AI

Karan
Graphic

Parveen
Graphic

Aman
Graphic

Graphic

Graphic

Abstract—

We

Keywords—

I. Introduction

Numerous
Citrus
A
The
Citrus
As
Image
Threshold
A
Annamalai
Categorization
Lu
In

II. Literature Review

Citrus

Research
In
Jianqiang

In

The

III. Methodology

Dataset:
In
.
The image of each dataset is given in .

Feature Extraction:
The
analyze
This

Model used:
A

1. K-Nearest Neighbour

K-

Class	Training images	Testing images
LIMON CRIOLLO	2256	250
LIMONMANDARINO	2439	270
LIMON TAHITI	3074	341
MANDARINA ISRAELI	2228	247
MANDARINA PIELDESAPO	2322	257
NARANJA VALENCIA	2375	263
TANGELO	1877	208
TORONJA	2098	233

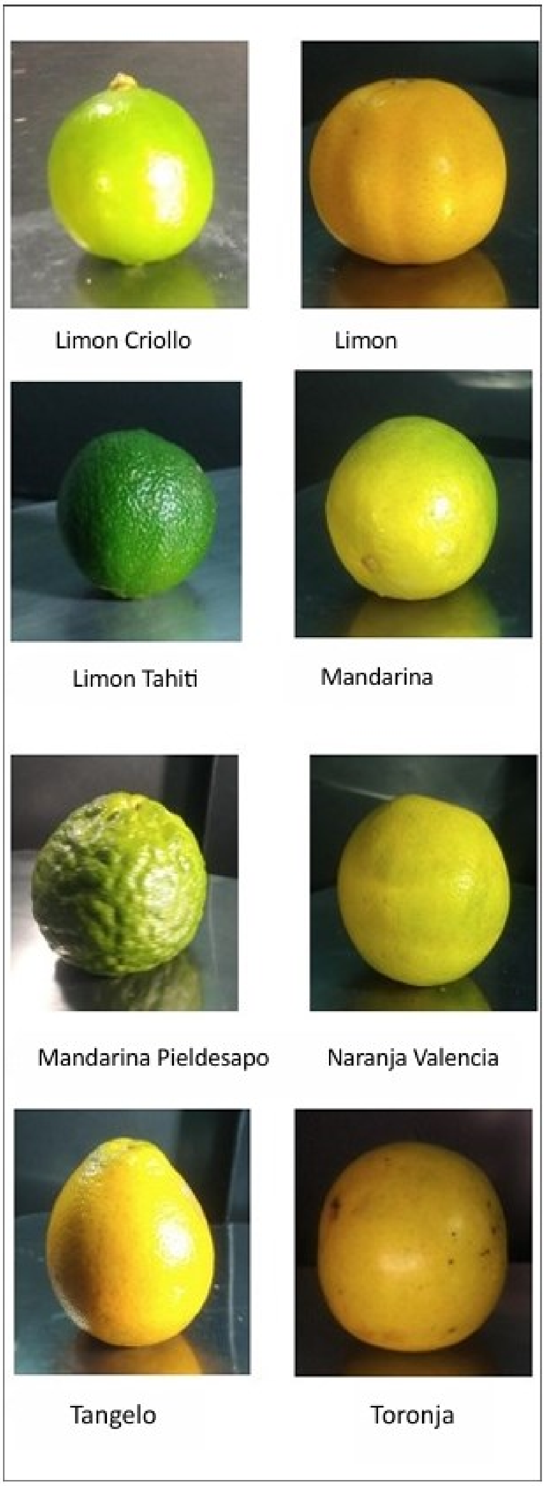

Figure 1: Citrus fruit dataset images

The

D
i
s
t

Euclidean:

n
c
e

$$\sqrt{\sum_{i=1}^{k}(xi-yi)^2} \quad \textbf{(1)}$$

Manhattan:

$$\sum_{i=1}^{k}|xi-yi| \quad \textbf{(2)}$$

s

Hamming distance:

c
a
l

$$D_H = \sum_{i=1}^{k}|x_i - y_i| \quad \textbf{(3)}$$

We

2. Decision Tree:

l
a
t
e

A

Table 1: Citrus classification and quantity of images

It

For every node, it calculates the entropy which is the degree of uncertainty in the dataset, and it is calculated by the equation $H(i) = -\sum_{\kappa\varepsilon K}^{n} P(i,k)\, log_2\, P(i,k)$ *(4.*

At every level, the uncertainty of the dataset reduces which is calculated by the equation and this is called information gain or variance. Refer to equation

Class	Training images	Testing images
LIMON CRIOLLO	2256	250
LIMONMANDARINO	2439	270
LIMON TAHITI	3074	341
MANDARINA ISRAELI	2228	247
MANDARINA PIELDESAPO	2322	257
NARANJA VALENCIA	2375	263
TANGELO	1877	208
TORONJA	2098	233

$$Information\ Gain(H,A) = H - \sum \frac{|Hv|}{|H|} H_v \quad (5).$$

$$H_i = -\sum_{\kappa\varepsilon K}^{n} P(i,k)\, log_2\, P(i,k) \quad (4)$$

Where H_i=entropy k= k is the class from K classes,
p(k)= proportion of the data points that belong to class k.

$$Information\ Gain(H,A) = H - \sum \frac{|Hv|}{|H|} H_v \quad (5)$$

Where A =class label or specific attribute, |H| is the entropy of dataset sample S, $|H_V|$ is the number of instances in the subset S *that have the value v for attribute A*

This is also used as an alternative to the KNN model for the detection of citrus fruits it is simple to understand and it needs less data cleaning compared to other models.

3. Random Forest:

Random Forest is an enhanced and better version of a decision tree however it is also a machine learning algorithm that is based on the supervised learning technique which is implemented through a decision tree and used in regression and classification problems. It is a collection of decision trees and it gives the result based on several votes. The more the decision trees the higher the accuracy.

We used the Random Forest for detection of citrus fruits and the results were best through this compared to other models. This is also used for detection problems and many other problems and provides high-accuracy prediction.

4. Gradient Boosting:

Gradient Boosting is based on a supervised learning technique and on an ensemble boosting technique which is highly versatile and can be used for classification and regression problems. Weights are assigned to the data values. It takes a learner or model that has better prediction than the preceding weak learner. Weak learners are decision stumps or decision trees with one node and 2 leaf nodes. The loss function is used for finding the loss i.e. the difference between the actual value and the predicted variable given in the equation $L(f) = \sum_{i=1}^{N} L(y_i, f(x_i))$ (6).

$$L(f) = \sum_{i=1}^{N} L(y_i, f(x_i)) \quad (6)$$

Where $L(f)$ = loss function, y_i=targetvalues, $f(x_i)$=function to map the input features x to target values y.

This was used for the classification of citrus fruit in our study and trained over 80% of i.e. 18669 jpg images and the results were optimal.

5. AdaBoost:

Adaptive Boost is a powerful machine learning ensemble boosting technique used for classification and regression problems. It iterates through all the weak learners which are decision stumps (a tree having one node and 2 leaf nodes) is trained on the given training dataset and combines these models to provide better results and predict the output of the problem. Steps of AdaBoost:

- Weight is initialized for each data point. Let's say if there are N datapoint then weight W{i}=1/N.
- Iteration through weak classifier is done and they are trained on the dataset.
- Error rate is calculated for each weak classifier and the importance of each weak model through equation
- $importance = 1/2 * log(1 - totalerror/totalerror)$ *(7.*
  $importance = 1/2 * log(1 - totalerror/totalerror)$ *(7)*
- Weights of the data points are updated through the given equation NewSampleWeight = sampleweight*exp^importance *(8.*

NewSampleWeight = sampleweight*exp^importance *(8)*

- To obtain the sum of weight 1 we normalize the weights using the equation $W\{i\} = W\{i\}/sum(W)$ *(9.*

$W\{i\} = W\{i\}/sum(W)$ *(9)*

- These steps are repeated until all stumps are iterated.

This model is also used in this study for the classification and detection of citrus fruit.

IV. Result and Discussion

A total of 8 citrus families (Limon Criollo, Limon Mandarino, Limon Tahiti, Mandarina Israeli, Mandarina Pieldesapo, Naranja Valencia, Tangelo, and Toronja) were classified based on texture parameters. Combinedly 2069 images taken in normal conditions were used for testing the models. Table 2 Shows the average accuracy, precision, recall, and F1 scores of all the algorithms used in this study. The brief discussion of accuracy of each model is given below:

Table 2: Average Accuracy, Precision, Recall and F1-score

Algorithm	Accuracy	Precision	Recall	F1-score
K-Nearest Neighbors	74.90%	75%	75%	75%
Random Forest	75.75%	76%	76%	76%
Decision Tree	63.18	63%	63%	63%
Gradient Boosting	73.86%	74%	74%	74%
AdaBoost	47.69%	47%	48%	46%

1. **K-Nearest Neighbours:**

A total of 20% of data was used for the testing of the model i.e. a total of 2069 images collectively of all classes of citrus were provided and we came up with an accuracy of 74.90% i.e. roughly 1550 images were detected correctly and a confusion matrix for the same is given in *fig 1*.

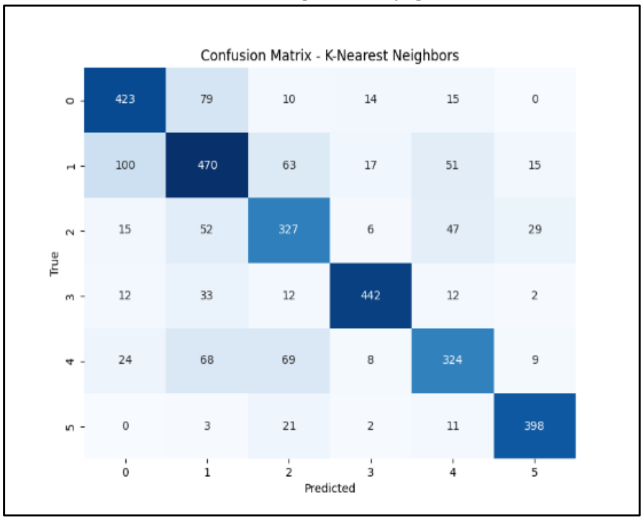

fig 1: KNN

2. **Decision Tree** In this also we gave 2069 images for model testing and it came up with an accuracy of 63.18% i.e. 1308 images were detected without any problem.

The confusion matrix for this is given in

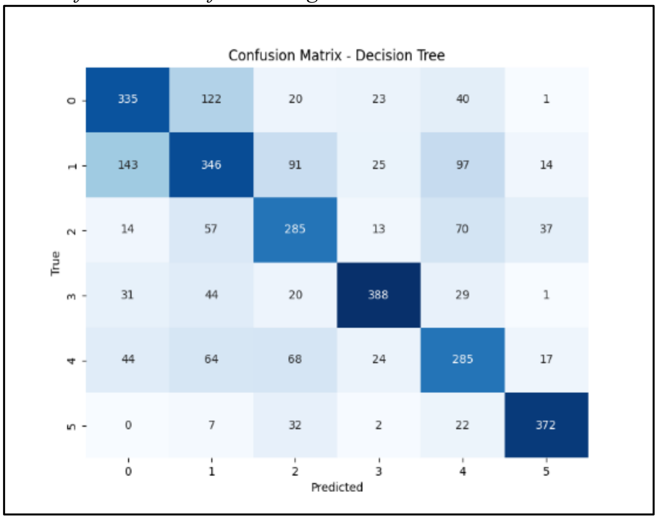

fig 2.

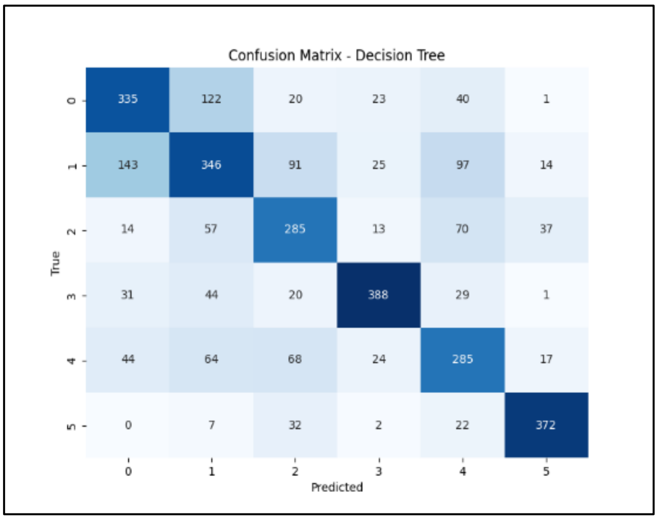

fig 2: Decision Tree

3. **Random Forest:**
   This model gave us the best accuracy compared to all the algorithms for the 2069 images of the citrus dataset. The accuracy was 75.75% i.e. 1568 images were detected without any problem.
   The confusion matrix for this is given in fig *4*.
4. **Gradient Boosting:**
   This algorithm also came up with 1528 successful detection out of 2069 images and gave an accuracy of 73.86%.
   The confusion matrix for gradient boosting is given in

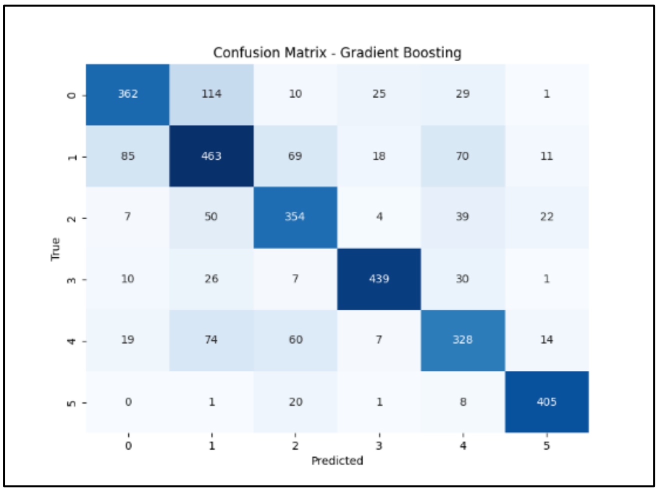

fig 3: Confusion Matrix-Gradient Boosting

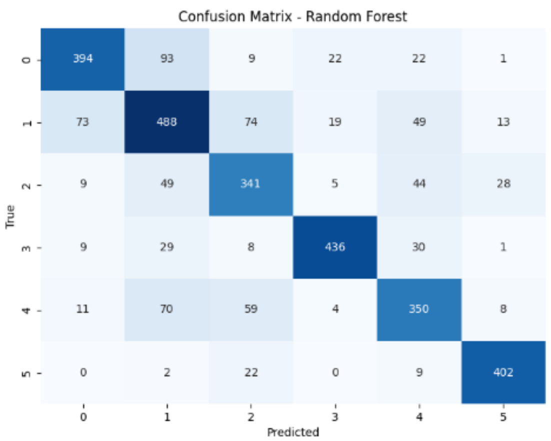

fig 4: Confusion Matrix-Random Forest

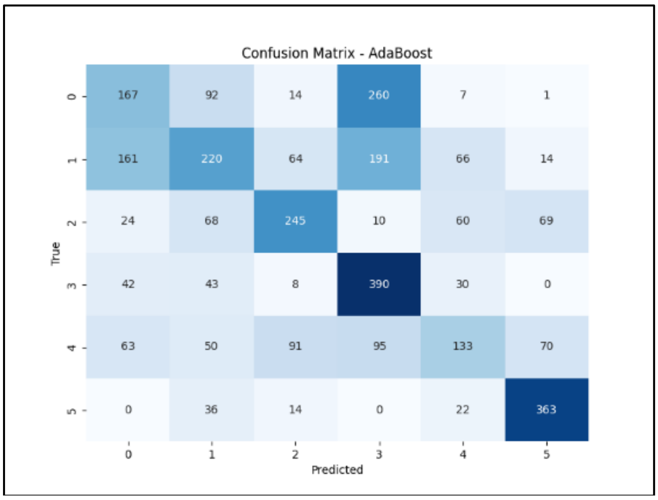

fig 5: Confusion Matrix-AdaBoost

V. CONCLUSION AND FUTURE SCOPE

In this research, we have proposed 5 different methods for the detection and classification of citrus fruits based on supervised learning techniques. We first acquired the data from Kaggle and used the LBP local binary pattern for the data feature extraction from the jpg images of the dataset. We used KNN, Decision Tree, Random Forest, Gradient Boosting, and AdaBoost. The average accuracies of the model were 74.90%,63.18%,75.75%,73.86%, and 47.69%.

It has so many applications in the future and presents also it can be used in predicting the yield of citrus fruit in the agricultural fields, we can use it in Citrus picking robots or in any (UAVs) i.e. unnamed aerial vehicles, detecting citrus is also necessary if any citrus is used as a raw material for making any final product. furthermore, it can be used in shops to detect the category of citrus fruit people buy or sell. it can also be used for the differentiation of healthy and bad citrus fruit.

.

5. **AdaBoost:**
   Again, this Boosting algorithm successfully came up with the detection of 987 images out of 2069. This gave the lowest accuracy of 47.69%.

The confusion matrix for AdaBoost is given in
fig *5***Error! Reference source not found.**.

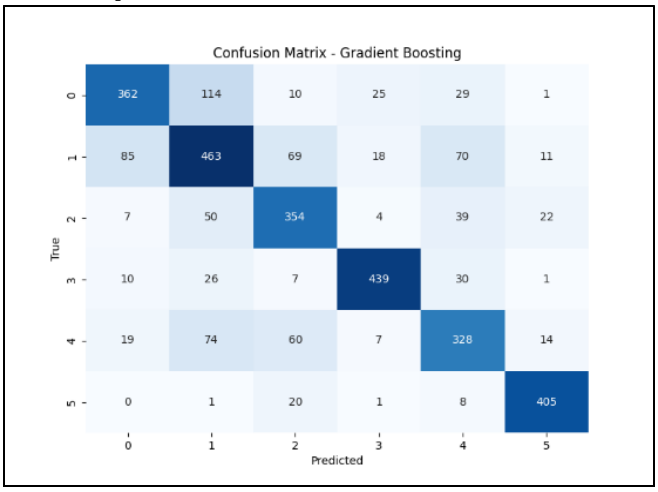

fig 3: Confusion Matrix-Gradient Boosting

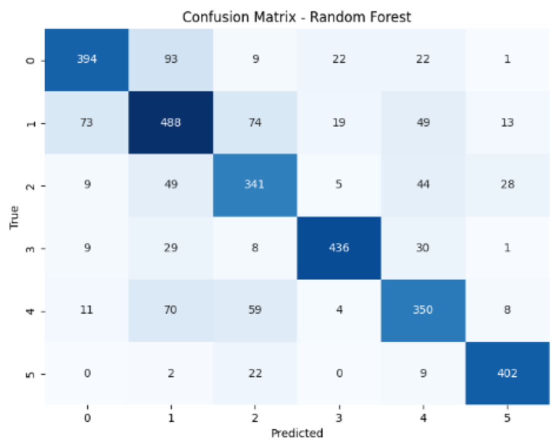

fig 4: Confusion Matrix-Random Forest

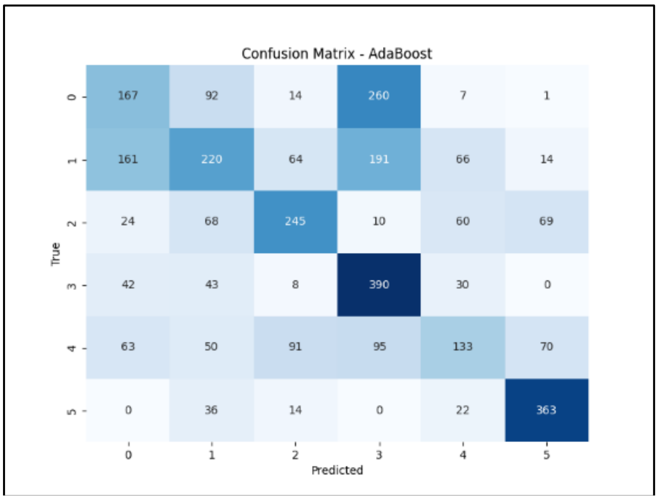

fig 5: Confusion Matrix-AdaBoost

VI. Conclusion and Future Scope

In this research, we have proposed 5 different methods for the detection and classification of citrus fruits based on supervised learning techniques. We first acquired the data from Kaggle and used the LBP local binary pattern for the data feature extraction from the jpg images of the dataset. We used KNN, Decision Tree, Random Forest, Gradient Boosting, and AdaBoost. The average accuracies of the model were 74.90%,63.18%,75.75%,73.86%, and 47.69%.

It has so many applications in the future and presents also it can be used in predicting the yield of citrus fruit in the agricultural fields, we can use it in Citrus picking robots or in any (UAVs) i.e. unnamed aerial vehicles, detecting citrus is also necessary if any citrus is used as a raw material for making any final product. furthermore, it can be used in shops to detect the category of citrus fruit people buy or sell. it can also be used for the differentiation of healthy and bad citrus fruit.

VII. Authors Contribution

All the members contributed to data collection, model design, feature extraction methods writing the research paper, and doing a study on the topic.

References

[1] S. M. Haldhar and N. Thaochan, “Citrus: production and management in NEH region.” [Online]. Available: https://www.researchgate.net/publication/370208165

[2] M. N. Sharif, U. Farooq, and W. A. Malik, “Citrus Marketing in Punjab: Constraints and Potential for Improvement,” *The Pakistan Development Review*, vol. 44, pp. 673–694, 2005, [Online]. Available: https://api.semanticscholar.org/CorpusID:56371982

[3] N. Akhter, M. Idrees, F. U. Rehman, M. Ilyas, Q. Abbas, and M. Luqman, “Shape and texture-based classification of citrus using principal component analysis,” *International Journal of Agricultural Extension*, vol. 9, no. 2, pp. 229–238, 2021, doi: 10.33687/ijae.009.02.3525.

[4] F. Pl~i, F. Juste, and F. Ferri, “Feature extraction of spherical objects in image analysis: an application to robotic citrus harvesting,” 1993.

[5] J. Lu *et al.*, “Citrus green fruit detection via improved feature network extraction,” *Front Plant Sci*, vol. 13, Nov. 2022, doi: 10.3389/fpls.2022.946154.

[6] S. Kumar, S. N. Jena, and N. K. Nair, “ISSR polymorphism in Indian wild orange (Citrus indica Tanaka, Rutaceae) and related wild species in North-east India,” *Sci Hortic*, vol. 123, no. 3, pp. 350–359, Jan. 2010, doi: 10.1016/j.scienta.2009.10.008.

[7] W. Fouad Abobatta, “Nutritional Benefits of Citrus Fruits,” *Am J Biomed Sci Res*, vol. 3, no. 4, pp. 303–306, Jun. 2019, doi: 10.34297/ajbsr.2019.03.000681.

[8] J. Lu and N. Sang, “Detecting citrus fruits and occlusion recovery under natural illumination conditions,” *Comput Electron Agric*, vol. 110, pp. 121–130, Jan. 2015, doi: 10.1016/J.COMPAG.2014.10.016.

[9] H. Okamoto and W. S. Lee, “Green citrus detection using hyperspectral imaging,” *Comput Electron Agric*, vol. 66, no. 2, pp. 201–208, May 2009, doi: 10.1016/j.compag.2009.02.004.

[10] Y. Lin, Z. Huang, Y. Liang, Y. Liu, and W. Jiang, “AG-YOLO: A Rapid Citrus Fruit Detection Algorithm with Global Context Fusion,” *Agriculture (Switzerland)*, vol. 14, no. 1, Jan. 2024, doi: 10.3390/agriculture14010114.

[11] L. Xu *et al.*, “Real-time and accurate detection of citrus in complex scenes based on HPL-YOLOv4,” *Comput Electron Agric*, vol. 205, p. 107590, Feb. 2023, doi: 10.1016/J.COMPAG.2022.107590.

[12] H. Huang, T. Huang, Z. Li, S. Liu, and T. Hong, “Design of citrus fruit detection system based on the mobile platform and edge computer device,” *Sensors*, vol. 22, no. 1, Jan. 2022, doi: 10.3390/s22010059.