diagnosis

Mango fruit diseases, such as Alternaria, Anthracnose, Black Mould Rot, and Stem and Rot, are prevalent in many mango-growing regions, causing substantial post-harvest losses and impacting marketability.1 These diseases manifest in various forms, such as discoloration, spots, and rot, which can reduce the aesthetic appeal of the fruit and render it unsuitable for consumption or export.1 Traditional methods for diagnosing these diseases involve visual inspection by agricultural experts.1 While effective in some contexts, these methods are often slow, expensive, and impractical for large-scale farming.1 Moreover, their reliance on expert knowledge creates accessibility barriers, particularly for smallholder farmers in resource-constrained regions.1

### Emergence of Machine Learning in Agriculture

With advancements in data science and machine learning, there has been a growing interest in using artificial intelligence (AI) to address agricultural challenges.1 Image-based disease detection has emerged as a promising field, leveraging computer vision techniques to automate the classification of diseased and healthy crops.1 Datasets play a pivotal role in the development of these models, as they provide the foundation for training and validating machine learning algorithms.1

MangoFruitDDS is a dedicated dataset designed specifically for the detection of mango fruit diseases.1 It consists of 1700 high-resolution images, categorized into five classes: healthy mangoes and four diseased categories—Alternaria, Anthracnose, Black Mould Rot, and Stem and Rot.1 The images were collected in an orchard in Senegal, reflecting real-world conditions, and were captured using a mobile phone camera.1 This makes the dataset particularly relevant for practical applications, as mobile-based disease detection systems could be deployed in the field.1

### Research Objective

This project focuses on applying multiple supervised machine learning algorithms to the MangoFruitDDS dataset to classify mango fruit images as healthy or diseased.1 The selected models include logistic regression, lasso regression, ridge regression, k-nearest neighbors (KNN), naive Bayes, and support vector machines (SVM).1 Each algorithm is chosen for its specific strengths in handling classification tasks.1

By systematically applying and comparing these techniques, the study aims to 1:

* Identify the most accurate and efficient ML algorithm for mango disease classification.1
* Explore the impact of image preprocessing (e.g., background removal) on model performance.1
* Contribute to the development of scalable, cost-effective, and reliable solutions for automated disease diagnosis in agriculture.1

The findings from this research have broader implications for the field of precision agriculture, particularly in low-resource settings where traditional methods are less viable.1 By demonstrating the potential of data-driven approaches, this study contributes to the growing body of work aimed at integrating AI and machine learning into agricultural practices.1

## II. LITERATURE REVIEW

The agricultural sector faces numerous challenges, including plant diseases, pests, and environmental stresses, all of which can adversely affect crop yield and quality.1 In recent years, the integration of artificial intelligence (AI) and machine learning (ML) techniques in agriculture has opened new possibilities for improving disease detection, optimizing crop management, and ensuring food security.1 This section provides an overview of current research on plant and fruit disease detection using machine learning, with a focus on algorithmic strategies, preprocessing techniques, and the gaps in existing research that the current study aims to address.1

### Machine Learning in Plant Disease Detection

Machine learning techniques have revolutionized plant disease detection by enabling automatic identification and classification of diseases from images.1 In the early stages, handcrafted features such as color, texture, and shape were the primary focus.1 For example, Al-Hiary et al. (2011) applied k-means clustering to extract features to classify tomato leaf diseases.1 While effective in controlled environments, such methods struggle to generalize across diverse field conditions.1

The limitations of feature engineering led to the adoption of deep learning techniques.1 Convolutional neural networks (CNNs) have emerged as the dominant approach due to their ability to learn hierarchical representations from raw image pixels.1 Mohanty et al. (2016) demonstrated the effectiveness of CNNs in classifying 26 diseases across 14 crop species, highlighting their potential to outperform traditional methods.1 However, the extensive computational requirements and the need for large, labelled datasets pose challenges.1 For practical deployment in low-resource settings, traditional ML algorithms can provide an attractive alternative due to their lower computational overhead.1

### Disease Detection in Fruits

While plant disease detection research has historically concentrated on leaves, fruit disease detection has received increasing attention.1 The significance is particularly important because fruit quality directly impacts commercial viability.1

Various studies have explored ML to detect diseases in fruit crops like citrus, bananas, and apples.1 Zhang et al. (2019) applied CNNs to classify citrus fruit diseases, demonstrating the advantages of preprocessing techniques such as background removal and image normalization.1 Arivazhagan et al. (2013) used support vector machines (SVMs) to detect fungal infections in banana fruits.1 Despite these advances, research focusing on mango fruit diseases remains limited.1 The MangoFruitDDS dataset provides a unique opportunity to study disease detection specifically in mango fruits.1

### The Role of Preprocessing in Image-Based Classification

Preprocessing plays a crucial role in preparing images for machine learning algorithms, especially in agricultural datasets where real-world conditions introduce noise and irrelevant features.1 Effective preprocessing enhances image quality and ensures models focus on the most relevant features.1

One of the most widely used techniques is background removal, which helps eliminate irrelevant regions of the image.1 Zhan et al. (2020) demonstrated that combining background removal with histogram equalization significantly improved model performance in detecting fungal infections in tomatoes.1 In the MangoFruitDDS dataset, the background-removed version (SenMangoFruitDDS\_bgremoved) allows for a direct evaluation of the impact of preprocessing on the performance of traditional machine learning algorithms.1

### Comparative Performance of Machine Learning Algorithms

The performance of ML algorithms in plant disease detection is influenced by the choice of model, dataset nature, and extracted features.1 Linear models like logistic regression are popular for their simplicity and interpretability.1 Regularized versions, such as lasso and ridge regression, help mitigate overfitting.1 Kumar et al. (2022) demonstrated the effectiveness of lasso regression in plant disease detection.1

Non-linear models such as K-nearest neighbors (KNN) are known for capturing local patterns, making them well-suited for smaller datasets.1 Sharma et al. (2020) applied KNN to classify rice pest infestations with competitive results.1 Naive Bayes classifiers are widely used where speed and scalability are important.1 Support vector machines (SVMs) have been widely adopted due to their ability to handle high-dimensional feature spaces and non-linear decision boundaries, as seen in the work of Arivazhagan et al. (2013) on banana fruit diseases.1

### Gaps in Existing Research

Despite significant advances, several gaps remain.1 Most research has focused on leaf disease detection, with relatively little attention paid to fruit-specific disease detection, especially for mangoes.1 Another key gap is the limited exploration of traditional machine learning algorithms in comparison to deep learning approaches, which can be computationally prohibitive in resource-constrained environments.1 This study aims to address these gaps by systematically evaluating a range of traditional machine learning models on the MangoFruitDDS dataset, comparing performance on both original and background-removed images.1

## III. METHODOLOGY

This section presents an in-depth exploration of the methodology used to assess various machine learning algorithms for the detection of mango fruit diseases.1 The primary objective is to apply logistic regression, lasso regression, ridge regression, k-nearest neighbors (KNN), naive Bayes, and support vector machines (SVM) on the MangoFruitDDS dataset, evaluating the effect of preprocessing techniques.1

### A. Description of the Dataset

The MangoFruitDDS dataset comprises a total of 838 images of mango fruits categorized into five classes: four disease categories (Alternaria, Anthracnose, Black Mould Rot, Stem Rot) and one healthy fruit class.1 The images have a resolution of $224 \times 224$ pixels in JPG format and are sourced from an orchard in Senegal, captured with a mobile phone camera.1 The dataset is available in two versions: SenMangoFruitDDS\_original (unaltered images) and SenMangoFruitDDS\_bgremoved (background-removed images).1

### B. Data Preprocessing

Given the visual nature of the data, preprocessing is crucial for enhancing the model's ability to accurately detect diseases.1

**Preprocessing Steps:**

* **Background Removal:** Background removal is performed to focus on the mango fruit itself, isolating it from the surrounding environment.1 This preprocessing step is essential because irrelevant background pixels could introduce noise.1
* **Image Resizing:** The raw images from the dataset are resized to a uniform dimension of $224 \times 224$ pixels to ensure consistency across all inputs.1
* **Normalization:** Normalization is applied to scale the pixel values to the range of .1 This step reduces the likelihood of any one pixel disproportionately influencing the model.1
* **Data Augmentation:** Given the relatively small size of the dataset (1700 images), data augmentation is applied to artificially increase the size of the training set.1 Techniques include random rotations, flipping, zooming, and adjusting brightness and contrast.1

Data Splitting:

The dataset is divided into three distinct subsets 1:

* **Training Set (80%):** Used to train the machine learning models.1
* **Validation Set (10%):** Used to tune hyperparameters and adjust the model's configuration.1
* **Test Set (10%):** Used exclusively to evaluate the model's performance after training and validation are complete.1

### C. Feature Extraction and Selection

Feature extraction is a crucial step in transforming the raw image data into a format suitable for traditional machine learning algorithms.1

* **Color Features:** The color of the mango fruit plays a significant role in identifying disease symptoms.1 Color histograms are computed for the three primary color channels (Red, Green, and Blue) to capture the distribution of pixel intensities.1
* **Texture Features (GLCM):** Texture features are derived from the Gray Level Co-occurrence Matrix (GLCM), a method for analysing spatial relationships between pixel intensities.1 Statistical measures of texture, such as contrast, correlation, energy, and homogeneity, are extracted.1

These combined color and texture features are formed into feature vectors and standardized by normalization to serve as inputs for the machine learning algorithms.1

### D. Model Implementation

The study applies six different machine learning algorithms to the MangoFruitDDS dataset.1

* **Logistic Regression (LR):** A foundational classification algorithm used for binary classification, extended to multi-class classification using the one-vs-all approach.1 It is simple, fast, and works well when the relationship between features and classes is relatively linear.1
* **Lasso Regression:** A regularized form of logistic regression that uses L1 regularization to enforce sparsity and perform feature selection by shrinking some coefficients to zero.1
* **Ridge Regression:** Another variant of logistic regression that applies L2 regularization to reduce model complexity and prevent overfitting by penalizing large coefficients.1
* **K-Nearest Neighbours (KNN):** A non-parametric algorithm that classifies new data points based on the majority class of their k nearest neighbours.1 It is effective for smaller datasets where local patterns are important.1
* **Naive Bayes:** A probabilistic classifier based on Bayes' theorem, assuming independence between features.1 It is simple, fast, and effective, especially with many features.1
* **Support Vector Machines (SVM):** A powerful classification model known for its ability to handle high-dimensional spaces and non-linear decision boundaries using kernels (e.g., radial basis function - RBF).1

### E. Model Evaluation and Performance Metrics

To evaluate the performance of the machine learning models, several performance metrics are used.1

* **Classification Accuracy:** The overall percentage of correctly classified images in the test set.1
* **Confusion Matrix:** An essential tool for providing a detailed breakdown of prediction results, showing the counts of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) for each class.1
* **Precision-Recall Curve:** This curve visualizes the trade-off between precision (the proportion of positive predictions that were actually correct) and recall (the proportion of actual positives that were correctly identified).1 It is particularly important for imbalanced datasets.1
* **Area Under the Curve (AUC):** AUC refers to the area under the Receiver Operating Characteristic (ROC) curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR).1 An AUC value of 1 represents a perfect model, while 0.5 indicates performance equivalent to random guessing.1
* **Box Plot:** A box plot is used to visualize the distribution of classification accuracy across different models, helping to understand consistency and stability.1

## IV. RESULTS

The evaluation of the machine learning algorithms on the MangoFruitDDS dataset revealed distinct performance patterns across the employed metrics.1

### A. Performance Metrics

* **Logistic Regression:** Achieved an overall accuracy of 80%.1 The confusion matrix showed it performed moderately well, particularly in identifying "Healthy" (98% recall) and "Stem and Rot" (90% recall), but struggled with "Alternaria" (61% precision and recall).1 The macro-averaged precision, recall, and F1-score were consistent at 80%.1
* **Ridge Regression:** Achieved a lower accuracy of 43%.1 The confusion matrix indicated struggles with classifying "Alternaria" (24% recall) and "Anthracnose" (31% recall).1 This suggests L2 regularization alone was insufficient for the dataset's complexity.1
* **Lasso Regression:** Exhibited poor performance with an accuracy of 38%.1 Recall for "Alternaria" and "Stem and Rot" was only 6% and 16%, respectively.1 The L1 regularization likely caused excessive feature elimination, leading to underfitting.1
* **Naive Bayes:** Demonstrated mixed performance with a moderate accuracy of 58%.1 It achieved high recall for "Alternaria" (61%) and "Anthracnose" (65%) but struggled with "Black Mould Rot" (32% recall).1 Its assumption of feature independence likely limited its performance.1
* **K-Nearest Neighbours (KNN):** Achieved an impressive accuracy of 90%.1 It showed strong performance in "Alternaria" (F1-score: 94%) but slightly lower results for "Anthracnose" (F1-score: 73%).1
* **Support Vector Machines (SVM):** Achieved an accuracy of 78%, similar to logistic regression.1 The confusion matrix showed high precision for "Healthy" (100%) and "Anthracnose" (85%), while recall was excellent for "Stem and Rot" (94%) and "Healthy" (95%).1 The model balanced precision and recall effectively.1

### B. Comparative Analysis

**Summary of Results** 1

| **Model** | **Accuracy** | **Weighted Avg Precision** | **Weighted Avg Recall** | **Weighted Avg F1-Score** |
| --- | --- | --- | --- | --- |
| Logistic Regression | 80% | 81% | 80% | 80% |
| Ridge Regression | 43% | 45% | 43% | 43% |
| Lasso Regression | 38% | 54% | 38% | 34% |
| Naive Bayes | 58% | 60% | 58% | 58% |
| KNN | 90% | 90% | 90% | 90% |
| SVM | 78% | 78% | 78% | 78% |

**Insights from the Comparison:** 1

1. **Best-Performing Model:** KNN achieved the highest accuracy (90%) and F1-score, indicating it excelled at capturing the local structure of the data.1
2. **Balanced Performance:** Logistic regression and SVM demonstrated balanced precision and recall, making them strong candidates for datasets with more uniform class distributions.1
3. **Impact of Regularization:** Both lasso and ridge regression underperformed, suggesting that regularization alone was insufficient to handle the complexity of the dataset.1
4. **Naive Bayes:** While simpler and faster, Naive Bayes struggled with overlapping features but performed moderately well for certain classes.1
5. **Confusion and Misclassification:** All models faced challenges in distinguishing between "Alternaria" and "Anthracnose," likely due to overlapping features.1

By comparing metrics, it is evident that KNN and SVM stand out as the most effective algorithms for this task, while logistic regression offers a simpler, interpretable alternative.1

## V. DISCUSSION

This section provides an in-depth analysis of the performance of the various machine learning models used for mango fruit disease detection.1

### A. Significance of Findings

Model Performance Analysis

Each model demonstrated unique strengths and weaknesses 1:

1. **Support Vector Machine (SVM):** SVM was a top-performing model (78% accuracy).1 The radial basis function (RBF) kernel's ability to handle non-linear separations allowed it to effectively classify disease patterns, corroborating previous studies on SVM's robustness in agricultural applications.1
2. **K-Nearest Neighbours (KNN):** KNN delivered the most competitive results (90% accuracy), particularly with small $k$-values.1 Its simplicity and reliance on feature similarity made it effective, though its computational inefficiency for larger datasets poses a limitation.1
3. **Naive Bayes:** Performed moderately well (58% accuracy) but was hindered by its assumption of feature independence, which does not hold for image data where pixel values are correlated.1
4. **Regression Models (Logistic, Ridge, Lasso):** Logistic regression (80% accuracy) served as a decent baseline classifier but was limited by its assumption of linear separability.1 Ridge (43%) and Lasso (38%) regression underperformed significantly.1 Lasso's feature selection likely eliminated too many relevant features, while Ridge's inability to handle non-linear relationships restricted its ability to classify complex patterns.1

Impact of Preprocessing

Preprocessing played a critical role in enhancing the performance of all models.1 The dataset was pre-processed using background removal, which eliminated irrelevant information and focused the models on the diseased areas of the fruit.1 This approach improved accuracy across classifiers.1 Techniques such as feature scaling, normalization, and data augmentation further contributed to better generalization by minimizing noise and environmental variations.1

### B. Implications for Agriculture

The findings of this study have several practical applications 1:

* **Mobile Diagnostic Tools:** The high accuracy of KNN and SVM makes them suitable candidates for integration into mobile applications for real-time disease detection.1 These tools could provide farmers with actionable insights, enabling timely interventions.1
* **Educational Applications:** Simpler models, with their interpretability, can be used as training aids for farmers and agricultural professionals to understand disease patterns and classification rules.1
* **Standardized Data Collection:** The importance of preprocessing highlights the need for standardized protocols in image acquisition, such as consistent lighting, controlled environments, and background removal, to enhance model reliability in real-world settings.1

### C. Limitations

Several challenges were encountered during the study 1:

1. **Dataset Size:** The relatively small size of MangoFruitDDS constrained the ability of models to generalize effectively.1 Data augmentation techniques partially addressed this issue.1
2. **Class Imbalance:** The dataset exhibited uneven representation of disease classes, which can lead to biased training, especially in models like Naive Bayes and logistic regression.1
3. **Feature Complexity:** Diseases such as Alternaria and Anthracnose displayed overlapping symptoms, making it difficult for simpler models to differentiate between them.1
4. **Computational Demands:** Models like KNN and SVM required significant computational resources.1

### D. Future Work

This study opens avenues for further research and development 1:

1. **Deep Learning Integration:** Convolutional neural networks (CNNs) could automate feature extraction, potentially outperforming traditional models.1 Transfer learning using pre-trained networks could address the dataset size limitation.1
2. **Dataset Expansion:** Expanding MangoFruitDDS to include more samples, diverse disease categories, and varying environmental conditions would enhance model robustness.1
3. **Real-Time Systems:** Deploying these models in IoT-enabled systems or mobile applications could facilitate real-time disease detection.1
4. **Hybrid Approaches:** Combining the strengths of multiple models, such as using regression for feature selection and SVM for classification, could yield better results.1

## VI. CONCLUSION

This study evaluated six machine learning algorithms—logistic regression, ridge regression, lasso regression, KNN, naive Bayes, and SVM—for mango fruit disease detection using the MangoFruitDDS dataset.1 KNN emerged as the best-performing model, achieving the highest accuracy (90%) and excelling in capturing local patterns, while SVM provided robust and balanced performance with a precision and recall of 78%.1

Background removal as a preprocessing step significantly improved model performance by reducing noise and emphasizing disease-specific features.1 However, linear models like ridge and lasso regression underperformed due to the dataset's complexity and overlapping features among disease classes.1

Challenges such as class imbalance and feature overlap impacted the models' ability to classify minority classes accurately.1 Expanding the dataset and exploring deep learning techniques, such as convolutional neural networks, could enhance future performance.1

This study underscores the potential of machine learning, particularly KNN and SVM, in developing reliable and cost-effective tools for mango disease detection.1 These findings contribute to advancing precision agriculture and improving sustainable farming practices.1

## REFERENCES

1. Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). Using deep learning for image-based plant disease detection. *Frontiers in Plant Science, 7*, 1419. 1
2. Zhang, S., Huang, W., & Liu, D. (2019). A review of advances in crop disease detection based on hyperspectral and multispectral imaging. *Computers and Electronics in Agriculture, 169*, 105223. 1
3. Arivazhagan, S., Newlin Shebiah, R., Ananthi, S., & Vishnu Varthini, S. (2013). Detection of unhealthy region of plant leaves and classification of plant leaf diseases using texture features. *Agricultural Engineering International: CIGR Journal, 15*(1), 211-217. 1
4. Sharma, P., Kumar, S., & Joshi, A. (2020). Applications of machine learning in agriculture: A review of the advancements in plant disease detection. *International Journal of Computational Science and Information Technology, 8*(1), 1-9. 1
5. Al-Hiary, H., Bani-Ahmad, S., Reyalat, M., Braik, M., & ALRahamneh, Z. (2011). Fast and accurate detection and classification of plant diseases. *Machine Learning in Plant Science, 12*(1), 1-8. 1
6. Preetham, H. D., Kumar, P. A., & Shwetha, S. (2021). Enhancing fruit disease detection with preprocessing techniques: A study on apples. *International Journal of Computer Applications in Agricultural Engineering, 23*(2), 45-52. 1
7. Ghosh, S., Debnath, D., & Sen, A. (2021). Convolutional neural networks for identifying mango leaf diseases. *Indian Journal of Agricultural Research, 55*(3), 383-390. 1
8. Kumar, R., Gupta, V., & Singh, D. (2022). Regularization techniques in logistic regression for plant disease prediction. *Journal of Data Science in Agriculture, 14*(4), 67-78. 1
9. Zhan, X., Wu, J., & Liu, H. (2020). Improving classification accuracy of tomato fungal infections using background removal. *Applied Computational Biology, 7*(2), 23-30. 1