ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2022.0387

PREDICTION OF SEED PURITY AND VARIETY IDENTIFICATION USING IMAGE MINING TECHNIQUES

M. Suganthi and J.G.R. Sathiaseelan
Department of Computer Science, Bishop Heber College, India

Abstract

Seed is a little embryonic plant that can be used to introduce plant infections to new areas while also allowing them to survive from one cropping season to the next. Seed health is a well-known component in modern agricultural science for achieving the required plant population and yield. Seed-borne fungus are a major biotic constraint in seed production around the world. The detection of seed-borne pathogens by seed health testing is a crucial step in the treatment of crop diseases. Speed and accuracy are critical requirements for long-term economic growth, competitiveness, and sustainability in agricultural output. Because human judgements in identifying objects and situations are variable, subjective, and delayed, seed prediction activities are costly and unreliable. Machine vision technology provides a nondestructive, cost-effective, quick, and accurate option for automated procedures. Seed variety, seed type (country seed or hybrid seed), seed health, and purity prediction were the four basic processes. We began the first procedure by aligning the seed bodies in the same direction using a seed orientation approach. Then, to detect atypical physical seed samples, a quality screening procedure was used. Their physical characteristics, such as shape, colour, and texture, were retrieved to serve as data representations for the prediction. This research introduces a new fuzzy cognitive map (FCM) model based on deep learning neural networks that predicts seed purity tests using data from biological investigations. The relevant data features from the seed test are extracted by FCM, which then effectively initialises the deep neural networks. The Levenberg–Marquardt (LM) technique for deep neural networks was discovered to improve seed purity test prediction. Four statistical machine learning algorithms (BP-ANN, Multivariate regression, and FCMLM deep learning). Furthermore, we demonstrated an improvement in the system's overall performance in terms of data quality, including seed orientation and quality screening. In independent numerical testing, the correlation coefficient between predicted values and true values acquired from experiments reached 0.9.

Keywords:

Fuzzy Cognitive Map, Deep Learning, FCMLM Deep Learning, BP-ANN, Multivariate Regression, Seed Purity

1. INTRODUCTION

Seed testing is the cornerstone of all other seed technologies. Seed testing is used for control of quality parameters during seed handling, and test results are submitted to customers as documentation on seed quality. It is the means by which the quality of seed can be measured and viability of seed is ensured. Seed testing is determining the standards of a seed lot namely physical purity, moisture, germination, vigor and thereby enabling the farming community to get quality seeds. One of the most important tasks of the agency is to control the seed quality. The contamination causes many problems such as variety impurity, seed mutation, or cross-breeding, which may result in poor quality production. In the traditional way, the examination of contamination in breeding seeds has been done by seed experts. The experts use personal skills to consider morphological structure, shape, texture, and color in many parts of the seed to make a decision. In the examination, they classified a specific type of seeds from a specific locality. Firstly, they put seed samples, which are supposed to be the same type on a table and examine them with tools such as a large magnifying glass, an illumination, and forceps. Then, they try to find and bring out seeds with different physical characteristics, which are contaminated seeds from other types. With the limitations of being human, a large number of seed inspectors take quite a long time in the process because it is difficult for the human eyes to find small differences in one seed among many seed samples.

Over the past decade, computer vision has been widely used in various domains. Several methods in the field of computer vision have been changed from statistical methods to deep learning methods because it offers greater accuracy for tasks such as object detection and image recognition. The technology can help computer scientists to develop tasks in various fields rapidly. It can automatically learn features from the given data while the traditional machine learning methods need feature engineering one at a time. It can handle the variability and deviations of the data that are very similar. However, deep learning technology is rather complicated. It has a large network structure and requires a large amount of training data, time-consuming, and high-performance computing resources. In order to compensate for the individual deficiencies of FCM and ANNs, this research combines FCM based on genetic algorithms (GA) with ANN based on deep learning to put forward a new model. The resulting model exploits the advantages of each technique to predict seed purity effectively.

2. RELATED WORK

Anami et al. [1] was the only researcher that proposed a rough assessment of rice quality instead of the one-grain classification. They attempted to classify the level of adulteration from the image of mixed bulk paddy samples varied between 10-30% of the adulteration levels.

Weihua Liu [2] propose a hyperspectral rice seed purity identification method based on the LASSO logistic regression model (LLRM). The feasibility of using LLRM for the selection of feature wavelength bands and seed purity identification are discussed using four types of rice seeds as research objects. The results of 13 different adulteration cases revealed that the value of the regularisation parameter was different in each case. The recognition accuracy of LLRM and average recognition accuracy were 91.67–100% and 98.47%, respectively. Furthermore, the recognition accuracy of full-band LRM was 71.60–100%. However, the average recognition accuracy was merely 89.63%.

Nadia Ansari [3] proposed the feasibility of paddy seed varietal purity inspection using machine vision combined with multivariate analysis based on color, morphological, and texture features became successful. Good classification performance for distinguishing among the paddy seed samples was achieved using selected features data, with a classification accuracy of 83.8% in PLS-DA, while 93.1% and 87.2% were performed for the SVM-C and KNN model respectively. This result indicates that machine vision can be used to monitor paddy seed varietal purity.

Xiaolong [4] propose a time-efficient method by pressing soybean seeds into rubber sand filled with culture plates through a ruler to ensure a relatively uniform surface height. The results showed that the support vector machine (SVM) obtained the optimal identification accuracy of 90% in the prediction set. In addition, PCA-ResNet (propagation coefficient adaptive ResNet) and PCSA-ResNet (propagation coefficient synchronous adaptive ResNet) were designed based on typical ResNet structure by changing the way of self-adaption of propagation coefficients. Combined with a new form of input data called spectral matrix, PCSA-ResNet obtained the optimal performance with the discriminate accuracy of 91.75% in the prediction set.

Qiu et al. [5] identified 4 varieties of rice seeds using hyperspectral imaging with three machine learning methods, namely, k-NN, SVM, and CNN. The experiment was studied on two different spectral ranges, and the numbers of training samples are varied. A hyperspectral camera was adopted to deal with the problem of rice varieties classification in many researches. However, the instrument was costly and complex. Moreover, it required a fast computer, sensitive detectors, and large data storage capabilities.

Lin et al. [6] proposed a comparison between two techniques, CNN and traditional methods, to distinguish rice grains between three different shapes (medium, round, and long grain). They studied 5,554 images for calibration purposes and 1,845 images for validation purposes. The experiments adjusted training parameters such as batch size and epochs in the CNN method. They also carried out an experiment using the traditional statistic methods that got a classification accuracy ranging from 89 to 92%. On the other hand, the experiment using the CNN method had given 95.5% classification accuracy, which was higher than the traditional methods.

From the above literature survey, most research identified seed varieties from a few species. Furthermore, they studied only few tenth image samples or few hundred images of each seed species. The limited number of samples might cause data bias, and insufficient variation may cause the trained model not to be general enough for practical uses.

3. PROPOSED METHODOLOGY

In this work, we presented a technical study of seed verity, seed identification, seed purity and seed quality inspection by evaluating over 14 varieties, as shown in Fig.1. We analysed more than 3,500 images in each species from various planting sources. The hardware consisted of a tray for conveying seed, a photographic part, a contaminated-seed detector, and a contaminated-seed elimination part. Therefore, many grain samples were collected to cover the diversity of each seed to assess the potential or the limitation of the efficiency of classification obtained from each seed identification. We plan to improve the technique in future efforts.

Our seed varieties perdition process consisted of the following steps: object orientation to align seed image in the same direction, image screening for outlier/irregular/abnormal seed or tilted seed, feature extraction for retrieving physical seed properties, and seed varieties. The system overview is presented the perdition performances were evaluated by comparing traditional machine learning technique. We investigated the performance of each seed variety in both subgroups and collective groups. We also presented preliminary results on data quality aspects such as quality screening and seed orientation.

4. MATERIALS AND METHODS

4.1 SAMPLE PREPARATION

Seed samples of 14 varieties (shown in Fig.1), obtained from various provinces in order to cover different characters which are varied depending on producing environments. The samples were prepared, and only complete seeds were selected, by experts from the seed department.

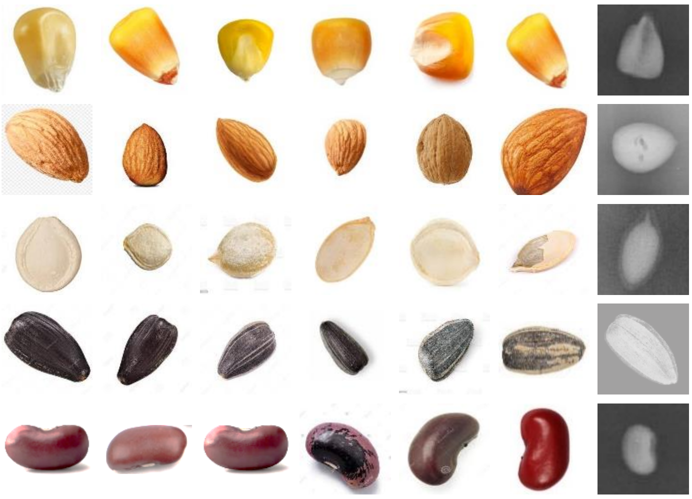
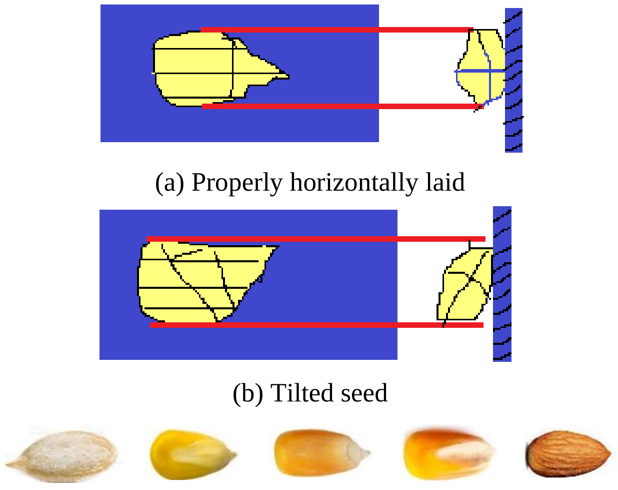


Fig.1 Sample of Collected seed images (Include X-Ray images)

4.2 TRAINING IMAGE ACQUISITION

Each training image of seed was acquired from a scanner with a special box tray, which could roughly separate each seed sample and usually adjusted each seed to be aligned horizontally as shown as in Fig.2(a), because a seed which was not horizontally laid (Fig.2(b)) did not show all of the seed features properly. The seeds, which were not horizontally laid, were tilted. Examples of tilted seeds are shown in Fig.2(c).

The obtained images were rechecked to get rid of images that contained more than one seed and images that the seed was not properly aligned horizontally. An object region in each input image was extracted by applying background-subtraction using Otsu thresholding method. Then, ellipse fitting with coordinates of object contour was used to calculate the object approximate size from the ellipse major axis and minor axis values. If the object size varied greatly from the average size, the object in the image was determined not properly aligned in the horizontal orientation.



(a) Properly horizontally laid

(b) Tilted seed

(c) Tilted seed examples

Fig.2. (a) Schematic of properly horizontally laid seed and the corresponding cross-sectional view; (b) tilted seed and the off dimensions; (c) photographs of tilted seed examples

4.3 PRE-PROCESSING

After getting a single seed, this section dealt with a preparation of quality seed image, which consisted of two parts: seed alignment and seed quality screening.

4.3.1 Seed Orientation/Alignment:

This process examined and rotated the seed body into the horizontal axis direction, so that all seeds' head-and-tail directions would be aligned in the same direction. Being aligned in the same direction simplified extracting features and analyzing data. This process is necessary because the grain might move or be misaligned during scanning. The procedure details are described as follows:

The seed image was processed based on coordinates of the object contour therefore the image needed to be rotated and flipped so that the image appeared as shown schematically in Fig.2(a). The seed head pointed to the left and the tail pointed to the top-right. After adjusting the alignment, shape features that expressed the head and the tail were easily extracted by the method.

The distances between each object contour points and the object centroid were calculated. The head point and the tail point were defined as tip points, which were the point furthest away from the centroid and the point locally furthest away from the centroid on the opposite side. A common physical shape of corn seed is shown in Fig.4. The shape around the head normally had a relatively symmetrical corner. In contrast, the shape around the tail had an unsymmetrical corner and might have two small corners due to the structure of lemma partly hanging over the palea. As a result, the head tip point could be determined by calculating the object area around the tip point, shown by the red triangle in Fig.3, and by comparing their sizes. After the head point was determined, the seed image could be rotated so the head point was on the left of the image and the line between the head point and the centroid was parallel to the X-axis in the image shown in Fig.2(a).

4.3.2 Seed Quality Screening:

It was important that input data images had high quality because we dealt with a lot of images and a large sample collection. A delivering of inappropriate data to be analyzed in the system should be avoided. There were two types of seed samples during the data preparation that should have been discarded. The first type, outlier seed, was caused by raw material itself while the other type, the tilted seed, was an error on the procedure in the seed scanning process.

- Outlier seeds were corn seeds that had different shapes from the standard one, e.g., very long tail, large crack, and smudged skin. Samples of outlier seeds are shown in Fig.5

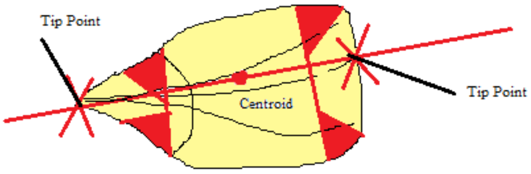


Fig.3. Seed orientation

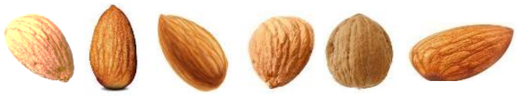


Fig.4. Outlier seed samples

- Tilted seeds were corn seeds that had an oval shape when viewed cross-sectionally. It might tilt up when laid on the flat surface of the scanner. Examples of tilted seeds are shown in Fig.2(b) and Fig.2(c) A seed quality screening process to examine the two cases is described as the following.

We extracted features (shape, color, and texture presented in feature extraction) from each sample and used the features as input data for the DBSCAN technique, one of the most popular clustering techniques, to detect outlier from sample data. DBSCAN uses a density-based clustering algorithm to detect abnormally of multidimensional data. The algorithm identified and clustered in a high-density region separated from a low-density region throughout the two parameters, eps (radius of neighborhood region), and MinPts (minimum number of points). A point was decided as a clustered region only if there were more neighbors than MinPts and was within the eps. Otherwise, a point that did not satisfy the condition was defined as an outlier point. In the tilted case, the shapes of the seed had more distinctive characteristics than the diverse shape of an outlier seed. Most of the tilted seed was more symmetrical in shape along the length of the body than the seeds laid horizontally. Here, the SVM technique with our shape features was applied to tackle this problem. A classify model was created to identify seed types between a tilted seed and a horizontal seed.

4.4 FEATURES EXTRACTION

After screening out inappropriate samples from the previous process, a proper seed sample image was rotated in the same direction and was easily facilitated to extract the data features on the various parts of the seed body.

4.4.1 Seed Verity Identification:

- **Shape**: We used basic physical shape features as referred in [13]. The extracted values are shown below. C = circularity was calculated from Eq.(1), while A = object area and P = object perimeter

$$C = 4\pi \frac{A}{P^2} \tag{1}$$

where, R = roundness and C_o = compactness was calculated from Eq.(2), while D_{max} = object maximum diameter.

$$R = \frac{4A}{(\pi D_{\max})^2}, \sqrt{\frac{4A}{\pi D_{\max}}} \tag{2}$$

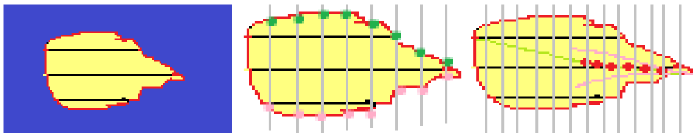
Shape factors were calculated from Eq.(3). F_1 = shape factor 1, F_2 = shape factor 2, F_3 = shape factor 3, d_1 = major axis length and d_2 = minor axis length.

$$F_1 = \frac{d_1}{A}; F_1 = \frac{A}{d_1^3}; F_3 = \frac{4A}{\pi d_1 d_2} \tag{3}$$

where, FA = area factor was calculated from Eq.(4) when A_{hull} = object convex hull area was shown as an area surrounded by a red line in Fig.6(a).

$$F_A = \frac{A}{A_{hull}} \tag{4}$$

Slope factors were calculated from contour pixel coordinates. The object was divided into N equal parts. From the experiment, N was determined to be 9. The point on the left of the object represented the head point P_{head}, and the point on the right of the object represented the tail point P_{tail}, shown by pink points in Fig.5(b). We defined S_{uppern} = n^{th} slope factor on the upper side, S_{lowern} = n^{th} slope factor on the lower side, P_{uppern} = n^{th} dividing point on contour on the upper side (blue points), and P_{lowern} = n^{th} dividing point on contour on the lower side (green points).



(a) Surrounded by red line (b) Slope factor (c) Tail Slope factor

Fig.5. Shape feature

When n=1,2,3,…,N-1,

$$S_{upper_n} = \frac{P_{upper_n}(y) - P_{upper_{n-1}}(y)}{P_{upper_n}(x) - P_{upper_{n-1}}(x)} \tag{5}$$

$$S_{lower_n} = \frac{P_{lower_n}(y) - P_{lower_{n-1}}(y)}{P_{lower_n}(x) - P_{lower_{n-1}}(x)} \tag{6}$$

The Fig.6(c) shows the illustration of S_{tail}, tail slope factors. Q_{midn} =n^{th} point dividing object into N_{tail} equal parts on the object middle line shown as the green line. Examples of Q_{midn} are shown as red points. The object middle line, L_{mid}, was calculated from the thin object area. Q_{uppern} (Blue point) and Q_{lowern} (green point) were calculated in the same way as the Q_{midn}, using the object upper contour and the lower contour instead of L_{mid}. N_{tail} was experimentally determined to be 21.

When m = 1,2,⋯,M and M was experimentally determined to be 7. While pos ∈ {mid,$upper$,$lower$}, $S_{tailposm}$ and $S_{tailavg}$ were calculated as in Eq.(7) and Eq.(8).

$$S_{tail\ pos_m} = \frac{P_{tail}(y) - Q_{pos(N_{tail}-m)}(y)}{P_{tail}(x) - Q_{pos(N_{tail}-m)}(x)} \tag{7}$$

$$S_{tail\ avg} = \frac{\sum_{m=1}^{M} S_{tail\ midm}}{M} \tag{8}$$

We also used object contour shape histogram 180 degree around tail tip as shape factors. Example is shown in Fig.7. From this histogram, we calculated a hair around tail tip frequency value and used as another feature.

5. SEED IMAGE DETECTION METHOD

In order to detect the seed with known shape in the seed image, the shape template or window of the object is used to match the seed country seed or hybrid seed, and the object seed is detected under a certain quasi-measurement. The template matching method can detect the lines, curves, patterns and so on in the image.

5.1 CROSS CORRELATION MATCHING

There are many ways to measure the degree of matching between two seed images f_1 and f_2 in regions Φ, such as mismatch can be expressed in the form of max_φ=$|f_1$-$f_2|$; $\iint_\varphi |f_1$-$f_2|$; $\iint_\varphi (f_1$-$f_2)^2$.

If mismatch is use $\iint_\varphi (f_1$-$f_2)^2$ and $\iint_\varphi (f_1$-$f_2)^2$=$\iint_\varphi f_1^2$+$\iint_\varphi f_2^2$+2$\iint_\varphi f_1 f_2$, Obviously, after given $\iint_\varphi f_1^2$, $\iint_\varphi f_2^2$ and 2$\iint_\varphi f_1 f_2$ it is a measure of matching, the bigger the item is $\iint_\varphi (f_1$-$f_2)^2$ the smaller, the lighter the mismatch degree is, the better the matching degree. Applying the Canchy-Schwarz inequality, for nonnegative f_1 and f_2, the following conclusions can be drawn:

$$\iint_\varphi f_1 f_2 \le \sqrt{\iint_\varphi f_1^2 \cdot \iint_\varphi f_2^2}$$

If and only if f_1 = cf_1 the equality sign holds (c is constant). In digital images, integrals are converted to summations, and the result is changed to

$$\sum_i \sum_j f_1(i,j) f_2(i,j) \le \sqrt{\sum_i \sum_j f_1(i,j) f_2(i,j)}$$

Similarly, if and only if $f_2(i,j)$=$cf_1(i,j)$ the equality sign holds (c is constant). When the target template f_1 is set and c is the image to be matched, obviously the f_1 should be assumed to be smaller than f_2. Then we will move f_1 at all possible positions in f_2 and calculate $\iint_\varphi f_1 f_2$ to each shift (u,v). According to Cauchy-Schwartz inequality, the following formula holds

$$cf_1(x,y) f_2(x+u, y+v) dxdy$$
$$\le \sqrt{\iint_\varphi f_1^2(x,y) dxdy \cdot \iint_\varphi f_2^2(x+u,y+v) dxdy}$$

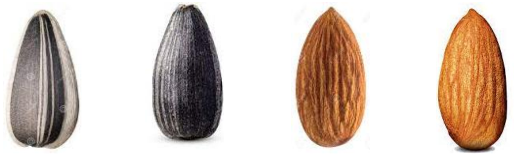
Because f_1 all of them are equal to 0 outside the region φ, the integral region can be expanded from φ to (-∞,+∞), so that the left part of the upper formula can be changed to:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_1(x,y)\, f_2(x+u,y+v)\,dxdy$$

so in upper formula f_1 and f_2 of cross correlation function cf_1f_2. On the right side of the analytic formula as in Eq.(1), although the $\iint_{\varphi} f_1^2$ term is a constant, but $\iint_{\varphi} f_2^2$ is not a constant. It is related to u and v. This is because in practice, the template f_1 is usually fixed and the image f_2 to be matched is moved. Therefore, the content of them image f_2 with the corresponding region of f_1 always varies with u and v. Simple apply cf_1f_2 as a matching measure is not suitable. Generally, the domesticated cross-correlation function is used as a matching measure, that is, the domesticated cross-correlation function is used as a matching measure, namely:

$$\frac{c_{f_1f_2}}{\sqrt{\iint f_2(x+u,y+v)\,dxdy}}$$

If makes $f_2=cf_1$ (c is constant) on a displacement (u,v), then (formula 6) has a maximum value $\iint_{\varphi} f_1^2$. At this time Eq.(2) or Eq.(3) holds, then Eq.(1) there is a minimum, that is, the minimum mismatch. In fact, because of the existence of noise, the above Eq.(2) will not occur, that is, it is impossible to match completely. Generally, the position of the maximum of formula Eq.(6) is chosen as the best matching point.



Normal Seed Hybrid seed Country Seed Hybrid seed

Fig.6. Seed match test (Country seed or Hybrid seed)

- **Color**: We used RGB color space in calculation and color features below from each pixel color value in the object area. When color ∈ $f_{R,G,Bg}$ and $p_{color}(x,y)$ = pixel value of the color in the object at coordinate (x,y) and N was the number of pixels of the object. We used the features below for each color.

$$\text{Max}(P_{color}(x,y))$$

$$\text{Min}(P_{color}(x,y)).$$

$$Avg\left(P_{color}(x,y)\right)=\sum P_{color}(x,y)/N$$

$$Var\left(P_{color}(x,y)\right)=\sum P_{color}(x,y)-Avg\left(\sum P_{color}(x,y)\right)^2/N$$

An example of Angle-Distance histogram graph around tail point as shown in Fig.7.

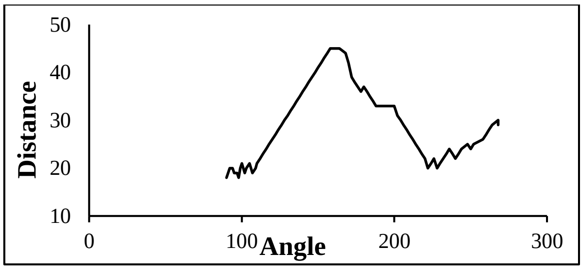
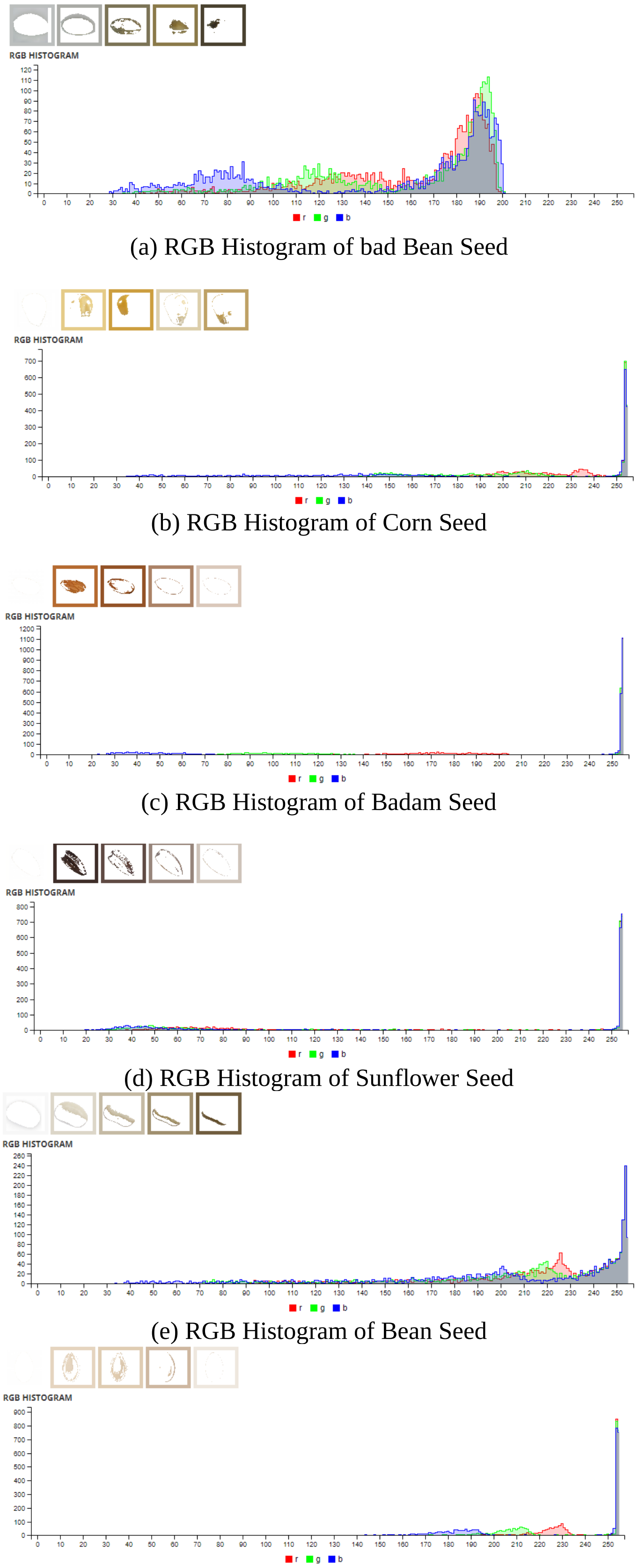


Fig.7. Example of a contour shape histogram around tail point

We also used color histograms on RGB color space which represented color distribution on the object shown in Fig.8.



(a) RGB Histogram of bad Bean Seed

(b) RGB Histogram of Corn Seed

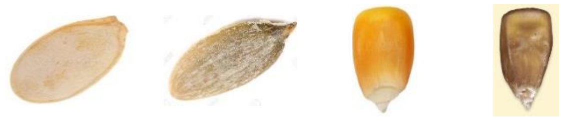
(c) RGB Histogram of Badam Seed

(d) RGB Histogram of Sunflower Seed

(e) RGB Histogram of Bean Seed

(f) RGB Histogram of Badam Seed

Fig.8. RGB Histogram



Normal Seed Bad Seed Normal seed Bad seed

Fig.9. Seed color test

5.2 X-RAY SEED PURITY TEST

We proposed an approach using X-ray images to investigate internal tissues because seed surface profile can be negatively affected, but without reaching important internal regions of seeds. A badam plant was used as a model species, which also serves as a multi-purposed crop of economic importance worldwide.

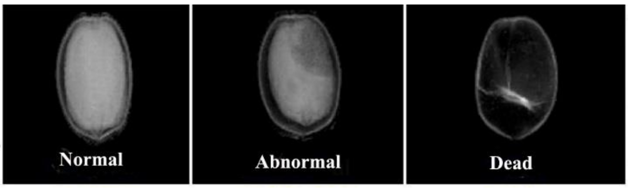


Fig.10. Seed Purity test

Badam seed has a thick and dark tegument. The Fig.10 shows X-Ray images obtained from ventral and dorsal surface of the seeds based on germination capacity and corresponding reflectance images captured at 940 nm (Wavelength Range) images.

6. FUZZY COGNITIVE MAP BASED ON DEEP LEARNING

A FCM model can extract the features from sample data and get the concept node and the connection weights. However, it lacks the ability to fine-tune the parameters and so it is not suitable for precise quantitative analysis. The new algorithm proposed in this paper combines the advantages of FCM and deep neural networks. First, a FCM model is trained; then, the FCM is mapped into a deep neural network, the hidden layers of neural network are initialized layer by layer using the connection weights of the FCM; Finally, the weights of the output layer of the network are trained using the Levenberg–Marquardt (LM) algorithm. This results in the FCM model based on deep learning which can predict seed purity rate. The Flow chart of FCM–LM algorithm is shown in Fig.11.

6.1 FCM BASED ON GA

GA is a searching algorithm which simulates biological genetic and evolutionary processes in the natural environment and forms an adaptive global optimization. It uses encoding technology to express a variety of complex structures, maps the original problem solution space to strand space, guides learning and determines the search direction by genetic and selection operations. GA has the ability of searching the approximate optimal solution in complex space. It is an iterative process of the population (candidate solutions), and Fitness functions can be used to evaluate the quality of these populations.

GA can be defined as an 8-tuple:

$$\text{GA} = \{I, F, P_0, N, S, C, M, O\},$$

where I is the coding for individuals, F is the individual fitness function, P_0 is the initial population, N is the size of population, S is the selection operator, C is the cross-over operator, M is the mutation operator, O is the termination condition for the genetic operation.

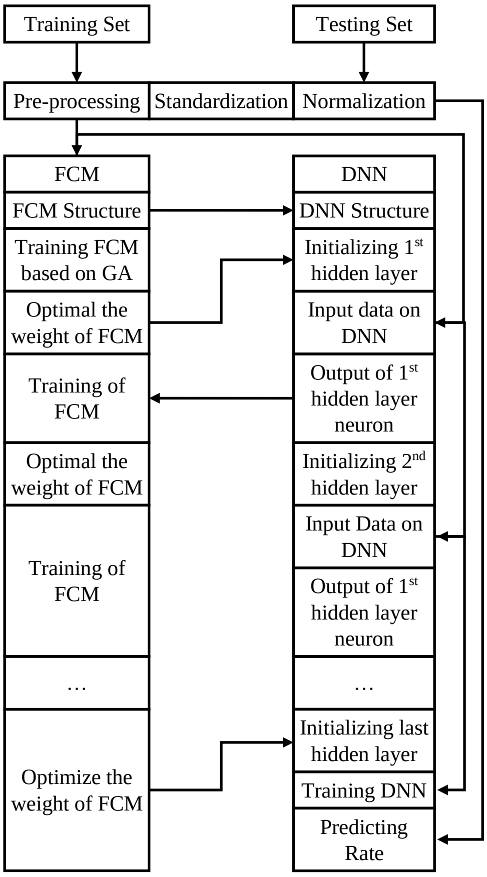

Training Set
Testing Set
Pre-processing
Standardization
Normalization
FCM
FCM Structure
Training FCM based on GA
Optimal the weight of FCM
Training of FCM
Optimal the weight of FCM
Training of FCM
…
Optimize the weight of FCM
DNN
DNN Structure
Initializing 1st hidden layer
Input data on DNN
Output of 1st hidden layer neuron
Initializing 2nd hidden layer
Input Data on DNN
Output of 1st hidden layer neuron
…
Initializing last hidden layer
Training DNN
Predicting Rate

Fig.11. Flow chart of FCM–LM deep learning algorithm

The GA for FCM is as follows:

Step 1: Code each connection weight of FCM in 16 bit binary (due to the characteristic of the Sigmoid function, set the initial weights be in [0, 1]).

Step 2: Construct a chromosome (individual) with all the binary weights in series.

Step 3: Generate N individuals randomly to constitute a population.

Step 4: Each individual was decoded into a corresponding FCM.

Step 5: Input the data from the samples and train each FCM, respectively.

Step 6: Calculate the error function $E(\omega)$ of each FCM

$$E(\omega) = 0.5\sum_{p=1}^{k}\sum_{i=1}^{m}\left(d_{pi} - x_{pi}\right)^2$$

where d_{pi} is the ith expected output of the pth sample, x_{pi} is the ith actual output of the pth sample.

Step 7: Transform the error function into the Fitness function of each individual, F=1/(E(w)+ε where ε is a small positive number that makes denominator of Fitness function nonzero.

Step 8: Select individuals according to their Fitness function value (arranged from the largest to smallest to consist the parent).

Step 9: Fitness proportional operator, one-point cross-over operator and simple mutation operator are used to produce a new population.

Step 10: Repeat steps 5–9 until a certain individual can meet the requirements of the error, or the number of the iteration has reached the maximum.

The individual with the maximum Fitness in the evolutionary process is decoded into an FCM.

6.2 MAPPING FCM TO DEEP NEURAL NETWORK (DNN)

Composed of a multilayer perceptron neural network, the depth is the number of nonlinear computations combined in network learning. Previous neural network learning algorithms were more focused on a shallow structure such as an input layer, a hidden layer, and an output layer. In contrast, this research employs a deep neural network which has more nonlinear computations and a structure of an input layer, two hidden layers, and an output layer.

6.3 FCMLM DEEP LEARNING PROCEDURE

Gradient Descent (GD) algorithms reused in neural network training widely. In addition to the basic BP algorithm, there is a momentum BP algorithm, resilient BP algorithm, conjugate gradient BP algorithm, quasi-Newton algorithm, and LM algorithm, to name a few. The LM algorithm is the fast algorithm to train the middle scale feed forward neural network (from tens to hundreds of connection weights) and can generally be used in single hidden layer network structures. To improve the accuracy of the network prediction, this research has developed a FCMLM deep learning algorithm which initializes the weights of hidden layer based on FCM and optimizes the weights of the output layer by the LM algorithm

6.4 NETWORK INITIALIZATION

Initialization is the pre-training of the network weights. In this model, the pretraining is completed by FCM and the connection weights of the trained FCM are mapped to the connection weights of the hidden layers of the neural network. The specific steps are as follows:

Step 1: Train the FCM based on GA by the sample data and gets the weight matrix $W_{n\times n}$ of the FCM.

Step 2: Remove w_{ij}^{o} (the weight of the output layer neuron) from $W_{n\times n}$ to get.

$$W'_{(n-1)\times(n-1)}, W'_{(n-1)\times(n-1)} = \left\{ w_{ij} \left| w_{ij} \in W_{n\times n} \bigcap w_{ij} \neq w_{ij}^{o} \right. \right\}$$

Step 3: Set $W'_{(n-1)\times(n-1)}$ as the initial connection weight matrix W^1 of the input and first hidden layer.

Step 4: The output of the hidden layer neurons is put into the FCM and train the FCM to get the weight matrix $W_{n\times n}$ of the FCM.

Step 5: Remove w_{ij}^{o} from $W_{n\times n}$ to get $W'_{(n-1)\times(n-1)}$ and set it as the initial connection weight matrix W_L of the hidden layer and the next hidden layer, where L is the number of hidden layers.

Step 6: Repeat steps 3 and 4 to the last hidden layer.

Step 7: Initialize the weights of the last hidden layer and the output layer with:

$$W_{(n-1)\times 1}^{o}, W_{(n-1)\times 1}^{o} = \left\{ w_{ij} \left| w_{ij} = w_{ij}^{o} \right. \right\}$$

6.5 PREDICTION RESULTS FROM STATISTICAL TECHNIQUES

After the seed object was aligned in the horizontal direction, the object physical characteristic information was extracted by using the proposed method described. In the prediction process, some data were discarded from the seed screening process, resulting in different proportions of the sample number of each class. Therefore, we cut the number of samples according to the classes with the least samples to provide balanced information. These samples were randomized in equal proportions at 2,900 samples per class and divided into a training set and a validation set with a proportion of 80% and 20%, respectively.

7. MODEL EVALUATION

In order to evaluate the predictive ability of the model, the effects of the FCMLM deep learning algorithm and other methods for predicting protein folding rate were compared. The correlation coefficients and number of protein sequences are given in Table.2.

From Table.2, it can be seen that the correlation coefficient of the FCMLM deep learning algorithm is the highest and the number of protein sequences of FCMLM deep learning algorithm is the most among all methods. In experiment for predicting protein folding rate change, the results of this model could not be compared with other sources because there was no predicted protein folding rate change presented in the literature. Hence, only a comparison of the FCMLM deep learning algorithm, BP ANN and Multivariate Regression was done here.

Table.3. Comparison of different methods for predicting seed purity

Model	MSE	Correlation Coefficients
FCM-LM	0.0014	0.9
BP-ANN	0.0221	0.52
Multivariate Regression	0.0062	0.79

From Table.3, the mean-square error (MSE) of FCMLM deep learning algorithm is the smallest and the correlation coefficient of FCMLM deep learning algorithm is the largest among all methods. In these two numerical tests, the correlation coefficients based on the FCMLM deep learning algorithm were both high. Thus, on the one hand, it can be thought that the predictive ability of this model is very strong. However, on the other hand, there were some overfitting phenomena because of the limited number of samples used, which made the accuracy of the predictions unrealistically high. In the future, with an increase of available biological data, the generalization ability of the model will be further verified.

8. CONCLUSION

In this study, the protein folding rates were predicted by FCM and deep learning techniques. The experimental results show that the FCMLM deep learning algorithm is better than other algorithms for its high prediction. The FCM based on GA can extract the features from the original data, undertake qualitative analysis, and initialize the deep neural network effectively. The LM algorithm can adjust network parameters to improve the accuracy of prediction rapidly. The FCMLM deep learning algorithm combines qualitative analysis with quantitative calculation and gives the explanatory ability and high learning efficiency to neural network making the results more realistic. In future research, this model will be consummated and applied in bioinformatics.

REFERENCES

[1] B.S. Anami, Naveen N. Malvade and Surendra Palaiah, “Automated Recognition and Classification of Adulteration Levels from Bulk Paddy Grain Samples”, *Information Processing in Agriculture*, Vol. 6, No. 1, pp. 47-60, 2019.

[2] W. Liu and Feifei Chen, “Rice Seed Purity Identification Technology using Hyperspectral Image with LASSO Logistic Regression Model”, *Sensors*, Vol. 21, No. 13, pp. 1-14, 2021.

[3] Nadia Ansari, Sharmin Sultana Ratri, Afroz Jahan, Muhammad Ashik-E-Rabbani and Anisur Rahman, “Inspection of Paddy Seed Varietal Purity using Machine Vision and Multivariate Analysis”, *Journal of Agriculture and Food Research*, Vol. 3, pp. 1-12, 2021.

[4] Xiaolong Li, Zhenni He, Fei Liu and Rongqin Chen, “Fast Identification of Soybean Seed Varieties using Laser-Induced Breakdown Spectroscopy Combined with Convolutional Neural Network”, *Frontiers in Plant Science*, Vol. 12, No. 2, pp. 1-12, 2021.

[5] P. Lin, X.L. Li, Y.M. Chen and Y. He, “A Deep Convolutional Neural Network Architecture for Boosting Image Discrimination Accuracy of Rice Species”, *Food and Bioprocess Technology*, Vol. 11, No. 4, pp. 765-773, 2018.

[6] J. Chen, Xudong Gao, Jia Rong and Xiaoguang Gao, “The Dynamic Extensions of Fuzzy Grey Cognitive Maps”, *IEEE Access*, Vol. 9, pp. 98665-98678, 2021.

[7] Taha Mansouri, Ahad ZareRavasan and Amir Ashrafi, “A Learning Fuzzy Cognitive Map (LFCM) Approach to Predict Student Performance”, *Journal of Information Technology Education: Research*, Vol. 20, pp. 221-243, 2021.

[8] A. Ali, Samreen Naeem, Sidra Rafique, Farrukh Jamal, Christophe Chesneau and Sania Anam, “Machine Learning Approach for the Classification of Corn Seed using Hybrid Features”, *International Journal of Food Properties*, Vol. 23, No. 1, pp. 1110-1124, 2020.

[9] Iqbal H. Sarker, “Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions”, *SN Computer Science*, Vol. 2, No. 6, pp. 1-20, 2021.

[10] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng and Jun Zhou, “A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects”, *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 68, No. 3, pp. 1-13, 2021.

[11] Z. Luan and Yan Yang, “Sunflower Seed Sorting based on Convolutional Neural Network”, *Proceedings of 11th International Conference on Graphics and Image Processing*, pp.1-12, 2020.

[12] K. Tatsumi and X. Mengxue, “Prediction of Plant-Level Tomato Biomass and Yield using Machine Learning with Unmanned Aerial Vehicle Imagery”, *Proceedings of International Conference on Graphics and Image Processing*, pp. 1-8, 2021.

[13] S.J. Symons and R.G. Fulcher, “Determination of Wheat Kernel Morphological Variation by Digital Image Analysis: I. Variation in Eastern Canadian Milling Quality Wheats”, *Journal of Cereal Science*, Vol. 8, No. 3, pp. 211-218, 1988.

[14] Jared Taylor, Chien-Ping Chiou and Leonard J. Bond, “A Methodology for Sorting Haploid and Diploid Corn Seed using Terahertz Time Domain Spectroscopy and Machine Learning”, *AIP Conference Proceedings*, Vol. 2102, No. 1, pp. 1-6, 2019.