ISSN 2394-739X
International Journal for Research in Science Engineering and Technology

FRUIT DISEASE IDENTIFICATION USING FEATURE EXTRACTION AND DATA MINING

[1]P. Kanjana Devi, [,2] Dr. M. Rathamani
[1]Research Scholar, [,2] Associate Professor
[1,2] Dept Of Master Of Computer Applications,
[1,2] N.G.M. College Of Arts And Science, Pollachi , Tamilnadu , India.

Abstract-Agriculture is one of the fields which produce data continually ensuring each of the four characteristics with fantastic development. There are various difficulties in preparing farming records which manages assortment of organized and unstructured configuration. One of the difficulties in farming industry involves fruit infection discovery and control. For this reason ranchers needed to screen fruits ceaselessly from collect till its development period. Yet, this errand is definitely not a simple one. Subsequently it requires proposing a proficient smart cultivating strategy which will help for better yield and development with less human endeavors. Data preprocessing is a procedure which will analyze and arrange outside disorder inside fruits through different pictures. In this paper we proposed feature selection algorithm based on association rules (ARFS), considering the way that the association rule can find the association between the features attributes and the classes in the dataset, it uses the most outrageous system to learn the certainty between feature attributes and classifications.

Keywords: Data preprocessing, Fruit diseases, Association Rules, Feature selection, Features attributes.

INTRODUCTION

Agriculture field is something beyond being a feeding source in the present world. However, because of climatic and different changes throughout the long term, crop yields and agriculture yield have become inclined to certain significant issues which are a subject of genuine concern. The world economy will rely especially upon the agriculture as nowadays the production is diminishing as compared to the increment in the interest and this proportion of interest versus production is projected to be high in the upcoming years.

Plants are as helpless by diseases as creatures. Citrus is a significant plant filled chiefly in the tropical areas of the world because of its wealth in vitamin C and other significant nutrients. The production of the citrus fruit has been generally influenced by citrus diseases which

eventually debases the fruit quality and makes financial loss the cultivators. During the previous decade, image processing and computer vision techniques have been extensively received for the detection and classification of plant diseases. Early detection of diseases in citrus plants causes in forestalling them to spread in the plantations which limit the financial loss to the farmers. In this article, an image dataset citrus fruits, leaves, and stem is introduced. The dataset holds citrus fruits and leaves images of solid and tainted plants with diseases, for example, Black spot, Canker, Scab, Greening, and Melanose.

In this stage have two phases for information preprocessing. In first stage fruit is recognized by an effective combination of color and texture features. After the fruit is recognized, the recognized fruit images are move to the second stage for preprocessing. The recognition is finished by the minimum distance classifier dependent on the statistical and co-occurrence features got from the Wavelet transformed sub-groups. Color and texture are the principal character of common images, and assumes a significant part in visual discernment. For example, color features of every pixel in images got in three components of RGB spaces could be effectively used to segment abandons. Three features investigation techniques color-based, shape based and size-based are combined together to expand accuracy of recognition.

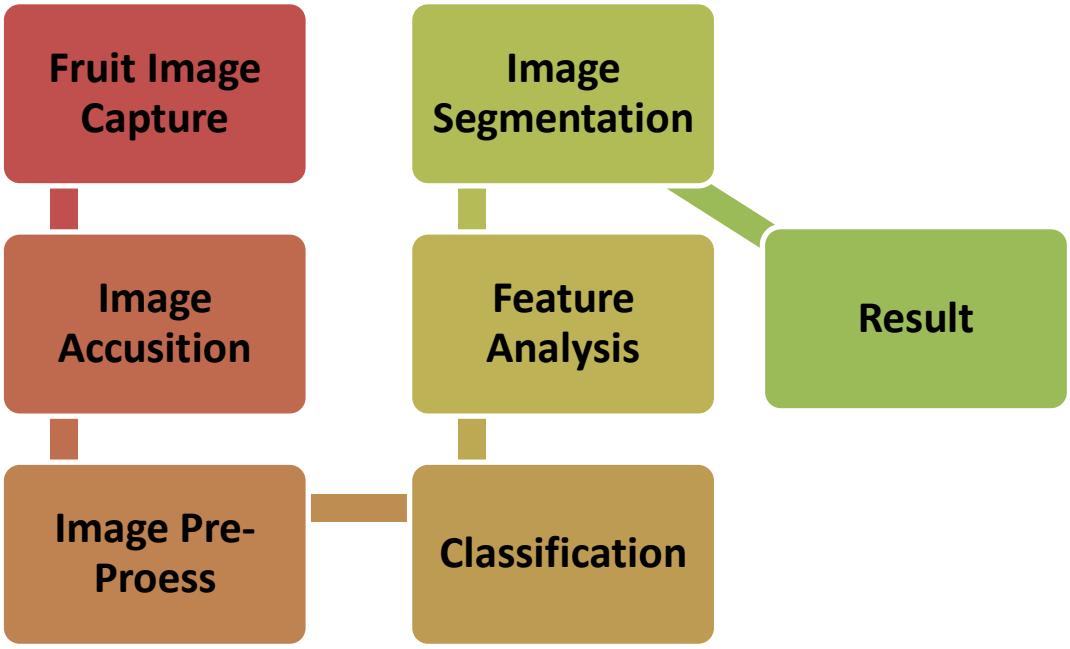


Figure 1.Fruit disease data pre-processing

Image acquisition is consistently the underlying condition for the work flow series of image processing on the grounds that as processing is conceivable just with the assistance of an image. For image segmentation, K-Means clustering method is utilized. Feature vectors, for example, image color, morphology, texture and construction of opening are applied for extricating features of each image and for conclusion of sickness morphology gives exact outcome. SURF algorithm utilized as locator and descriptor for extricating the features. To diminish the size of the feature subset and improve the effectiveness of the feature selection algorithm without lessening the accuracy, this paper proposes the ARFS. The algorithm utilizes affiliation rules to mine the regular 2-things set of the feature ascribes and class in the dataset. At that point the algorithm sorts the features according to the confidence of the continuous 2-things set.

For the development of the database, the images were taken both physically and on the web. For manual collection of images certain angles and distance was predefined to take the picture within the sight of suitable vision sources. The images hence collected were resized to the size of 250 * 250. Two datasets, separately of both ailing and un-sick images were framed. These images were utilized for both the preparation just as the testing phases.

2. Existing Methodology

2.1. SURF

SURF is a key point extraction and description algorithm that gives a comparable choice to SIFT and requires substantially less processing time for recognizing and coordinating key points. This is because of utilizing integral image technique and the more modest descriptor size (commonly, the SURF descriptor consists of 64-receptacles which is a large portion of the size of the SIFT descriptor). The concept of classification of the SURF fascinating points by the machine learning approach has been set up, actualized and confirmed. For a picked illustration of the tag detection in an image, the Naive Bayes Multivariate algorithm gave the best and promising outcomes, firmly followed by the Decision Tree algorithm.

2.2. CART

Making a CART model includes choosing input factors and split points on those factors until a reasonable tree is constructed. The selection of which input variable to utilize and the particular split or cut-point is picked utilizing a greedy algorithm to limit a cost function. Tree construction closes utilizing a predefined halting standard, for example, a minimum number of

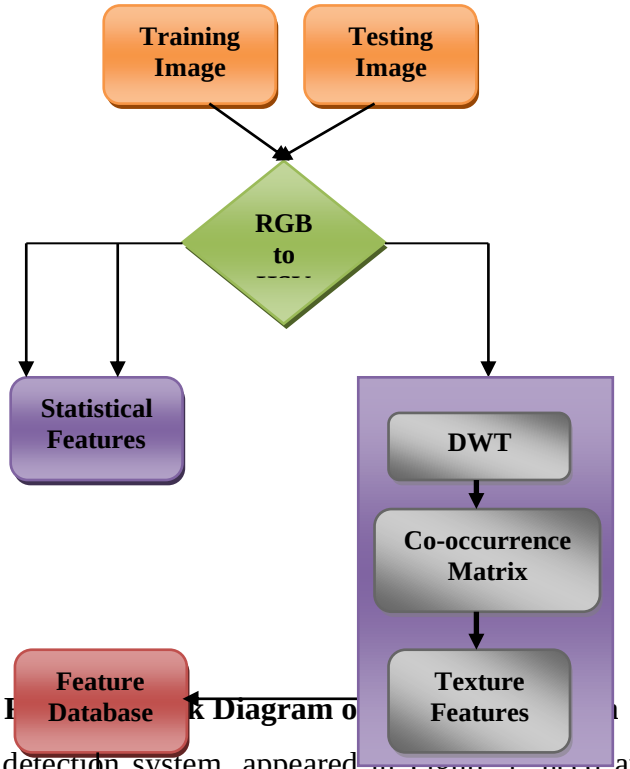
preparing occasions allocated to each leaf hub of the tree. Traditionally, this algorithm is alluded to as "decision trees", yet on certain platforms like R they are alluded to by the more current term CART. The CART algorithm gives an establishment to significant algorithms like packed away decision trees, arbitrary backwoods and supported decision trees. Truck doesn't need any uncommon information readiness other than a decent portrayal of the issue.

3. Proposed Methodology

In this stage have two phases for information preprocessing. In first stage fruit is recognized by a proficient combination of color and texture features. After the fruit is recognized, the recognized fruit images are move to the second stage for preprocessing.

Stage 1: Fruit Recognition

The two areas that engaged with this work are Training and Classification. The block chart of the Fruit Recognition is given in Figure 1.



F k Diagram o

The proposed Fruit detection system, appeared in Figure 1, need an adjustment in the color space of the images, to get one channel containing the luminance information and two different channels containing information. esentation is frequently chosen for its invariant pro de is invari rientation of an object

regarding the illumination and camera direction and thus more appropriate for object retrieval. Texture features are computed from the luminance channel 'V', and color features are computed from the chrominance channels 'H' and 'S'. The component which corresponds to brightness of the color (V) is decomposed utilizing Discrete Wavelet Transform and the co-occurrence matrix is constructed from the approximation sub-band by assessing the pair shrewd insights of pixel force. The utilization of the co-occurrence matrix depends on the theories that a similar dark level configuration is rehashed in a texture. Further, co-occurrence features, for example, contrast, energy, nearby homogeneity, cluster shade and cluster conspicuousness are determined from co-occurrence matrix C(i,j), inferred for transformed sub-bands and put away in the features library. There exist 5 co-occurrence features i.e., texture features for an image. Statistical features, for example, Mean, Standard Deviation, Skewness and Kurtosis are gotten from H and S components. Henceforth there will be 8 chrominance or color statistical features for an image. Accordingly a sum of 13 features portrays one fruit image. The test fruit image, color and texture features are inferred as that of the preparation stage and compared with corresponding feature esteems, put away in the feature library. The classification is finished utilizing the Minimum Distance Criterion. The image from the preparation set which has the minimum distance when compared with the test image says that the test image has a place with the class of that preparation image.

Mosambi Pomelo Orange Lemon

Algorithm for Extracting Region of Interest

1. The info fruit image is converted to HSV color space.
2. Perform thresholding operation on the S component, since S is substantially less delicate to lighting variety.
3. Close little openings utilizing the Closing morphological administrator with a disk structuring component.

4. Find the territory of the Region of Interest from the paired image.
5. Crop the Region of Interest and supplant the twofold qualities with the first pixel intensity.

STAGE 2: IMAGE DATA PREPROCESSING

Image data preprocessing is finished by three stages, i.e., Image acquisition, and Feature Selection. Image acquisition is consistently the underlying condition for the work flow series of image processing in light of the fact that as processing is conceivable just with the assistance of an image. Feature vectors, for example, image color, morphology; texture and construction of opening are applied for extracting features of each image. The ARFS algorithm is utilized for feature selection. The ARFS utilizes affiliation rules to discover regular 2-things set of feature credits and classifications, and afterward figures their confidence for feature selection. The related meanings of affiliation administer and continuous 2-things set mining algorithms utilized.

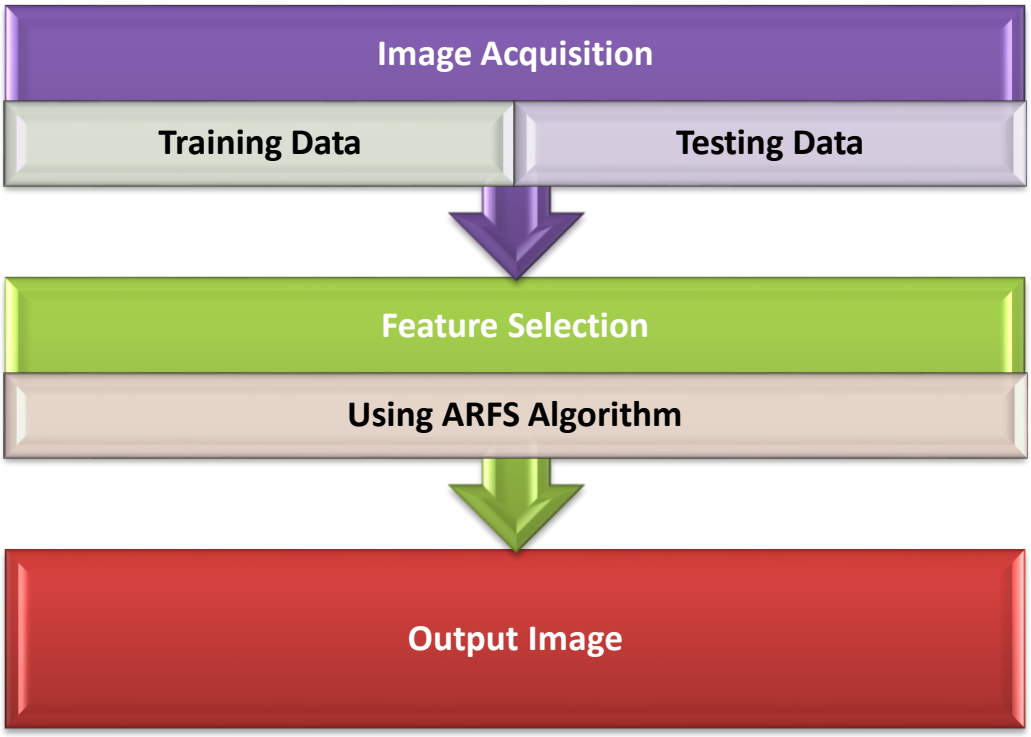


Figure 3.Image Data Preprocessing

Step 1: Image Acquisition

Image acquisition in image processing can be comprehensively characterized as the activity of recovering an image from some source, normally an equipment based source, so it tends to be gone through whatever cycles need to happen a short time later. Performing image

acquisition in image processing is consistently the initial phase in the workflow sequence in light of the fact that, without an image, no processing is conceivable. The image that is procured is completely natural and is the consequence of whatever equipment was utilized to produce it, which can be vital in certain fields to have a consistent pattern from which to work. One of a definitive goals of this cycle is to have a source of information that works inside such controlled and estimated guidelines that a similar image can, if important, be almost impeccably duplicated under similar conditions so odd components are simpler to find and kill.

Step 2: Feature Extraction

The ARFS utilizes association rules to discover successive 2-things set of feature ascribes and classes, and afterward ascertains their confidence for feature selection. The related meanings of association rules and successive 2-things set mining algorithms utilized.

Item set

Association rule is a technique for acquiring association information between data items, In the ARFS, I = {i1,i2,… , im} is a collection of items. In I, each ik is known as a venture, the quantity of items k is the length of thing set I. Every subset in thing set I is marked as T, and all correspond to a remarkable identifier, indicated as TID, the entire of T constitutes the dataset D of thing set I, and the estimation of |D| is equivalent to the quantity of T.

Association rules

The association rule R is an implication:

$$R : X \Rightarrow Y \quad \text{----------(1)}$$

$X \subset I$, $Y \subset I$ *and* $X \cap Y = \emptyset$, it means that thing set X shows up in a specific T, and makes Y show up in T with a specific probability. Among them, X and Y are called precursor and consequent of association rules individually.

Support

For thing sets X, Y, where X ⊂ I and Y ⊂ I, Let count (X ∩ Y) indicate the proportion of the convergence of the feature subsets of X and Y to |D|, at that point the help of association rule R is:

$$Support\ (X \Rightarrow Y) = count\ (X \cap Y) / |D| \quad \text{----------(2)}$$

The help of association rule R reflects the probability of simultaneous occurrence of X and Y, the minimum help in R is the minimum help threshold of the thing set, signified as

sup_min, it addresses the minimum standard for filtering association rules. The thing set whose help is more noteworthy than or equivalent to sup_min is called frequent thing set L, and the frequent set whose length is k means as Lk. The help of Lk is equivalent to the help of its association rule.

Confidence

For the association rule R, the confidence alludes to the proportion of the quantity of subsets containing X and Y to subsets of X. which is:

Confidence (X ⇒ Y) = support (X ⇒ Y) / support (X) ----------(3)

Confidence reflects the probability that T contains Y in the event that it contains X. The minimum confidence is additionally the minimum confidence threshold of the thing set, meant as conf_min. All in all, solitary association rules with high help and confidence are the rules that are in the long run discovered.

L2 Mining algorithm

In the ARFS proposed in this paper, first we need to discover L2 that is more prominent than or equivalent to sup_min between feature credits and classes in D, at that point ascertain the confidence of these frequent 2-items set. To mine frequent 2-items set L2, this paper applies the standard of Apriori algorithm and proposes a L2 mining algorithm dependent on Apriori.

L2 Mining algorithm.

```
Input ：Dataset D, Minimum support threshold sup_min
Output ：confidence conf of frequent 2-items set
1.   init T2, sup_min
2.   n = count(T2)
3.   FOR i < n DO
4.   BEGIN
5.        sup_T2i = support(T2i)        // T2i = (Xi, Y)
6.        IF sup_T2i > sup_min THEN
7.        BEGIN
8.             L2.add(T2i)
9.             conf2i = count(L2i) / count ( Xi )
10.       END   // end of if
11.  END   // end of for
12.  RETURN L2 。 Conf
```

In the above algorithm, T2 is a 2-items subset of feature attributes and categories, numerous redundant feature attributes in it. To improve the effectiveness of the algorithm and reduce the time complexity of the algorithm, in this paper, the parameter sup_min is set to sift through a piece of random redundant features ahead of time, and afterward complete the subsequent feature selection work.

Feature Selection Algorithm Based on Association Rules

In the ARFS proposed in this paper, in light of the fact that the association rule can discover the association between the feature attributes and the categories in the dataset, it utilizes the most extreme strategy to ascertain the confidence between feature attributes and categories, and utilizations this confidence to assess the correlation among features and categories. At that point sort the features by the size of their importance loads, lastly get an arranged sequence of features whose significance is from largest to littlest. The feature sequence is arranged by the pertinence loads, so choosing the SFS strategy on this premise can choose the feature subset with minimal scale however much as could reasonably be expected. Since the ARFS adopts the SFS technique to coordinate the arranged feature sequence, it can likewise reduce the time complexity of the hunt strategy.

Feature Selection Algorithm based on association rules

Input :frequent 2-items set L_2 and its confidence conf
Output: Feature subset is feature_vector

```
1.  Init Ft, feature_vector          //Ft is CART classifier
2.  Vec = Sort(Max(L2。 conf))
3.  WHILE divide == true DO
4.  BEING
5.      Set d = divide_length
6.      feature_vector.add(Vecd)
7.      accuracy_max = Max( Ft （feature_vector） )
8.      Δacc = | accuracy_max - accuracy_i |
9.      IF Δacc < 0 & i > β THEN
10.     BEGIN
11.         divide = false
12.     END  // end of if
13. END  //end of while
14. RETURN feather_vector
```

In the above algorithm, when sorting according to the confidence, L2 adopts the Max strategy that is to sort the features by the maximum confidence of feature attributes. In figuring the maximum accuracy accuracy_max of the feature subset, the Max strategy is additionally received. The accuracy_max is refreshed with the maximum accuracy determined during every iteration. In the SFS, according to the arranged feature sequence, the feature is bit by bit added to the feature subset from front to back with a specific advance length partition length. The default estimation of the gap length is 1. During the time spent iteratively choosing the ideal feature subset, the halting models is that the accuracy distinction between the current feature subset and the current ideal feature subset is under 0, $\Delta acc < 0$; It should likewise meet the condition that the frequency of $\Delta acc < 0$ is not exactly the estimation of parameter β, and the default estimation of β is the length of the first feature vector. At the point when the over two conditions are fulfilled, the iteration can be halted, and the feature subset addressed by the ARFS is at last output.

4. EXPERIMENTAL RESULT

1. Experimental results on number of attributes

Data set	SURF	CART	ARFS
Data set1	8	7	6
Data set2	11	9	7
Data set3	24	22	11
Data set4	39	31	23

Table 1. Number of Attributes in Data sets

Table 1 represents number of attributes in this table. Proposed (ARFS) values are compared with Existing values of SURF and CART. Their proposed values are higher than compare with other existing values.

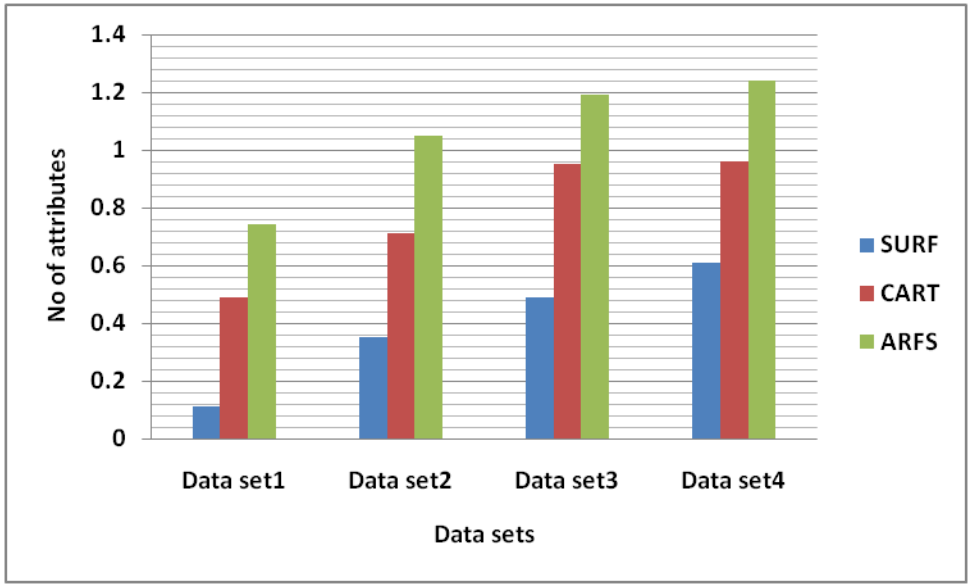


Figure 4.Number of Attributes in Data Sets

Figure 4 represents number of attributes is compare with them. All values are only positive. The proposed values are higher than in this diagram. Existing 1 is a lower than compare with existing 2 and proposed values.

2. Experimental results on classification accuracy

Data set	SURF	CART	ARFS
Data set1	75.36	85.58	96.45
Data set2	41.56	49.64	56.87
Data set3	60.42	66.53	71.589
Data set4	74.26	81.57	90.09

Table 2.Classification Accuracy

Table 2 represents classification accuracy values in this table. Proposed (ARFS) values are compared with Existing values of SURF and CART. Their proposed values are higher than compare with other existing values.

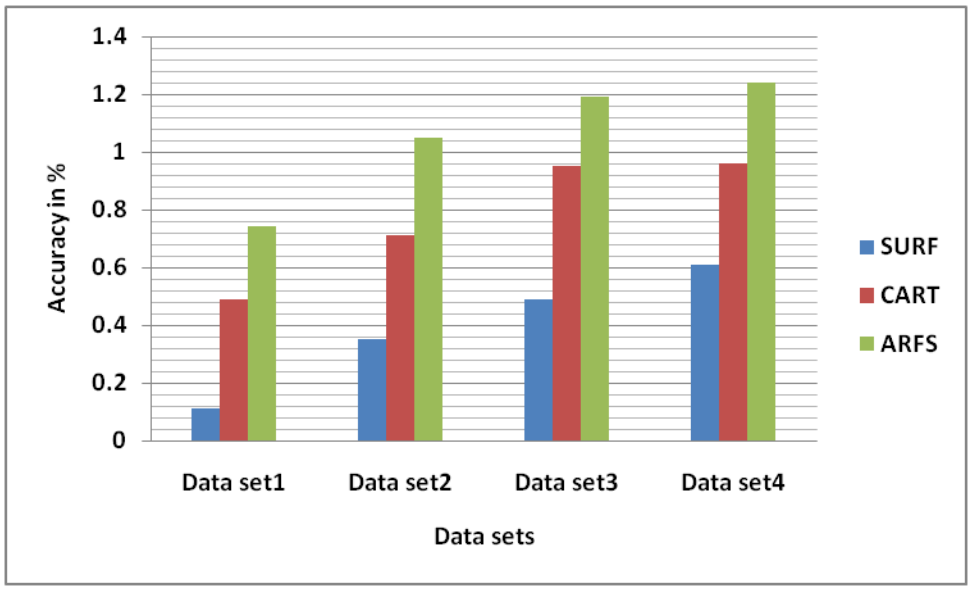


Figure 5. Classification Accuracy

Figure 5 represents the classification accuracy values are compare with them. All values are only positive. The proposed values are higher than in this diagram. Existing 1 is a lower than compare with existing 2 and proposed values.

3. Experimental results on runtime

Data set	SURF	CART	ARFS

Data set1	0.11	0.49	0.74
Data set2	0.35	0.71	1.05
Data set3	0.49	0.95	1.19
Data set4	0.61	0.96	1.24

Table 3.Results on Runtime

Table 3 represents runtime values in this table. Proposed (ARFS) values are compared with Existing values of SURF and CART. Their proposed values are higher than compare with other existing values.

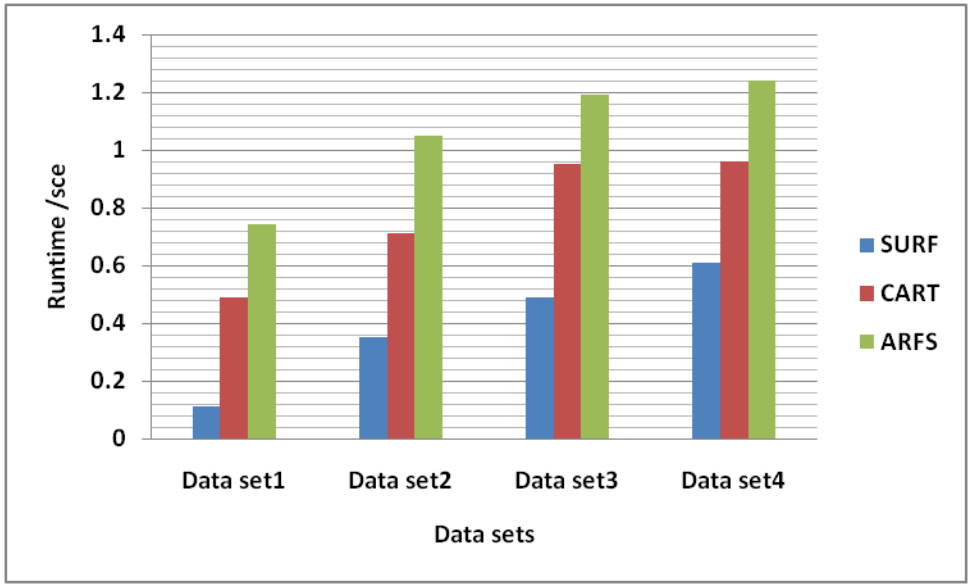


Figure 6. Classification Accuracy

Figure 6 represents the runtime values are compare with them. All values are only positive. The proposed values are higher than in this diagram. Existing 1 is a lower than compare with existing 2 and proposed values.

5. CONCLUSION

The utilization of computers to investigate images has numerous possible applications for mechanized agricultural tasks. However, the fluctuation of the agricultural objects makes it hard

to adjust the current industrial algorithms to the agricultural space. The proposed strategy can measure, examine and recognize fruits dependent on color and texture features. To improve the functionality and adaptability of the recognition system shape and size features can be combined along with color and texture features. Further, by expanding the quantity of images in the database the recognition rate can be expanded. This algorithm can be utilized for keen self help scales.

Reference

1. Akira Mizushima, Renfu Lu. “An image segmentation method for apple sorting and grading using support vector machine and Otsu’s method”.
2. Awate, A., Deshmankar, D., Amrutkar, G., Bagul, U., Sonavane, S.: Fruit disease detection using color, texture analysis and ANN. International Conference on Green Computing and Internet of Things (ICGCIoT), pp. 970–975, 8–10 Oct 2015
3. Chandra Sekhar Nandi, Bipan Tudu, Chiranjib Koley. “An Automated Machine Vision Based System for Fruit Sorting and Grading”.
4. Cihan Akin, Murvet Kirci, Ece Olcay Gunes, Yuksel Cakir “Detection of the Pomegranate Fruits on Tree Using Image Processing”
5. CROPSAP (Horticulture) team of ’E’ pest surveillance: 2013: Pests of Fruits (Banana, Mango and Pomegranate) ’E’ Pest Surveillance and Pest Management Advisory (ed. D.B. Ahuja), jointly published by National Centre for Integrated Pest Management, New Delhi and State Department of Horticulture, Commissionerate of Agriculture, Pune, MS. pp 67.
6. Dah-Jye Lee, James K. Archibald and Guangming Xiong, “Rapid Color Grading for Fruit Quality Evaluation Using Direct Color Mapping”.
7. Dewliya, S., Singh, P.: Detection and classification for apple fruit diseases using support vector machine and chain code. Int. Res.J. Eng. Technol. (IRJET) 02, 04 Aug 2015
8. Dhakate, M., Ingole, A.B.: Diagnosis of pomegranate plant diseases using neural network. In: Fifth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), pp. 1–4, 16–19 Dec 2015
9. Hadha Afrisal, Muhammad Faris, Guntur Utomo P, Lafiona Grezelda, Indah Soesanti, Mochammad Andri F, “Portable Smart Sorting and Grading Machine for Fruits Using Computer Vision”.

10. Ilaria Pertot, Tsvi Kuflik, Igor Gordon, Stanley Freeman, Yigal Elad, Identificator: A web-based tool for visual plant disease identification, a proof of concept with a case study on strawberry, Computers and Electronics in Agriculture, Elsevier, 2012, Vol.88, p.144-154.
11. Lu, J., Wu, P., Xue, J., Qiu, M., Peng, F.: Detecting defects on citrus surface based on circularity threshold segmentation. In: 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), pp. 1543–1547 (2015)
12. Manisha Bhangea, H.A.Hingoliwala, "Smart Farming: Pomegranate Disease Detection Using Image Processing".
13. Monika Jhuria, Ashwani Kumar, Rushikesh Borse, Image Processing For Smart Farming: Detection of Disease and Fruit Grading, IEEE Proceedings of the 2013 IEEE Second International Conference on Image Information Processing, 2013, p.521-526.
14. Mrunmayee Dhakate, Ingole A. B, "Diagnosis of Pomegranate Plant Diseases using Neural Network".
15. Padol, P.B., Yadav, A.A.: SVM classifier based grape leaf disease detection. In: Conference on Advances in Signal Processing (CASP), pp. 175–179, 9–11 June 2016
16. Parag Shinde, Amrita Manjrekar, Efficient Classification of Images using Histogram based Average Distance Computation Algorithm Extended with Duplicate Image Detection Elsevier, proc. Of Int. Conf. On advances in Computer Sciences, AETACS, 2013.
17. R.Gonzalez, R. Woods, Digital Image Processing, 3rd ed., Prentice- Hall, 2007.
18. Revathi, P., Hemalatha, M.: Classification of Cotton Leaf Spot Diseases Using Image Processing Edge Detection Techniques. In: International Conference on Emerging Trends in Science, Engineering and Technology (INCOSET), pp. 169–173, 13–14 Dec 2012
19. S. B. Ullagaddi, Dr. S.Vishwanadha Raju, "A Review of techniques for Automatic detection and diagnose of mango Pathologies".
20. Shiv Ram Dubey, Anand Singh Jalal, Detection and Classification of Apple Fruit Diseases using Complete Local Binary Patterns IEEE, Third international conference on Computer and communication Technology, 2012, p.247-251.
21. Xiaoou Tang, Fang Wen, IntentSearch: Capturing User Intention for One-Click Internet Image Search, IEEE transactions on pattern analysis and machine intelligence, 2012, vol.34, p.1342-1353.