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 Razzaq1, Faisal Shahzad, 1Dost 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.

salmqadri@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-overlapping 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-LS) 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 Monjadraro 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 Zhaogmian-16 and J-4B. The growth of cotton leaf functionality with following rate such as 25% affected in length, 24.1% in height of the plant and 37.5% of sympodia. The different varieties are affected significantly in the result namely; Uniformity ratio, fiber fineness, and maturity index increase about 20.5%, 14.4%, and 0.9%, 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% affected in length, 24.1% in height of the plant and 37.5% of sympodia. The different characteristics are improved slightly inaccurate 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 insects, 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 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 performed 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].

\[
\text{Area of Leaf} = M \times N \times K
\]

<table>
<thead>
<tr>
<th>Province</th>
<th>District</th>
<th>Area / Tehsil</th>
<th>Union Area or Council (Owner)</th>
<th>Rural / Village</th>
<th>Growers (Farmers)</th>
<th>Cotton Leaves</th>
<th>Total Leaves Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Punjab</td>
<td>Bahawalpur</td>
<td>Khan Pur</td>
<td>Basti Punraan</td>
<td>45</td>
<td>55</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(M. Shehzad)</td>
<td>Basti Chuhan</td>
<td>70</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Noor Pur</td>
<td>Noor Pur Pull</td>
<td>40</td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Mashood)</td>
<td>Wahi Jogiyaan</td>
<td>45</td>
<td>60</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mubarak Pur</td>
<td>Basti Kaloo Khan</td>
<td>80</td>
<td>85</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(M. Imran)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dera Nawab</td>
<td>Basti Bhattiyani</td>
<td>60</td>
<td>75</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sahib (Abdullah)</td>
<td>Basti Balouch</td>
<td>50</td>
<td>60</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sukhail</td>
<td>Basti Khamisa</td>
<td>40</td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Rashid Dharala)</td>
<td>Basti Dharala</td>
<td>70</td>
<td>65</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Different Area or Zone for Data Collection
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_overlapping 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.
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_{RGB} of the cotton leaf

**Phase 2:** RGB leaf image I_{RGB} is converted into gray_scale_level P_{gray}

\[
P_{gray} = (0.299 \times \text{Red}) + (0.587 \times \text{Green}) + (0.114 \times \text{Blue})
\]

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_{RGB} - transpose (Binary_Values)

**Phase 4:** H_S_V value is calculated form ROI

\[
\text{Red}' = \frac{\text{Red}}{255}\\
\text{Green}' = \frac{\text{Green}}{255}\\
\text{Blue}' = \frac{\text{Blue}}{255}\\
\]

\[
D_{max} = \max (\text{Red}', \text{Green}', \text{Blue}')\\
D_{min} = \min (\text{Red}', \text{Green}', \text{Blue}')\\
\]

The value conversion method is given as:

\[
\Delta = \Delta = D_{max} - D_{min}\\
H = \{60 \times \left(\frac{\text{Green}' - \text{Blue}'}{\Delta} \mod 2\right)\}
\]

### Phase 5:

\[
H = \{60 \times \left(\frac{\text{Blue}' - \text{Red}'}{\Delta} + 2\right)\}
\]

S = \{ O \ D_{max} = 0 \} 

V = D_{max}

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 2735600 (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**

<table>
<thead>
<tr>
<th>Input</th>
<th>(A^t_j, \ldots, A^t_n, T^s), the parameter of (\tau_\theta, \tau_r, \epsilon).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data:</td>
<td>(\theta_j, x_i = 1, 2, \ldots, n)</td>
</tr>
<tr>
<td>Phase 1:</td>
<td>Feature selection procedure</td>
</tr>
<tr>
<td>Phase 2:</td>
<td>Feature optimization or computation procedure</td>
</tr>
<tr>
<td>Phase 3:</td>
<td>Evaluate the machine learning (ML) classifiers and calculate the target ML classifiers (M^t).</td>
</tr>
<tr>
<td>Given (M^t), evaluate (\theta^t = \hat{f}(T^s))</td>
<td></td>
</tr>
<tr>
<td>Output:</td>
<td>(\theta^t) (ML Results)</td>
</tr>
</tbody>
</table>

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.

\[
H_{\text{istogram}} = L(m) = \frac{K(o)}{N} \tag{1}
\]

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 evaluating the dark and brightness value of an image, which is shown in equation 2.

\[
\text{Mean} = \bar{x} = \frac{1}{N} \sum_{n=0}^{N-1} sn(s) = \sum_{p} \sum_{q} \frac{M(p,q)}{M} \tag{2}
\]

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.

\[
\text{SD} = \sigma_n = \sqrt{\sum_{n=0}^{N-1} (n - \bar{n})^2 D(n)} \tag{3}
\]

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

\[
\text{Skew} = \frac{1}{\sigma_n^3} \sum_{n=0}^{N-1} (n - \bar{n})^3 D(n) \tag{4}
\]

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

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

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

\[
\text{Entropy} = -\sum_{n=0}^{N-1} p(n)\log_2[p(n)] \tag{6}
\]

**5.2 Histogram or 2ND Order (Texture) Features**

The 2ND order feature is also called static 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.

\[
\text{Energy} = \sum_{a} \sum_{b} (D_{ab})^2 \tag{7}
\]

Correlation calculated the similarity between pixels
with pixel distance. It is evaluated in equation 8.

$$\text{Corr} = \frac{1}{\sigma_c\sigma_d} \sum_{i} \sum_{j} (i - \mu_i)(j - \mu_j)D_{ij}$$  \hspace{1cm} (8)

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

$$\text{Entropy} = -\sum_{c} \sum_{d} K_{cd}\log_{2}K_{cd}$$  \hspace{1cm} (9)

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

$$\text{Inver-Diff} = \sum_{i} \sum_{j} \frac{K_{ij}}{|i-j|}$$  \hspace{1cm} (10)

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

$$\text{Inertia} = \sum_{c} \sum_{d} (c-d)^2 K_{cd}$$  \hspace{1cm} (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 jth object was described in equation 12.

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

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

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

$$\bar{D} = \frac{1}{P_j} \sum_{p=0}^{\text{height-1}} \sum_{q=0}^{\text{width-1}} D_{ij}(p, q)$$  \hspace{1cm} (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 feature 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.

$$S_{R.P} = \sum_{a \in \text{region area}} \sum_{m} |M(m, n)|^2$$  \hspace{1cm} (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,..., M) has described the set of training values. Here a_j is describe the dimensional vector of features, and b_j ∈ (-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) ∈ R^2) 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.

$$D(a) = |\bar{a} - P(a, \bar{a})| \leq S_k$$  \hspace{1cm} (16)

Here S_k is the kth is static order of [D(a, \bar{a})] M, and [D(a, \bar{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.

$$T(n|a) = \frac{T(a|n)T(n)}{T(a)} = \frac{S_k}{S}$$  \hspace{1cm} (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.

$$h(a) = \begin{cases} 1 & S_k [K = 1] \geq S[n = 1] \\ 1 & S_k [K = 1] \geq S[n = -1] \end{cases}$$  \hspace{1cm} (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.

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

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

$$M_k = \frac{M_{k} \mathbf{a}_{B}}{\sqrt{M + M (S-1) \mathbf{a}_{B}}} \quad (20)$$

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

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

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

<table>
<thead>
<tr>
<th>Cotton Classes</th>
<th>BS-15</th>
<th>S-32</th>
<th>Z-31</th>
<th>Z-32</th>
<th>Total Images</th>
<th>(%) Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>BS-15</td>
<td>560</td>
<td>10</td>
<td>27</td>
<td>3</td>
<td>600</td>
<td>93.33</td>
</tr>
<tr>
<td>S-32</td>
<td>17</td>
<td>572</td>
<td>2</td>
<td>9</td>
<td>600</td>
<td>95.33</td>
</tr>
<tr>
<td>Z-31</td>
<td>32</td>
<td>3</td>
<td>556</td>
<td>9</td>
<td>600</td>
<td>92.67</td>
</tr>
<tr>
<td>Z-32</td>
<td>1</td>
<td>7</td>
<td>15</td>
<td>577</td>
<td>600</td>
<td>96.17</td>
</tr>
</tbody>
</table>

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)

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

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

<table>
<thead>
<tr>
<th>Cotton Classes</th>
<th>BS-15</th>
<th>S-32</th>
<th>Z-31</th>
<th>Z-32</th>
<th>Total Images</th>
<th>(%) Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>BS-15</td>
<td>587</td>
<td>2</td>
<td>8</td>
<td>3</td>
<td>600</td>
<td>97.83</td>
</tr>
<tr>
<td>S-32</td>
<td>597</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>600</td>
<td>99.50</td>
</tr>
<tr>
<td>Z-31</td>
<td>594</td>
<td>2</td>
<td>2</td>
<td>600</td>
<td>99</td>
<td></td>
</tr>
<tr>
<td>Z-32</td>
<td>596</td>
<td>2</td>
<td>2</td>
<td>600</td>
<td>99.33</td>
<td></td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Source</th>
<th>Features</th>
<th>Algorithm</th>
<th>Results / Overall Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>[2]</td>
<td>ANN, MAPE</td>
<td>CNN, LSTM, DNN</td>
<td>60%</td>
</tr>
<tr>
<td>[3]</td>
<td>MLP, J48, NB</td>
<td>CFS Best Search Algorithm</td>
<td>98%, 97.5%</td>
</tr>
<tr>
<td>[18]</td>
<td>SVM, ANN, MLP</td>
<td>KNN ,Robust Algorithm</td>
<td>43.9%, 88%</td>
</tr>
<tr>
<td>[12]</td>
<td>NIGAB, ABRI</td>
<td>CRS, MTN, CCRI, APTMA</td>
<td>39.4%</td>
</tr>
<tr>
<td>[20]</td>
<td>PNN</td>
<td>PCA, ANN</td>
<td>90.312%</td>
</tr>
<tr>
<td>Proposed Approach</td>
<td>K-NN, K*, RF, NB</td>
<td>CFS Genetic Search Algorithm</td>
<td>98.2583%</td>
</tr>
</tbody>
</table>
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, 

Catal, "Crop Yield Prediction Using Machine 
Learning: A Systematic Literature Review," 
Computer and Electronic in Agriculture 

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 

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 

and Q. Ma, "Overview of the Cotton in 
Pakistan and its FutureProspects," Pakistan 
Journal of Agricultural Research, vol. 31, 
no. 4, pp. 396-407, November 2018.

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 

Wan-Li and M. I. Ahmad, "Estimating the 
Economic Viability of Cotton Growers in 
Punjab Province, Pakistan," SAGE Open 

Luqman, "Future of Cotton Sector in 
Pakistan: A 2025 Outlook," Pakistan Journal 
145-150, June 2018.

Abdullah and Y. Zafar, "Cotton Germplasm 
of Pakistan," World Cotton Germplasm 
Resources INTECH, pp. 137-166, June 2014.

M. T. H. Mehmood and M. A. , 
"Development of Cotton Leaf Curls Virus 
Tolerance Varieties Through Interspecific 
Hybridization," African Journal of 
1612-1618, March 2015.

and A. Q. , "Evaluation of Release of New 
Cotton Variety "RH-668"," International 
Journal of Biology, Pharmacy and Allied 
Science (IJBAPAS), vol. 9, no. 2, pp. 270-281, 
February 2020.


