MAIZE SEED CLASSIFICATION WITH MACHINE LEARNING APPROACH

Sudesh Kumar Mittal  
Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India, sudesh.mittal@chitkara.edu.in  
Ridhi Jindal  
Research Scholar, Rayat Bahra University, Mohali, India  
ridhijindal136@gmail.com

ABSTRACT
Sorting seeds is important in agriculture for both productivity and commercial reasons. Low-quality seeds may result in poor plant development, disease, and poor crop yields. A quick identification and classification approach for maize seeds is developed in this work using machine learning and machine vision. The traditional experiments’ knowledge cannot reduce agriculture’s present adverse effects on the biosphere. There is an increase in the gap between the slowly expanding knowledge base and the adverse impacts of the environment. Crop seed quality can be significantly indicated by the seed purity. Also, in the modern agricultural industry, maize is a significant crop which is having worldwide production greater than 40%. All over the world, maize is among the significantly cultivated grains. In the context of genetic programs, modern crop improvement and advanced maize breeding, the most significant technique is doubled-haploid as with the help of this technique breeding efficiency is increased as well as breeding period is shortened. The study aims in examining the ML (machine learning) approach’ feasibility in various types of maize seeds’ classification. The digital images (DI) of seeds of 6 maize varieties were ICI 339, Neelam Makkai, Pioneer P-1429, Kashmiri Makkai, Sygenta ST-6142, and Desi Makkai. The classification outcomes meet the demands of both producers and consumers.

Keywords: Seed, Environment, Machine learning, Agriculture, Virtual Reality, Quality, Random Forest, Confusion Matrix.

1. INTRODUCTION
Annual variability and climate change greatly affect the agriculture sector, specifically impacting agricultural production. Most commonly grown crops in the growing season have up to 30% annual variations caused by meteorological conditions that include alterations in temperature and precipitation variables [1]. Implementing genotype analysis and classification, trait determination, and the breeding of new varieties is one of the crucial steps in agricultural production [19]. This may enhance a plant's capacity to endure stress, which is essential for its development and growth. Also, there exist various other factors like socio-economic factors, topography (aspect, slope, and elevation), and soil conditions that affect crop yields. In agriculture for better decision-making and planning detailed crop yield prediction approaches are required by farmers and other decision-makers [12]. Determining seed purity is now a major problem for the market and farmers; this is significant since seed quality affects maize diseases.
and output. Additionally, since maize types are planted near to one another, it is simple for them to get mixed up unintentionally during harvest. Therefore, when maize seed obtained from group plantings is put into the market without being categorized, the market value considerably decreases [20]. Seeds are a key component in crop production from the viewpoint of sustainable agricultural development, and seed technology is important to agriculture [21]. Particularly, farmers with the help of crop yield predictions can decide on scheduling as well as planning seasonal crops, and an event’s future outcomes can be possibly determined. [1]

Artificial neural networks (ANN), expert systems, simulation, and regression are considered the methods used for yield prediction. Specifically for prediction purposes, several studies widely use regression models. The reason for using these is the fact that they usually produce reliable standard tests as well as can be easily used [13]. Furthermore, in some complex cases like non-linear relationships and extreme data values, regression models’ use is limited. Interrelated factors’ diversity which influences the production of crops hardly describes their relationship through traditional approaches [14]

Among the various important agricultural products, maize (Zea mays L.) is mostly utilized as industrial raw materials, animal feed and human food (Cerit et al., 2016). Due to climate change as well as the increasing world population, it is required that new maize varieties must be developed that cannot be affected by abiotic and biotic stress conditions as well as are high-yielding similar to other cultivated plants. Only with the help of maize breeding programs, these goals can be achieved. [3]

There exist two types of maize seeds: open-pollinated and hybrid. The plants produced from open-pollinated maize seeds further can produce more plants with their seeds which are analogous to the parent maize plant. Generally, pollinated plants generate these seeds. Such plants are planted due to the reason that these seeds have better taste. Also, they are generally utilized for feeding animals [2]. Furthermore, hybrid maize seeds are a highly specialized expense and unique crop. In comparison to other cereal crops, more experience, expense, and time is required for hybrid maize seed production. This seed production consists of 2 inbred lines crossing which intersect for producing a different variety of seeds through hybridization. The process uses “male” plants which are 2 inbred lines which are responsible for pollen production as well as “female” plants which generate the hybrid seeds. Also, for a new variety, the inbred lines are crossed which signifies some qualities like dry season protection. For ensuring the seed’s purity as well as quality, throughout the process various measures are acquired. [5]

This study focused on introducing a framework for hybrid features classification, for maize seed varieties classification with the help of hybrid features. It comprises 5 phases: (1) digital image dataset of maize seed are preprocessed for removing noise; (2) eliminating image background and extracting the region of interest threshold-based segmentation is utilized; (3) extracting hybrid texture features or statistical information; (4) for obtaining optimized features
as well as removing an extra feature of not worth, hybrid feature dataset is fused, and (5) finally, high classification accuracy is obtained by deploying the optimized hybrid feature dataset on ML classifiers.

2. LITERATURE REVIEW
A system has been proposed by Zhang et al., which is used for maize seed verities’ classification and utilizes a remote sensing dataset along with an RF classifier. For capturing the hybrid information, texture features and spectrum are utilized as well as it results in an accuracy of 97.65%. [10]
For damage maize seeds’ classification, a hyperspectral imaging framework has been represented by Zhang et al. Several image processing approaches were used by them which resulted in achieving 90% accuracy as the mean spectrum classifier is implemented with mean spectrum. [11]

The forecast of maize yield in Jilin, China is based on climate data and fertilizer [6]. The authors stated that the estimated yield was closely related to the observed yield. Although the application of ANNs to maize output has proved accurate, some researchers utilized such models to forecast maize output in South Africa. Many of the studies currently underway have depended on the use of crop modelling which is most costly and data intensive. The purpose of this thesis is to establish an ANN for maize development in South Africa's major corn production regions (FS, NW, MP and KZN).

The two pathways to ML were discussed by Folberth et al. (2019). ML has been conditioned on the GGCM (Global Gridded Crop Model) global maize simulation. The system makes high-precision estimates (R2>0.96) for the yield of maize, evapotranspiration as well as using water for crop production.

The ANN is known as the best way to derive information from non-linear and imprecise data, according to M. Caselli et al. [7] ANN techniques have proved very important tools in many disciplines for a wide range of applications, like crop production projections. They were then used to forecast maize yield on the basis of soil and weather data with differing performance rates [8, 9].

O’Neal et al. [4], in east-central Indiana, USA, envisaged maize yields at 3 scales with yield data spanning and local crop-stage weather from 1901-1996 utilizing regression models with a completely connected back-propagation ANN.

3. MATERIAL AND METHODS
Maize seed’s 6 varieties were gathered that includes hybrid maize seed (Pioneer P-1429, ICI339, and Sygenta ST-6142) and open-pollinated (Kashmiri Makkai, Neelam Makkai, and Desi Makkai). For the experiment, every variety is equally weighed and 30g is utilized. Further with the help of a DSLR camera above-mentioned maize seed varieties’ digital images are
acquired. Figure 1 presents the images of 6 different varieties of maize seeds. A machine vision system was utilized to examine maize seeds of all types to understand them more intuitively.

**MACHINE LEARNING SCHEMES**

The goal of machine learning is to develop and use mathematical "data structures" that enable a computer to behave in ways that would typically need a human. To develop a model utilizing training data and categorize the maize seeds, the following ML classifiers are utilized in classification problems.

**A Multilayer Perceptron**

A feedforward ANN model called a multilayer perceptron transfers a set of input data into sets of suitable output. The input and output layers of the MLP each include one or more hidden layers of nonlinearly activating nodes. Each node in a layer links to each node in the layer below with a specific weight $w_{ij}$. Universal approximators are MLPs.

**B Random Forest**

An integrated method of the Bagging type is called random forest (RF). Voting is used to combine numerous weak classifiers, which increases the accuracy and generalizability of the final model outcome. While there are several samples and features, random forests perform well when processing high-dimensional data. A multi-layer data cube was created in this work using many multi-temporal single-band VI image series to facilitate further analysis.

**C LogitBoost (LB)**

A modified form of boosted decision trees for classification is called LogitBoost (LB). Boosting is a different strategy from bagging for enhancing the predictions made by decision trees. The difference between boosting and RFs is that boosting grows the trees sequentially by utilizing data from earlier grown trees, while RFs use a bagging approach to replicate the original training dataset with the bootstrap (weak learners). The LB classifier employs one-node decision trees (decision stumps) as weak learners and applies a feature pre-selection technique similar to MR, with a score that is equal to the test statistic of the Wilcoxon two-sample test.

![Image of maize seeds]
Figure 1: Six varieties of maize seed

Table 1 presents the sunlight intensity’s recorded value as well as the time when maize seeds’ images were captured i.e., recorded time. Results can be affected by the different sizes’ of captured images. Thus, resizing every image is performed. Therefore, every image is cropped to a similar size and then “CVIPtools version 5.7e” an image processing software [15] is used for converting these images into Gray-Level (GL) images as represented in figure 2 as well as an image is acquired that shows the desired object. The study involves maize seed’s 6 different varieties for classification. Models help in categorizing characteristics into different classes. The classification method uses a variety of algorithms, and the classification outcomes on the dataset are varied.

**Image Processing**

This process takes place before the images are analyzed to feature extraction. It directly affects the classification outcomes. Figure 2 illustrates the image preprocessing procedure, which involves segmenting each seed and eliminating shadows from the maize. The images must be converted into grayscale since they utilize shape and dimensional characteristics.

**Table 1: Sunlight intensity and time while acquiring 6 maize seed varieties’ digital images**

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Maize varieties</th>
<th>Time</th>
<th>Intensity of sunlight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ICI-339</td>
<td>3:00 PM</td>
<td>340 Lux</td>
</tr>
<tr>
<td>2</td>
<td>Desi Makkai</td>
<td>2:35 PM</td>
<td>365 Lux</td>
</tr>
<tr>
<td>3</td>
<td>Neelam Makkai</td>
<td>2:15 PM</td>
<td>369 Lux</td>
</tr>
<tr>
<td>4</td>
<td>Kashmiri Makkai</td>
<td>1:45 PM</td>
<td>377 Lux</td>
</tr>
<tr>
<td>5</td>
<td>Pioneer P-1429</td>
<td>1:20 PM</td>
<td>384 Lux</td>
</tr>
<tr>
<td>6</td>
<td>Sygenta ST-6142</td>
<td>1:00 PM</td>
<td>385 Lux</td>
</tr>
</tbody>
</table>

Figure 2: Gray-Level (GL) images of maize seeds’ 6 varieties
Proposed methodology

This section discusses the proposed methodology’s detailed description. Initially, preprocessing of maize seed images is performed. As per the knowledge, for seed segmentation, there does not exist any ideal approach and most commonly for the segmentation approach ROI extraction is utilized because there exist some limitations to human-based segmentation. To its solution, a new “TRGS (threshold-based region growing segmentation) technique” was proposed by researchers. In this approach, it depends on the seed value which is compared to the entire image’s adjacent pixels, in the case of a GL (gray level) value higher than 5, the area/region is increased known as “ROI (region of interest)”. For every ROI, for obtaining spectral features, 2nd order statistical texture features, 1st order histogram parameters, “computer vision and image processing (CVIP)” that is image processing software tool is utilized. For every ROI, 55 features were extracted that are categorized as 19 spectral features, 26 texture features with 5 average texture values in 4-dimensions whereas in 10 first-order histograms. Statistically, it represented that for six maize seed varieties 330,000 (6000 × 55) features vector space input data is used. Figure 3 shows the classification framework for maize seed types.

Hybrid-feature acquisition

In maize seed classification, hybrid features proved to help overcome the noise effect of a pixel. Some distinctive hybrid information such as geometrical, spectral, and statistical information is held by digital images. Furthermore, medical image classification and leaf classification greatly implemented the hybrid feature [16, 17].

Figure 3: Classification framework for maize seed varieties
Histogram features
The object that is concerned with columns, as well as rows, are selected with the histogram features. For feature extraction, this binary object is utilized as the original image’s mask. Individual pixels’ intensity is utilized for calculating the histogram features and these intensities are the object’s part. Histogram features depend on histograms. These are also known as “statistical features” or “first-order histogram”. [16]

Spectral features
The frequency domain-dependent features are called spectral features. While images are classified depending on their texture, these features are proved to be very useful. In various areas, they are measured as a power form, as well as these areas, are known as “sectors” and “rings”.

Texture feature
These features are often called “second-order statistical features”. For obtaining the texture features angle as well as distance among pixels is determined. Texture features depend on the “GLCM (gray level co-occurrence matrix).” For the present study, 5 texture features are used in 4 dimensions that are, 135°, 90°, 45°, and 0° as well as a 5-pixel distance among pixels. Furthermore, 5 second-order texture features are obtained that are energy, inverse difference, correlation, inertia, and entropy.

Feature selection
It was observed that for maize seed classification, there does not present an equal significance of 55 extracted features for every ROI. It is not easy handling large-scale datasets which means 6000 × 55 = 330,000 hybrid-feature data space. Therefore, the vector space of extracted feature must be reduced. For acquiring this reduction, the best first search algorithm along with the supervised feature selection method known as “correlation-based feature selection (CFS)” were utilized in this dataset. In the dataset, the significant features are extracted with the help of CFS. [18]

Classification
There exist various traditional machine learning classifiers for classification, and it has been found that BN (Bayes Net), RF (Random Forest), LB (LogitBoost), and MLP (Multilayer Perceptron) classifiers perform well on hybrid features dataset of maize seed. It has been observed that for ROIs size (50 × 50), (75 × 75), (100 × 100) very low accuracy (<85%) is performed by the above-mentioned classifiers that were not suitable. Therefore, for acquiring effective classification, the ROI size is increased to (125 × 125) which results in better accuracy of nearly 93.76%. Furthermore, by increasing the ROI size, the accuracy of classification is also improved. Figure 4 represents the hybrid-feature analysis MLP framework. The “GREEN” colored features are represented in the first layer which is at the input layer, whereas the “RED”
colored second layer represented 15 neurons in the invisible/hidden layer. Furthermore, “YELLOW” nodes at the third layer represented hidden layers’ weights at the output layer.

![Figure 4: Hybrid-feature analysis MLP framework](image)

4. RESULTS
In the present study, 6 maize seed varieties’ comparative analysis is performed with the help of 4 supervised classification algorithms BN, RF, LB, and MLP by utilizing a (10-fold) cross-validation approach on fused optimized hybrid features datasets. There exist 2 reasons behind this, firstly: maize seeds’ 6 varieties are pre-defined; secondly, naturally acquired noisy input data. [32] Table 2 represents various evaluating parameters that include RMSE (root mean squared error), MAE (mean absolute error), KS (kappa statistics), ROC (receiver operating characteristic) area, FP (false positive) rate, and TP (true positive) rate.
Figure 5: Proposed Methodology

Table 2: ML classifier implementation on maize seed dataset having ROI size (125 × 125)
Figure 6: Classification results of MLP classifier for ROI size (125 × 125) of maize seed varieties

Table 3: ROI size (125 × 125) for maize seed classification by using MLP classifier’s confusion matrix (CM)

<table>
<thead>
<tr>
<th>Classes</th>
<th>Desi Makkai</th>
<th>Syngenta ST-6142</th>
<th>Kasmri Makkai</th>
<th>Pioneer P-1429</th>
<th>Neelam Makkai</th>
<th>ICI 339</th>
<th>Total</th>
<th>Accuracy %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Desi Makkai</td>
<td>990</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>8</td>
<td>1000</td>
<td>1000</td>
<td>99%</td>
</tr>
<tr>
<td>Syngenta ST-6142</td>
<td>6</td>
<td>954</td>
<td>10</td>
<td>0</td>
<td>30</td>
<td>1000</td>
<td>1000</td>
<td>95.4%</td>
</tr>
<tr>
<td>Kasmri Makkai</td>
<td>2</td>
<td>10</td>
<td>935</td>
<td>12</td>
<td>41</td>
<td>1000</td>
<td>1000</td>
<td>93.5%</td>
</tr>
<tr>
<td>Pioneer P-1429</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>979</td>
<td>0</td>
<td>20</td>
<td>1000</td>
<td>97.9%</td>
</tr>
<tr>
<td>Neelam Makkai</td>
<td>53</td>
<td>1</td>
<td>41</td>
<td>931</td>
<td>39</td>
<td>891</td>
<td>1000</td>
<td>93.1%</td>
</tr>
</tbody>
</table>

For improving the results of classification, ROIs size 150 × 150 maize seed images are used by above mentioned classifiers with a similar strategy as well as results in obtaining better results with BN, RF, LB, and MLP as 96.67%, 97.22%, 97.78%, and 98.83 respectively, as table 4 represented these results. Figure 7 represents the maize seed’s 6 varieties’ namely, ICI 339, Pioneer P-1429, Syngenta ST- 6142, Neelam Makkai, Kasmri Makkai, and Desi Makkai, distinctive classification results as 99.4%, 99.9%, 98.6%, 98.5%, 97%, and 99.8%, respectively. The average classification ratio rate ranged from 99.8% – 90%. Furthermore, the MLP classifier’s confusion matrix (CM) utilizing ROIs size 125 × 125 on the maize seed varieties dataset is represented in table 5.

It has been found that in comparison to existing techniques, the proposed methodology proved to be better. It is satisfactory, reliable, and enhanced in comparison to available seed classification methodologies. Table 6 presents the proposed methodology’s comparative analysis with existing methodologies. The last row represents the proposed methodology.

Table 4: ML classifier implementation on maize seed dataset having ROI size (150 × 150)
Figure 7: Classification outcomes of MLP classifier for ROI size (150 × 150) of maize seed varieties

Table 5: ROI size (150 × 150) for maize seed classification by using MLP classifier’s confusion matrix (CM)
5. CONCLUSION

The present study proposed a system that with the help of ML approaches classifies maize seeds’ 6 varieties with a digital image dataset. The study aims at developing as well as selecting appropriate classifiers (BN, RF, LB and MLP) along with efficient features for effective classification. Therefore, 6 maize seed varieties that include ICI 339, Pioneer P-1429, Syngenta ST- 6142, Neelam Makkai, Kasmiri Makkai, and Desi Makkai are used for exploring classification techniques. Also, various environmental factors like moisture, illumination, temperature, and sunlight along with several dataset features affect the results. Most efficient
classification outcomes among implemented classifiers are provided by the MLP classifier. Moreover, reduction, selection, feature extraction, and image processing techniques are also used. Classifier’s overall performance is significantly affected by the feature’s selection and extraction in every classification method. Furthermore, the total execution time can be maximized as well as classification accuracy can be minimized due to the large no. of features. This is the reason that the maize seed varieties are classified using an optimized hybrid-feature dataset.

6. REFERENCE