A Spatial Model of K-Nearest Neighbors for Classification of Cotton (Gossypium) Varieties based on Image Segmentation

Tanveer Aslam¹, Salman Qadri * ¹, Najia Saher¹, Syed Ali Nawaz¹, Muhammad Shehzad¹, Syed Furqan Qadri ², Abdul Razzaq³, Faisal Shahzad, ¹ Dost Muhammad Khan.

¹Department of IT, Islamia University of Bahawalpur, 63100 Bahawalpur Punjab, Pakistan.

²School of Computer Science and S.E, Shenzhen University, Shenzhen, China.

³Department of Computer Science MNS-UAM Multan, Punjab Pakistan.

salman.qadri@iub.edu.pk

ABSTRACT

In this study, we describe a technique that used a machine learning (ML) approach to classify four (4) different cotton leaf varieties namely; BS-15, S-32, Z-31, and Z-32. Each variety of cotton leaves were collected from 500 Farmers. These image datasets are captured by using the cell phone camera in the open agricultural field area, and every image was captured from both sides (Front and Back) of the cotton leaf. Each variety of cotton has used over 300 (150 Front Side and 150 Back Side of the leaves) leaf images and the total calculated cotton leaves are 1200 (300 x 4) as leaf image samples. These sample datasets have analyzed through image preprocessing and image segmentation process. Each image was employing four different non-over-lapping regions of interest (ROI's) and calculated a total of 4800 (1200 x 4) ROI's. The acquired datasets are employed different machine learning features such as Scalability, Texture, Spectral, Binary, Histogram, Rotational, and translational (R-S-T). A total of fifty-seven (57) machine learning features were evaluated on each ROI and a total calculated 273,600 (4800 x 57) features. Furthermore, the Correlation-Based Feature Selection (CFS) genetic algorithm technique was employed for feature optimization. It has been evaluated 22 optimized features and applying different machine learning (M-L) classifiers namely; K-Nearest Neighbor (K-NN), K, Random Forest (RF) Tree, and Naive Bayes (NB) Tree. The resulting accuracy produced by K-NN presented is 98.9167% on (512 x 512) ROI's. The individually overall result accuracy dataset values by using K-NN classifier on the four varieties of cotton leaf namely; BS-15, S-32, Z-31, and Z-32 were evaluated 97.83%, 99.50%, 99%, and 99.33%, respectively.*

KEYWORDS: Cotton Leaves, image & classification process, feature optimizing, machine learning process

1. INTRODUCTION

In the field of the agricultural system, machine learning plays an important role in the classification of plant categories [1] [2]. The structure of leaves are the primary or main section of plants and every leaf area is different according to size, shape, texture, color, and structure. This primary section was very helpful for the classification of different plant leaf varieties [3], and cotton leaf is one of them. Nowadays, the identification of cotton (Gossypium) leaves was performed by different researchers in Southern Asia (sub-continental) countries, and Pakistan is one of them. Gossypium is the most important source or crop of vegetable oil, pollen fiber, and produce high-quality protein meals for livestock [4]. Cotton (Gossypium Hirsutum L.) is also called the 'White Gold' and 'Fiber King'. Gossypium is also a most beneficial source for the textile industry, and 60 million (approx.) peoples are engaged with the cotton industry and agricultural system [5]. In Pakistan, cotton (Gossypium) is the second-largest crop cultivated locally and ranked fourth (4th) cultivation

country of Gossypium world-wide. Pakistani Gossypium is highly produced fiber and vegetable oil, which is most beneficial for the oil industry [6].

Gossypium crop is sowing on the area of 2527 (hectares) during 2019-20, and it produces 9.178 billion cotton bales [7]. Gossypium is cultivated among the Sindh and Punjab (Pakistan) provinces [8]. Punjab produces 82% of cotton, and the Sindh production rate is 16% [9]. Cotton (Gossypium Arboreum L.) variety was cultivated in Monjadharo Sindh (Pakistan) in early 6000 B.C [10]. The different genotypes of cotton varieties were cultivating locally such as BS-15, S-32, Z-31, Z-32, CIM-610, CIM-608, Bt.CIM-63 and Bt.CIM-600 etc. These different genotypes of Gossypium are very famous categories and nearly 90% of growers are produced cotton by using the CIM-608 variety of Gossypium in Pakistan [11]. In [12], Pakistan is the biggest exporter of cotton, and it is also earned a million dollars in the fiber and oil industry.

Therefore, we need an automated or artificial

intelligence system that recognized automatically different varieties of *Gossypium* plants, rather than the complete area of the cotton leaf. The open eye or visual-based identification of the cotton leaf is very challenging due to the same features of leaf such as texture, color, geometrical, size, and shape feature between its variant cotton varieties. The confusing arguments about cotton varieties describe major problems such as the high use of pesticides (chemicals) and urea (Fertilizer) without proper study or knowledge and research. That is why the cotton production rate is a little low based on purity and quantity. Here, we discuss three different phases for resolving the following issues

- a) the pre-processing technique was employed on raw image dataset for evaluation of features;
- b) employed different methodologies for feature optimization;
- c) different machine learning classifiers were implemented on available optimized features datasets.

2. LITERATURE REVIEW

Seed quality is the most important element for the cotton cultivation process in the agricultural system. The procedure of seed storage was an essential part to secure seeds from deterioration. Non-Consumer-Seed-Growers (NCSG) and Consumer-Seed-Growers (CSG) were used for seed storage in district Vehaari of Pakistan. CSG techniques are more beneficial for seed storage as compared to NCSG growers. [13]. Cotton-Leaf-Curl-Virus (c-l-c-v) diseases were described in Punjab, Pakistan early 1990s. Four different cotton varieties were used for the experimental process namely; FH-900, CIM-448, FH-901, and CIM-1100. The random Amplified Polymorphic DNA (RAPD) approach was describing the genetic distance as only 3% [14]. The choice of cotton seeds verity was based on geographical environment, crop requirement, and agricultural resources. Two different varieties of cotton are used in the cultivation process such as Zhaongmian-16 and J-4B. The growth of cotton leaf functionally accuracy is 15.9% - 10.7% and 12.3% - 10.5% based on two years, respectively [15].

The cotton (*Gossypium*) leaf verity is affected by Mosaic disease, which is a cause of the reduction in plant growth and genetic characteristics. CRIS-168 variety of cotton is identifying the Mosaic disease and c-l-c-v virus. Mosaic diseases described the different issues on the plant are morphology with following rate such as 25% effected in length, 24.1% in height of the plant and 37.5% of sympodia. The different characteristics are improved slightly inaccuracy of result namely; Uniformity ratio, fiber fineness, and maturity index increase about 20.5%, 14.4%, and 0.9%, respectively [16].

C-L-C-V disease has been the most serious issue of the cotton cultivation process in Pakistan. It is also identified in some areas of plants in Pakistan (1967). Climate and temperature changed were also infected the cotton crops. Fiber index, length, height, fineness, and maturity ratio also affect cotton crop yield and

production. Different methods were employed to reduce C-L-C-V diseases along with fertilizer (urea), control insets, agronomic and biotechnology methods [17]. Leaf varieties classification leads to multiple applications. Many researchers describe enormous techniques for pattern recognition in the leaf of the plant. CNN based techniques were employed on different Indian leaves varieties [18]. Ethylene Insensitive-3 (EIN-3) and Ethylene Insensitive-3 like (EIL) protein values are managed as a major part of plant development and growth during diverse cultivation environment. EIN-3 and EIL proteins are the main part of ethylene signaling regulators and also helped in cotton production stages [19].

In [20], the researchers describe 32 plants by employing among twelve (12) different features of the leaf. A Probabilistic Neural Network (PNN) approach was deploying on the available 1800 sample image dataset, and it's evaluating 90% result accuracy based on PNN. A genetic algorithm was used for the feature optimization process by using apply fruit. Support Vector Machine (SVM) classifier was achieved a classification accuracy is 98.10% [21]. Two different varieties of fused data of 5 variant types of whets were described for the leaf classification process. These datasets were collected from photographic and radiometric information, it is achieved accuracy result is 93.14% and 96%, respectively [22]. The authors [23] introduced a novel based dataset of features, which describe the shape and image color feature. Artificial Neural Network (ANN) approaches were employed on 8 different pepper seed varieties and evaluate 84.94% result accuracy. The different leaves of citrus have been used to describe texture-based features on the color image. Discriminant Analysis (DA) technique deployed for feature optimization, and its observed 95% accuracy based on available features [24].

In this study [25], Deep Learning (DL) technique was used for leaf feature classification. That also describes a new model technique of hybrid, which helps in exploring the hidden information about leaf features. DL approach was improving feature identification in image processing. The shape and texture-based segmentation or object identification techniques were deployed on mango variety detection [26]. The accuracy of the result was compared to an existing one and same technique report in [27], author recognize different method for feature extraction such as binary, histogram and texture. The authors [28] describe, two dimensional (2D) images of the leaf dataset were used for the plant classification process. It's also used two different methods for feature optimization such as one-dimensional (1-D) and two-dimensional (2-D) based on the classifier of Bagging. The five different machine learning (ML) techniques were employed on Flavia's publically colored leaf images dataset for the experimental process namely; Principal Component Analysis (P-C-A), Direct Linear Discriminate Analysis (D-L-D-A), LDA + PCA, 2-DLDA, and 2-DPA. It was evaluated that, the 2-DLDA and 2-DPCA result accuracy much better than all other techniques. In this

study [29], many researchers have adopted different machine vision (M-V) techniques for the varieties identification based on ten different varieties of potatoes. PCA approach was deployed for the feature optimization process. It is observed that the Artificial Neural Network (ANN) was achieved some better accuracy in the result.

It was concluding based on the above literature discussion, huge research or experimental work, it was adopted for the identification of cotton varieties by using different feature techniques. So, there is necessary required a reliable or automatic system for the classification of cotton leaf varieties, that is cost-effective, efficient in work, reliable, and time-saving. The objective of this research or study was to describe the novel technique for the identification of cotton leaves by using optimized features.

3. MATERIALS AND METHOD

3.1. Data Collection

The proposed technique is employed on the Pakistan cotton leaf samples [30]. These leaf datasets were collected from 500 Farmers' cotton land of agricultural and 2 tehsil of District Bahawalpur Punjab (Pakistan) such as Bahawalpur and Ahmed Pur East zone, which is described in Table 1 and Figure 1. The geographical or global position location of cotton was field area in Bahawalpur (29° 23' 44" (North) latitude and 71° 41' 1" (East) longitude) [31] and Ahmad Pur East (29° 5'

13" (North) latitude and 71° 16' 61" (East) longitude).

3.2. Image Capturing and Acquisition

In this experimental process, about 500 growers of cotton field plants with 600 fresh leaves samples have been collected for every single type of cotton. The leaves samples were managed or found in average height and width of 17.5 ± 0.7 cm and 9.05 cm. The sample image acquisition was performing by using a cell phone (OPPO) camera with a resolution of 1080 x 1920 pixels. A cell phone camera has been adjusting vertically at the height of 2 feet on each leaf image, and eliminate shadow leaf. So, all image samples were collected at the open climate and temperature of 40° C (July, 2020) from 1:00 (P.M) to 3:00 (P.M) according to the Pakistani time zone.

All images were captured from the front and backside of 4 different Cotton leaves varieties, and evaluate 600 (150 x 4) front and 600 (150 x 4) backside image of leaf, which is shown in figure 2. A total calculated of 1200 (600 x 2) cotton leaves image datasets were acquired.

3.3. Leaf Area Method

Information about leaf area and Index of Leaf Area (ILA) has been collected from the grower's cotton land and it's calculated by using the following method [32].

$$\text{Area of Leaf (ab}^2\text{)} = M \times N \times K$$

Table 1: Different Area or Zone for Data Collection

Province	District	Area / Tehsil	Union Area or Council (Owner)	Rural / Village	Growers (Farmers)	Cotton Leaves	Total Leaves Sample
Punjab	Bahawalpur	Bahawalpur	Khan Pur (M. Shehzad)	Basti Punraan	45	55	1200 (600 Front Side, 600 Back Side of leaf)
				Basti Chuhan	70	100	
			Noor Pur (Mashood)	Noor Pur Pull	40	50	
				Wahi Jogiyaan	45	60	
		Ahmed Pur East	Mubarak Pur (M. Imran)	Basti Kaloo Khan	80	85	
			Dera Nawab Sahib (Abdullah)	Basti Bhatiyar	60	75	
				Basti Balouch	50	60	
			Sukhail (Rashid Dharala)	Basti Khamisa	40	50	
				Basti Dharala	70	65

Figure 1: (A) Field Area of Crop Image and (B) Geographical or Study Area Map in Pakistan

Figure 2: Gossypium Leaf Variety Image Dataset; (A) BS-15 (B) S-32 (C) Z-31 (D) Z-32

Here, M is the length of the leaf; N is the highest width of the cotton leaf; and K is the value of correction factor of the cotton leaf is 0.75.

Index of Leaf Area (ILA):

$$ILA = \frac{\text{Total Area of Leaf from all Cotton Leaf}}{\text{Field Area Land Where Data Collected}}$$

3.4. Image Preprocessing

In this section, initial processes on images were described in the layout to obtain feature optimization, and classification techniques are very important procedures or ways [33]. Image accuracy or the processing is difficult as it directly expresses our results of feature optimization and classification [34]. So, we describe a better technique during the image processing

step.

Furthermore, all leave images of cotton were arranged in the same shape (Width and Length) by using image resizer software. All the collected images of cotton leaves were resized to resolution 512 x 512 pixels and employed a Median filter to evaluate the prominent area of the cotton leaf

3.5. Selection or Segmentation ROI

In this proposed technique, the most prominent area of the cotton leaf is considered for ROI selection with four (4) non_over_lapping ROIs. The segmentation of cotton leaf is done by using gray_scale_level image, and it's described statistically feature [35]. To classification of the leaf as ROI, the real color is Red, Green, and Blue (RGB) image is exchanged into H_S_V color space structure or model.

Figure 3: Block-Based Diagram for ROI selection Procedure

Three different coordinate values were evaluated such as shade, brightness, and shade [36]. ROI is developing in gray_scale_level describe in algorithm 1, which describes the all necessary steps of ROI selection and segmentation. RGB image conversion into H_S_V or gray_scale_level is shown in Figure 3.

Algorithm 1: Segmentation of Leaf ROI

Phase 1: Input the RGB based image $I_{R_G B}$ of the cotton leaf

Phase 2: RGB leaf image $I_{R_G B}$ is converted into gray_scale_level P_{gray} .

$$P_{gray} = (0.299 * Red) + (0.59 * Green) + (0.114 * Blue)$$

1: Leaf co-ordinate is evaluate by employing gray_scale_level

2: Evaluate the circular shape of image by use equation

$$(a - a_0)^2 + (b - b_0)^2 - M^2 = 0$$

Here, $(a_0 - b_0)$ is the mid-point value of the circle and (a, b) are expected coordinate

Phase 3: Optimize ROI's are evaluate from P_{gray} by employing binary values

$$ROI = I_{R_G B} - \text{transpose}(\text{Binary_Values})$$

Phase 4: H_S_V value is calculated form ROI

$$Red' = Red / 255$$

$$Green' = Green / 255$$

$$Blue' = Blue / 255$$

$$D_{maxi} = \maxi(Red', Green', Blue')$$

$$D_{mini} = \mini(Red', Green', Blue')$$

The value conversion method is given as:

$$\Delta = Delta = D_{maxi} - D_{mini}$$

$$H = \{60 * \left(\frac{Green' - Blue'}{Delta} * MOD2 \right)\}$$

$$H = \{60 * \left(\frac{Blur' - Red'}{Delta} * +2 \right)\}$$

Phase 5:

$$H = \{60 * \left(\frac{Red' - Green'}{Delta} + 4 \right)\}$$

$$S = \{O_{D_{maxi}} = 0\}$$

$$V = D_{maxi}$$

Extracted ROI

4. PROPOSED METHODOLOGY

At the start, the image dataset of cotton leaves are input and employing the preprocessing procedure. Furthermore [37], the Transductive Parameter Transfer (TPT) approach was used for image segmentation and machine learning classification successfully, this approach was evaluated in three phases, which is shown in Algorithm 2.

In Phase 1, the ROI selection procedure is employed for image segmentation. The complete procedure for ROI selection is discussed in Algorithm 1.

In Phase 2, we employed the correlation feature selection (CFS) genetic search algorithm for feature optimization or feature reduction procedure. Phase 3 and finally, these optimized feature datasets were employed different machine learning classifiers for output or result.

5. FEATURES SEGMENTATION OR ACQUISITION

The binary, rotational, spectral, translational (R_S_T), first and second-order (GLCM), and scalability features were deployed on each cotton leaf sample image dataset [38]. These extracted features were based on 28 binary based on 10 pixels (width and height) distance, 6 spectral (3 sectors and 3 rings) features, 7 R_S_T features, 5 first order

(texture), and second-order (GLCM) features [39], [40]. A total of fifty-seven (57) features were extracted on each ROI's, and it's calculated 273600 (57 x 4800) features dataset on available image sample. Computer

Vision and Image Processing (CVIP) tools were deployed for feature extraction.

Algorithm 2: Proposed Transductive Parameter Transfer (TPT) Method	
Input Data:	$A_1^c, \dots, A_n^c, T^x$, the parameter of τ_B, τ_C, ϵ .
Phase 1:	ROI selection procedure, which is describe in Algorithm 1.
Phase 2:	Feature optimization or computation procedure $\theta_j, x_i = 1, 2, \dots, n$
Phase 3:	Evaluate the machine learning (ML) classifiers and calculate the target ML classifiers M^t . $M_j^t = r(x_j, T^x)$ Given M^t , evaluate $\theta^t = \hat{f}(T^x)$
Output:	θ^t (ML Results)

All the statistical feature segmentation was discussed below:

5.1 Histogram or 1ST Order Features

Histogram (1ST order) features are calculated by applying the intensity value on each pixel. Histogram features are also called the 1ST order feature. The probability of the Histogram is L (m), which is shown in equation 1.

$$\text{Histogram} = L(m) = \frac{K(a)}{N} \quad (1)$$

Here, N evaluates the total value of a pixel, and L (m) represents all instances of gray_scale_level value of m. First-order (Histogram) features were calculated five different statistical features namely; Entropy, Mean, Energy, Skewness, and Standard Deviation. Mean is evaluate the dark and brightness value of an image, which is shown in equation 2.

$$\text{Mean} = \bar{s} = \sum_{s=0}^{n-1} sn(s) = \sum_p \sum_q \frac{M(p, q)}{M} \quad (2)$$

Here, n represents the value of gray_scale_level from range 0 to 255, and m and n describe the pixel value of horizontal (rows) and vertical (columns) area. **Standard Deviation (SD)** is evaluating the contrast of the image; this is shown in equation 3.

$$SD = \sigma_n = \sqrt{\sum_{n=0}^{D-1} (n - \bar{n})^2 D(n)} \quad (3)$$

Skewness is describing the center value of symmetry or image, which is shown in equation 4.

$$\text{Skew} = \frac{1}{\sigma_n^3} \sum_{n=0}^{D-1} (n - \bar{n})^3 D(n) \quad (4)$$

Energy is describe the value of gray_scale_level, and is describe in equation 5.

$$\text{Energy} = \sum_{n=0}^{D-1} [D(n)]^2 \quad (5)$$

Entropy is calculating the contents of image data, which is shown in equation 6.

$$\text{Entropy} = - \sum_{n=0}^{D-1} p(n) \log_2[p(n)] \quad (6)$$

5.2 Histogram or 2ND Order (Texture) Features

The 2ND order feature is also called statics texture feature or gray-level co-occurrence matrix (GLCM) based feature. Five different texture features are calculated with 4 dimensional (0°, 45°, 90°, and 135°) areas by using a five-pixel distance. These features are energy, correlation, entropy, inverse-difference, and inertia.

Energy is evaluating the gray_level value in equation 7.

$$\text{Energy} = \sum_a \sum_b (D_{ab})^2 \quad (7)$$

Correlation calculated the similarity between pixels

with pixel distance. It is evaluated in equation 8.

$$\text{Corr} = \frac{1}{\sigma_c \sigma_d} \sum_i \sum_j (i - \mu_c)(j - \mu_d) D_{ij} \quad (8)$$

The total content of the image is described by using **Entropy**, which is shown in equation 9.

$$\text{Entropy} = - \sum_c \sum_d K_{cd} \log_2 K_{cd} \quad (9)$$

The **inverse difference** is also called local homogeneity of image, which is described in equation 10.

$$\text{Inver} - \text{Diff} = \sum_i \sum_j \frac{K_{ij}}{|i - j|} \quad (10)$$

Inertia is evaluating the contrast of image, which is shown in equation 11.

$$\text{Inertia} = \sum_c \sum_d (c - d)^2 K_{cd} \quad (11)$$

5.3 Binary Feature

The binary feature is also called image shape-based feature such as a least second moment of an axis, area, the area of the center, number of Euler, thinness, projection, and aspect of ratio. Twenty Eight (28) binary shaped features were evaluated by using height and width (10, 10) pixels of projection. The area of the j th object was described in equation 12.

$$A_j = \sum_{m=0}^{\text{height}-1} \sum_{n=0}^{\text{width}-1} J_j(m, n) \quad (12)$$

The rows and column coordinates have calculated the centroid of the image by using the j th object, which is shown in equations 13 and 14.

$$\bar{A} = \frac{1}{P_j} \sum_{p=0}^{\text{height}-1} \sum_{q=0}^{\text{width}-1} A P_j(p, q) \quad (13)$$

$$\bar{D} = \frac{1}{P_j} \sum_{p=0}^{\text{height}-1} \sum_{q=0}^{\text{width}-1} D J_j(p, q) \quad (14)$$

In this binary feature also evaluated the orientation, perimeter Euler number and aspect ratio of image.

5.4 Translation, Rotation and Scaling (R_S_T) Features

R_S_T dataset or values are the segmentation of invariant features which are evaluated by employing architectural knowledge of histogram, that is described seven R_S_T invariant segmentation features. Spectral features are based on the frequency domain and it describes successfully when an image is classifying on texture values. Spectral features were evaluated in the different regions and these are based on sector and region, which is defined in equation 15.

$$S.R.P = \sum_{\sigma \in \text{region area}} \sum |M(m, n)|^2 \quad (15)$$

5.5 K- Nearest Neighbors (K-NN) Classifier

Different ML classifiers were available and K-NN is one of them [41]. K-NN algorithm is the simplest and very helpful for the deep learning process [42]. In the K-NN algorithm, an element of the object is identifying by using the maximum number of its nearest or neighbors. The element of the object has continuously evaluated to the class, which was the nearest value between K-NN, here the value of K is positive, which represents a small value. If the value of K is 1 (K=1), then the nearest value of neighbors is assigned to the selected class.

The first employed algorithm of K-NN was evaluating some statistical notation $N = (a_j, b_j)$, $j = 1, 2, 3, 4, \dots, M$ has described the set of training values. Here a_j is describe the dimensional vector of features, and $b_j \in (-1, +1)$ was evaluated with the label of class. Simply we consider this as a classification of binary class. The previous sample labeled as a training value of N. So, the algorithm of K-NN builds a sub-region locally ($D(a) \in R^c$) as input, that is described at the point of a. The estimated region or area D(a) measure the nearest point of training is a, which is shown in equation 16.

$$D(a) = [\hat{a} | P(a, \hat{a}) \leq S_k] \quad (16)$$

Here S_k is the k th is static order of $[D(a, \hat{a})]_1^M$, and $D(a, \hat{a})$ is the metric distance and k is denoted the value of the region in sample D(a). The algorithm of K-NN described the probability of statistics $T(n|a)$, which is shown in equation 17.

$$T(n|a) = \frac{T(a|n)T(n)}{T(a)} = \frac{S_k}{S} \quad (17)$$

For the basis of a, the decision of $h(a)$ is evaluating the data value of S_k and electing the class, which describe the value of the height of S_k , which is shown in equation 18.

$$h(a) = \begin{bmatrix} -1 & S[K=1] \geq S[n=1] \\ 1 & S[K=1] \geq S[n=-1] \end{bmatrix} \quad (18)$$

Therefore the heights probability value is deployed in the algorithm of K-NN. Here, $b_j \in (-1, +1)$ is evaluating the binary problem of classification, which is shown in equation 19.

$$h(a) = nsg \left(\text{aver}_{a_j \in D(a)} b_j \right) \quad (19)$$

6. FEATURE ACQUISITION OR SELECTION

It is perceived that the evaluation of fifty-seven (57) variant features on each ROIs by using cotton leaf sample images. It is not satisfactory of ML classification process, and to manage this 273,600 large amount of features are taken more processing time for model building [43]. So, this problem is resolved by using a feature optimization process as possible, as discussed in [3]. So, that is the easiest way for the classification process with a minimum rate of error. The Principal Component Analysis (P-C-A) technique described good results accuracy on repeated data. But this optimized feature dataset was not providing better accuracy on the whole dataset [44].

Furthermore, we employed the represent correlation-based feature selection (C-F-S) genetic search algorithm to express optimized features from a huge size dataset, which is shown in equation 20

$$M_k = \frac{M\bar{\mu}_{aB}}{\sqrt{M + M(S-1)\bar{\mu}_{aB}}} \quad (20)$$

After employed CFS genetic search on the large size dataset, it evaluated 22 optimized features on each cotton leaf image, which is shown in table 2. It was produced that 273600 (4800 x 57) huge features were minimized into 105,600 (4800 x 22) features.

Finally, different machine learning classifiers are plugged into available optimized features. 10-k fold cross-validation technique has been employed to eliminate the difficulties about the testing and training ratio. The experimental layout on the framework of cotton types is described in figure 4. In this experimental process, there are different ML classifiers were applied to our proposed features dataset such as

K-NN, K*, Random Forest Tree, and Naive Bayes Tree.

It has been evaluated that according to the above discussion classifier evaluates a very low accuracy (below 90%) result on the backside image dataset of the cotton leaf. However, the same classifier set was employed on the front side image dataset of cotton leaf and it is showing a better accuracy result (above 90%).

Table 2: CFS Genetic Algo Based Optimized Features

Sr. No.	Optimized Features
1.	Objects_row_coordinate
2.	Objects_column_coordinate
3.	Centroid (row_column)
4.	Orientation (Axis_of_least_second_moment)
5.	Histro-Standard-Deviation
6.	Histro-Skew
7.	Histro-Entropy
8.	Inertia-range
9.	Inertia-average
10.	Texture-energy-range
11.	Texture-energy-average
12.	Correlation-range
13.	Inverse-diff-average
14.	Inverse-diff-range
15.	Spectral-DC
16.	Texture-entropy-range
17.	Texture-entropy-average
18.	Ring1
19.	Ring2
20.	Ring3
21.	Sector2
22.	Sector3

Figure 4: Feature Dataset Layout or Framework of Cotton varieties classification based on Leaves

7. RESULT AND DISCUSSION

In this proposed study, we employed 4 main machine learning classifiers such as K-Nearest Neighbor (K-NN), K*, Random Forest Tree, and Naive Bayes Tree by using 4 Gossypium varieties with different features dataset. Initially, we employed the 10-k fold cross_validation technique for the classification process. After this, for the creating ROI's of used 60 x 60-pixel size and with image resolution 512 x 512 pixels. Here we discuss some important factors or an element, that's helped us for result accuracy.

1. The acquired dataset of the cotton image is healthy and clear
2. Pre-processing or image accuracy phases and Transductive Parameter Transfer (TPT) method of image segmentation helped us for image clarification
3. A feature optimization algorithm was used for extracted prominent features

In section 6.1, the Backside image dataset of cotton leaf-based employed different machine learning classifiers (ML) were discuss and section 6.2, the Front

side of image dataset of cotton leaf-based employed ML classifiers were discussed.

7.1. Back Side Image Dataset Result

For this proposed experimental process, WEKA 3.6.12 software is employed as a tool and used a different algorithm along with it. The resulting accuracy was described by using 52800 (2400 x 22) total optimized features dataset as input and four different machine learning classifiers were employed for the experimental process such as K-Nearest Neighbor (K-NN), K*, Random Forest Tree, and Naive Bayes Tree, and shown following accuracy result of 94.375%, 93.8333%, 92.875%, and 84.25% respectively. It has been shown that only K-NN achieved better accuracy results, which is shown in Table 3. The individually overall result accuracy of the 4 cotton varieties namely; BS-15, S-32, Z-31 and Z-32 were 93.33%, 95.33%, 92.67% and 96.17% respectively, which is shown in figure 5 and the miss classification result was shown in figure 6. The confusion-matrix (C-M) of K-NN classifier on ROI's image resolution 512 x 512 of four cotton leaves varieties were described in Table 4.

Table 3: Performance of Accuracy Result of Different Machine-Learning Classifiers (Back Side Image of Cotton)

Classifiers	Kappa-Static Value	TP Values	FP Values	R.O.C Values	R-M-S-E Values	Time (Second)	Miss (%) Classify	(%) Accuracy
K-NN	0.925	0.944	0.019	0.963	0.1676	0.03	5.625	94.375%
K*	0.9178	0.938	0.027	0.992	0.1608	0	6.1667	93.8333%
Random Forest	0.905	0.929	0.024	0.987	0.1986	3.25	7.125	92.875%
Naive Bayes	0.79	0.843	0.052	0.939	0.2578	20.55	15.75	84.25%

Table 4: Confusion-Matrix (C-M) of K-NN classifier (Back Side Image of Cotton)

Cotton Classes	BS-15	S-32	Z-31	Z-32	Total Images	(%) Accuracy
BS-15	560	10	27	3	600	93.33
S-32	17	572	2	9	600	95.33
Z-31	32	3	556	9	600	92.67
Z-32	1	7	15	577	600	96.17

Figure 5: Result Accuracy of Cotton Leaves Image classification on Four Varieties (Back Side Image of Cotton)

Figure 6: Miss Classification of Result on K-NN based Classifier (Back Side Image of Cotton) and Color Represents are Blue= BS-15, Red= S-32, Green = Z-31 and Light Blue = Z-32

7.2. Front Side Image Dataset Result

To achieve better accuracy results, the same dataset of machine learning classifiers was also employed on the cotton leaves image dataset with 512 x 512 ROI's. The resulting accuracy is being described by using 52800 (2400 x 22) total optimized features dataset input and four different machine learning classifiers were employed for an experimental process such as K-Nearest Neighbor (K-NN), K*, Random Forest Tree and Naive Bayes Tree, and shown following accuracy result of 98.9167%, 98.2083%, 97.0417%, and 90% respectively. It is showed that only K-NN achieved a better accuracy result, which is shown in Table 5. The individually overall result accuracy of the 4 cotton varieties namely; BS-15, S-32, Z-31, and Z-32 were 97.83%, 99.50%, 99%, and 99.33% respectively, which is shown in figure 7 and the miss classification result was shown in figure 8. The confusion-matrix (C-M) of the K-NN classifier on ROI's image resolution 512 x 512 of four cotton leaves varieties were described in

Table 6.

Finally, the K-NN feature is given better overall accuracy results based on open climate and noisy image dataset, as compared to the following approach [12]. Complete descriptions of the following techniques with existing techniques were shown in Table 7. Our presented system was the ability to discriminate the cotton varieties leaves based on multiple features, which was very beneficial for growers and industrial sectors to accurately identifying the exact variety of cotton. It is an efficient and robust technique to eliminate human error. All the available digital cameras were their resolution and change in the pixel resolution of cameras may become the variation in results. The image dataset used in this experimental process was very less. We positively encourage to researcher and researchers in developing an open or public distribution network for agricultural areas or datasets, where we describe the problems, that is related to the agricultural study.

Table 5: Performance of Accuracy Result of Different Machine-Learning Classifiers (Front Side Image of Cotton)

Classifiers	Kappa-Static Value	TP Values	FP Values	R.O.C Values	R-M-S-E Values	Time (Second)	Miss (%) Classify	(%) Accuracy
K-NN	0.9856	0.989	0.004	0.889	0.0735	0	1.0831	98.9167%
K*	0.9761	0.982	0.006	1.00	0.0846	0	1.7917	98.2083%
Random Forest	0.9606	0.97	0.01	0.998	0.1459	3.04	2.9583	97.041%
Naive Bayes	0.8667	0.9	0.033	0.965	0.2074	18.67	10	90%

Table 6: Confusion-Matrix (C-M) of K-NN classifier (Front Side Image of Cotton)

Cotton Classes	BS-15	S-32	Z-31	Z-32	Total Images	(%) Accuracy
BS-15	587	2	8	3	600	97.83
S-32	0	597	1	2	600	99.50
Z-31	2	2	594	2	600	99
Z-32	2	0	2	596	600	99.33

Figure 7: Result Accuracy of Cotton Leaves Image classification on Four Varieties (Front Side Image of Cotton)

Figure 8: Miss Classification Result on K-NN based Classifier (Front Side Image of Cotton) and Color Represents are Blue= BS-15, Red= S-32, Green = Z-31 and Light Blue = Z-32

Table 7: Comparison Result Accuracy between Proposed and Existing Approach

Source	Features	Algorithm	Results / Overall Accuracy
[2]	ANN, MAPE	CNN, LSTM, DNN	60%
[4]	SVM, LR, PLS-DA	PCA, ResNet, CNN	80%
[5]	MLP, SVM, K-NN	GLCM, HoG, HIST	96%
[3]	MLP, J48, NB	CFS Best Search Algorithm	98%, 97.5%
[18]	SVM, ANN, MLP	KNN ,Robust Algorithm	43.9%, 88%
[12]	NIGAB, ABRI	CRS, MTN, CCRI, APTMA	39.4%
[20]	PNN	PCA, ANN	90.312%
Proposed Approach	K-NN, K*, RF, NB	CFS Genetic Search Algorithm	98.2583%

8. CONCLUSION

In this proposed study, we described the classification of four cotton varieties using multiple features leaves dataset. For this experimental process, four different machine learning techniques were employed successfully namely; K-Nearest Neighbor (K-NN), K*, Random Forest Tree, and Naive Bayes Tree. Different methodologies were applied in the description of the

image dataset for the classification of the cotton leaves. The entire machine learning classifier described the efficient resultant accuracy but the K-NN's resultant accuracy was better as compared to the other machine learning classifiers' results by achieving 98.2583% accuracy on the four cotton varieties. The achieved overall accuracy results evaluate that the proposed technique is efficient and can be employed in real-life applications. In this proposed study, it was evaluated

that the feature optimization technique is more helpful for result accuracy and time-saving procedure. In the featured study, we used the variation among the texture features dataset with the addition of extra attributes to the same cotton leaves.

Acknowledgments: The authors would like to especially thank Dr. Salman Qadri, Ali Nawaz Shah and Muhammad Shehzad for their technical support during this experimental process. Muhammad Abdullah Ismail, Muhammad Mashood, Muhammad Imran and Muddasir Ijaz works as a field assistant for the collection of cotton leaf process.

Funding: "This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors".

Author Contributions: Tanveer Aslam supervised all the fieldwork of data collection, wrote the manuscript, and description of the available dataset. Dr. Salman Qadri supervised this study process, provides all technical support, and analyzed all datasets.

Conflicts of Interest: The authors declare no conflict of interest.

ORCID:

Salman Qadri <https://orcid.org/0000-0002-3503-6535>
TANVEER ASLAM <https://orcid.org/0000-0001-5625-6206>

REFERENCES

- [1] R. Barth, J. IJsselmuiden, J. Hemming and E. V. Henten, "Synthetic bootstrapping of convolutional neural networks for semantic plant part segmentation," *Computers and Electronics in Agriculture*, vol. 161, no. 2019, pp. 291-304, November 2017.
- [2] T. V. Klompenburg, A. Kassahun and C. Catal, "Crop Yield Prediction Using Machine Learning: A Systematic Literature Review," *Computer and Electronic in Agriculture Elsevier*, pp. 1-12, August, 2020.
- [3] S. Qadri, S. F. Qadri, M. Husnain, M. M. S. Missen, D. M. Khan, M.-u.-R. A. Razzaq and S. U. , "Machine Vision Approach for Classification of Citrus Leaves using Fused Features," *International Journal of Food Properties*, pp. 2071-2088, December, 2019.
- [4] S. Zhu, L. Zhou, P. Gao, Y. Bao, Y. He and L. Feng, "Near-Infrared Hyperspectral Imaging Combined with Deep Learning to Identify Cotton Seed Varieties," *Molecules MDPI*, vol. 24, no. 3268, pp. 1-17, September 2019.
- [5] K. Prashar, D. R. Talwar and D. C. Kant, "CNN based on Overlapping Pooling Method and Multi-layered Learning with SVM & KNN for American Cotton Leaf Disease Recognition," *Journal of Engineering Research*, vol. 3, no. 4, pp. 1511-1514, 2014.
- [6] F. Shuli, A. H. Jarwar, X. Wang, L. Wang and Q. Ma, "Overview of the Cotton in Pakistan and its Future Prospects," *Pakistan Journal of Agricultural Research*, vol. 31, no. 4, pp. 396-407, November 2018.
- [7] D. I. Ahmad, M. S. Hanif, M. Asif, M. S. Malik, M. Z. Husnain, M. N. Gul, M. O. Farooq, M. A. Shah, M. N. Mazhar, M. S. Khatoon, M. A. Nawaz, M. T. Islam, H. S. M. Azeem, M. R. Akbar, M. U. Zahid, M. S. Saddiq, M. A. M. S. Khan, M. S. Ameer, M. F. Anwar and M. F. Shamim, "Pakistan Economic Survey 2019-20," Islamabad, 2020.
- [8] W. Wei, Z. Mushtaq, A. Ikram, M. Faisal, Z. Wan-Li and M. I. Ahmad, "Estimating the Economic Viability of Cotton Growers in Punjab Province, Pakistan," *SAGE Open Original Research*, pp. 1-12, June, 2020.
- [9] S. Ashraf, A. H. Sangi, Z. Y. Hassan and M. Luqman, "Future of Cotton Sector in Pakistan: A 2025 Outlook," *Pakistan Journal of Agricultural Research*, vol. 31, no. 2, pp. 145-150, June 2018.
- [10] M. Ur-Rehman, Z. Rahmat, A. Mahmood, K. Abdullah and Y. Zafar, "Cotton Germplasm of Pakistan," *World Cotton Germplasm Resources INTECH*, pp. 137-166, June 2014.
- [11] Z. Anjum, K. H. S. C. T. Azhar, U. S. F. A. M. T. H. Mehmood and M. A. , "Development of Cotton Leaf Curls Virus Tolerance Varieties Through Interspecific Hybridization," *African Journal of Agricultural Research*, vol. 10, no. 13, pp. 1612-1618, March 2015.
- [12] Y. M. A. A. M. S. A. W. Ali L., A. B. K. A. and A. Q. , "Evaluation of Release of New Cotton Variety "RH-668"," *International Journal of Biology, Pharmacy and Allied Science (IJBPAS)*, vol. 9, no. 2, pp. 270-281, February 2020.

- [13] M. F. Saeed, A. Jamal, I. Ahmad, S. Ali, G. M. Shah, S. K. Husnain, A. Farooq and J. Wang, "Storage Condition Deteriorate Cotton and Wheat Seed Quality: An Assessment of Farmers Awareness in Pakistan," *Agronomy MDPI*, vol. 10, no. 1246, pp. 1-19, August 2020.
- [14] I. Ullah, M. Nawaz, N. Akram, M. Younas, M. Z. Iqbal and H. L. Shaheen, "PHYLOGENETIC ANALYSIS OF FIRST GENERATION COTTON LEAF CURL VIRUS (CLCuV) RESISTANT COTTON VARIETIES," *Biotechnol. & Biotechnol. eq. 2011 (A&EB)*, vol. 25, no. 4, pp. 2610-2612, April 2011.
- [15] A. Khan, J. Zheng, D. K. Yuen Tan, A. Khan, K. Akhtar, X. Kong, F. Munsif, A. Iqbal, M. Z. Afridi, A. U. S. Fahad and R. Zhou, "Changes in Leaf Structural and Functional Characteristics when Changing Planting Density at Di under a New Planting Model," *Agronomy MDPI*, vol. 9, no. 859, pp. 1-15, December 2019.
- [16] K. P. Akhtar, M. H. W. Ishaque, M. R. Khan and A. I. Khan, "Effect of Cotton Leaf Mosaic Disease on Morphology, Yield and Fibre Characteristics of Upland Cotton in Pakistan," *The Plant Pathology Journal*, vol. 21, no. 2, pp. 137-141, February 2005.
- [17] J. F. A. F. M. R. M. S. F. S. M. S. Iqbal, T. A. B. and A. M., "Cotton Leaf Curl Virus Disease a Principle Cause of Decline in Cotton Productivity in Pakistan (A Mini Review)," *Candian Journal of Plant Protection*, vol. 2, no. 1, pp. 9-16, 2014.
- [18] M. V. and P. R., "CNN Approaches for Classification of Indian Leaf Species Using Smartphones," *Computer, Materials & Continua Techscience Journal*, vol. 62, no. 3, pp. 1445-1472, 2020.
- [19] H. Salih, S. He, H. Li, Z. Peng and X. Du, "Investigation of the EIL/EIN3 Transcription Factor Gene Family Members and Their Expression Levels in the Early Stage of Cotton Fiber Development," *Plants MDPI*, vol. 9, no. 128, pp. 1-16, January 2020.
- [20] S. G. Wu, F. S. Bao, E. Y. Xu, Y.-X. Wang, Y.-F. Chang and Q.-L. Xiang, "A Leaf Recognition Algorithm for Plant Classification Using Probabilistic Neural Network," *Agricultural Journal*, July 2007.
- [21] M. A. Khan, M. I. LaLI, M. Sharif, K. Javed, K. AURANGZEB, S. I. HAIDER, A. S. ALTAMRAH and T. AKRAM, "An Optimized Method for Segmentation and Classification of Apple Diseases based on Strong Correlation and Genetic Algorithm based Feature Selection," *igital Object Identifier IEEE Access*, vol. 1, pp. 1-18, 2019.
- [22] S. Qadri, M. D. Khan, F. Ahmad, S. F. Qadri, M. E. Babar, M. Shahid, M. Ul-Rehman, A. Razzaq, S. S. Muhammad, M. Fahad, S. Ahmad, M. T. Pervez, N. Naveed, N. Aslam, M. Jamil, E. A. Rehmani, N. Ahmad and N. A. Khan, "A Comparative Study of Land Cover Classification by Using Multispectral and Texture Data," *BioMed Research International Hindwai*, pp. 1-12, April 2016.
- [23] F. Kurtulmuş, İ. Alibaş and I. Kavdir, "Classification of pepper seeds using machine vision based on neural network," *International Journal of Agricultural and Biological Engineering*, vol. 9, no. 1, pp. 51-61, January 2016.
- [24] M. Clerc and S. Mallat, "The Texture Gradient Equation for Recovering Shape from Texture," *IEEE TRANSACTION ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, vol. 24, no. 4, pp. 536-549, April 2002.
- [25] S. H. Lee, C. S. Chan, S. J. Mayo and P. Remagnino, "How deeplearning extracts and learns leaf features for plant Classification," *Pattern Recognition Elsevier*, vol. 71, pp. 1-13, May 2017.
- [26] N. Turner, A. Hearn, J. Begg and G. Constable, "COTTON (GOSSYPIMUM HI-RSUTUM L.): PHYSIOLOGICAL AND MORPHOLOGICAL RESPONSES TO WATER DEFICITS AND THEIR RELATIONSHIP TO YIELD," *Field Crops Research Elsevier*, pp. 153-170, 1986.
- [27] T. Aslam, S. Qadri, M. Shehzad, S. F. Qadri,

- A. Razzaq, S. S. Muhammad, S. Ahmad, S. A. Nawaz and N. Ahmad, "Emotion Based Facial Expression Detection Using Machine Learning," *Life Science Journal*, vol. 17, no. 8, pp. 35-43, August 2020.
- [28] A. Tharwat, T. Gaber and A. E. Hassanien, *University of Salford Manchester*, pp. 1-31, April 2016.
- [29] A. Azizi and Y. Abbaspour-Gilandeh, "Identifying Irregular Potatoes by Developing an Intelligent Algorithm Based on Image Processing," *Journal of Agricultural Sciences*, vol. 22, pp. 32-41, December 2014.
- [30] A. Rehman, L. Jingdong, A. A. Chandio, I. Hussain, S. A. Wagan and Q. U. A. Memon, "Economic perspectives of cotton crop in Pakistan: A time series analysis (1970–2015) (Part 1)," *Journal of the Saudi Society of Agricultural Sciences*, pp. 1-7, December 2016.
- [31] T. Aslam and S. Qadri, "The Islamia University of Bahawalpur," December 2020. [Online]. Available: <https://www.iub.edu.pk/>.
- [32] V. J. Pascual and Y.-M. Wang, "Impact of Water Management on Rice Varieties, Yield, and Water Productivity under the System of Rice Intensification in Southern Taiwan," *Water MDPI*, pp. 1-15, December, 2016.
- [33] F. Khan, M. A. Khan, S. Abbas, A. Athar, S. Y. Siddiqui, A. H. Khan, M. A. Saeed and M. Hussain, "Cloud-Based Breast Cancer Prediction Empowered with Soft Computing Approaches," *Journal of Health Care Engineering Hindwai*, pp. 1-16, May, 2020.
- [34] R. H. Ajaz and L. Hussain, "Seed Classification using Machine Learning Techniques," *Journal of Multidisciplinary Engineering Science and Technology (JMEST)*, vol. 2, no. 5, pp. 1098-1102, May 2015.
- [35] P. Chaudhary, A. K. Chaudhari, D. A. N. Cheeran and S. Godara, "Color Transform Based Approach for Disease Spot Detection on Plant Leaf," *International Journal of Computer Science and Telecommunications*, vol. 3, no. 6, pp. 65-70, June 2012.
- [36] N. Sengar, M. K. Dutta and B. Sarkar, "Computer vision based technique for identification of fish quality after pesticide exposure," *International Journal of Food Properties Taylor & Francis*, pp. 2192-2206, August, 2017.
- [37] G. Zen, L. Porzi, E. Sangineto, E. Ricci and N. Sebe, "Learning Personalized Models for Facial Expression Analysis and Gesture Recognition," *IEEE TRANSACTIONS ON MULTIMEDIA*, pp. 775-788, April, 2016.
- [38] A. Humeau-Heurtier, "Texture Feature Extraction Methods: A Survey," *IEEE Digital Object Identifier*, vol. 7, no. 2019, pp. 8975-8986, January 2019.
- [39] G. M. M. E Elahi, "Gradient Based Novel Texture Feature Extraction Methods For Texture Classification," 2016.
- [40] D. Kaur and Y. Kaur, "Various Image Segmentation Techniques: A Review," *International Journal of Computer Science and Mobile Computing*, vol. 3, no. 5, pp. 809-814, May 2014.
- [41] G. Guo, H. Wang, D. Bell, Y. Bi and K. Greer, "KNN Model-Based Approach in Classification," in *Lecture Notes in Computer Science*, 2003.
- [42] C. Li, S. Zhang, H. Zhang, L. Pang, K. Lam, C. Hui and S. Zhang, "Using the K-Nearest Neighbor Algorithm for the Classification of LymphNodeMetastasis inGastric Cancer," *Computational and Mathematical Methods in Medicine*, vol. 2012, pp. 1-12, September 2012.
- [43] G. Chandrashekar and F. Sahin, "A Survey on Feature Selection Methods," *Computer and Electrical Engineering Elsevier*, vol. 40, pp. 16-28, December 2013.
- [44] M. Shehzad, S. Qadri, T. Aslam, S. F. Qadri, A. Razzaq, S. S. Muhammad, S. Ahmad, S. A. Nawaz and N. Ahmad, "Machine Vision Based Identification of Eye Cataract Stages Using Texture Features," *Life Science Journal*, pp. 44-50, August, 2020.