International Journal of Electrical and Computer Engineering (IJECE)
Vol. 13, No. 5, October 2023, pp. 5770~5781
ISSN: 2088-8708, DOI: 10.11591/ijece.v13i5.pp5770-5781

Analysis and prediction of seed quality using machine learning

Raghavendra Srinivasaiah[1], Meenakshi[2], Ravikumar Hodikehosahally Channegowda[3], Santosh Kumar Jankatti[4]

[1]Department of Computer Science and Engineering, CHRIST Deemed to be University, Bengaluru, India
[2]Department of Computer Science and Engineering, RNS Institute of Technology, Bengaluru, India
[3]Department of Electronics and Communication, Dayananda Sagar Academy of Technology and Management, Bengaluru, India
[4]Department of Computer Science and Engineering, Dayananda Sagar University, Bengaluru, India

Article Info

Article history:

Received Nov 12, 2022
Revised Feb 2, 2023
Accepted Mar 9, 2023

Keywords:

Agriculture
Classification
Convolution neural network
Prediction
Seed quality

ABSTRACT

The mainstay of the economy has always been agriculture, and the majority of tasks are still carried out without the use of modern technology. Currently, the ability of human intelligence to forecast seed quality is used. Because it lacks a validation method, the existing seed prediction analysis is ineffective. Here, we have tried to create a prediction model that uses machine learning algorithms to forecast seed quality, leading to high crop yield and high-quality harvests. For precise seed categorization, this model was created using convolutional neural networks and trained using the seed dataset. Using data that can be used to forecast the future, this model is used to learn about whether the seeds are of premium quality, standard quality, or regular quality. While testing data are employed in the algorithm's predictive analytics, training data and validation data are used for categorization reasons. Thus, by examining the training accuracy of the convolution neural network (CNN) model and the prediction accuracy of the algorithm, the project's primary goal is to develop the best method for the more accurate prediction of seed quality.

Corresponding Author:

Raghavendra Srinivasaiah
Department of Computer Science and Engineering, CHRIST Deemed to be University
Kanmanike, Kumbalgodu, Mysore Road, Bangalore-560074
Email: raghav.trg@gmail.com

1. INTRODUCTION

The quality of the seeds that are spread to new growth is one of the crucial factors in agriculture to ensure that the yields are excellent. Unfortunately, some crops fail to generate the necessary quantity of seeds that may sprout, and between 35% and 40% of the seeds are discarded. Later improving the properties of the seeds generated at the time of their growth and aggregating using different techniques, the ensuing phase is to handpick the seeds before they are carried out to the field [1]–[3].

Machine vision and near-infrared spectroscopy are two non-critical testing methods that have been utilized extensively in the last 20 years to anticipate the quality of foods and agricultural products. These non-destructive techniques are useful because they encourage simultaneous assessment of food's chemical as well as physical data without causing the ingredient any harm. The main advantage of these techniques is that they consume less time and low cost and these utilizations provide good benefits to food manufacturing. In general, non-destructive approaches have a promising future for evaluating the nutritional value of food and crops, and the development of more accurate and efficient imaging systems [4], [5].

The prediction algorithm is developed to help people to maintain a validated procedure and structure in determining seed quality by evaluating seeds they want to grow for production or research purposes in the agriculture industry. Because in today's world, we are still underdeveloped due to a lack of technological advancement. People are practicing agriculture activities and businesses on human intelligence and wishful thinking. Today we have a validation mechanism to grow bigger in terms of technology, power, resources, capital, mass production, and good-quality products.

Seeds are used as an important source of food to serve the human population and they also act as a preliminary material for the growing of crops. The harvest of the crop is mainly dependent on the quality of the seed and slightly dependent on the environment. So, the prediction of seed germination is a significant and the most important task. This is also required to improve the efficiency of the food chain and measure the performance of a variety of seeds [6]. To satisfy the need of the growing population, global crop production should be doubled by 2050 [7].

During the selection of good quality seeds, the evaluation of the seed, vigor is very important and has a lot of insinuations. Right from the 1960s the properties of the seed vigor to determine the characteristics of the seeds have been recognized worldwide. But now the advent of modern technologies such as biotechnology, biophysics, seed image analysis, and molecular biology resulted in reducing the use of a traditional technique such as the vigor test since they are clumsy and consumes more time [8]. In addition to this most of the seed test developed by the International Seed Testing Association uses manual testing that employs consistent techniques and it varies from one crop to another.

The world was in its greatest time of difficulties due to the pandemic coronavirus disease 2019 (COVID-19) during the 2019 to 2021. People are struggling to survive; they do not have enough food to eat, and money to meet their necessities of life, countries are on the verge of poverty because of lockdowns and zero businesses. All of this is happening because of the virus which was spread from person to person by touch, sneezing, and air. Therefore, this is the time when it is very necessary to make sure that whatever production occurs, it should occur with maximum productivity and money-making. The public and private agricultural firms and government should take the responsibility to carry out this process in the most effective way to produce high-quality crops with much higher production to stabilize the country’s economy and to help the poor, the farmers getting what they deserve in this cruel time. This could only be achieved when there will be good quality seeds. However, to reduce the impact of the pandemic, several government agencies and private firms are allowed a limited number of agricultural activities to be resumed post the number of new cases of COVID-19 will drop below a certain level. Now the situation is coming to normal and all agricultural activities resuming to normal as it was before the pandemic.

The current approach to agricultural operations is undeveloped for a number of reasons, but the absence of contemporary technology is the main one. Currently, the ability of human intelligence to forecast seed quality is used. Because it lacks a validation method, the existing seed prediction analysis is ineffective. The final objective is to develop an algorithm employing machine learning methods. With the use of machine learning algorithms, we are attempting to create a predictive model to forecast seed quality, which will lead to high crop yield and high-quality harvests.

2. LITERATURE REVIEW

Advances in seed science have been made possible by the study of optical sensors in conjunction with machine learning techniques. Because of these developments, reliable strategies have made it easier to make decisions in the seed sector on the marketing of seed lots. This study introduces a brand-new method for categorizing seed quality. To predict seed germination and vigour, classifier models were created utilizing Fourier transform near-infrared (FT-NIR) spectroscopy and X-ray imaging methods. As a model species, a forage grass (Urochloa Brizantha) was chosen. The radiographic images and FT-NIR spectroscopic data were collected from individual seeds, and the models were built using the linear discriminant analysis (LDA), partial least squares discriminant analysis (PLS-DA), random forest (RF), naïve Bayes (NB), and support vector machine with radial basis (SVM-r) kernel techniques. The models’ individual accuracy for seed categorization (germination) using FT-NIR and X-ray data was 82% and 90%, respectively. Using FT-NIR and X-ray data, the models’ accuracy for seed prediction (seed vigour) was 61% and 68%, respectively. In terms of accuracy, the classification model’s performance using the combined FT-NIR and X-ray data achieved 85% for predicting germination and 62% for predicting seed vigour. Overall, the models created utilizing NIR spectra and X-ray imaging data were created using machine learning techniques because they are rapid, non-destructive, and accurate at determining a seed’s ability to germinate. The LDA method and the utilization of X-ray data shown significant promise as a practical option to aid in quality categorization [9].

An extensive study based on multispectral imaging applications for seed phenotyping and quality monitoring is carried out to comprehend the potential of multispectral imaging with various chemometrics algorithms in describing physicochemical quality attributes, forecasting physiological parameters, variety identification, and classification, and detection of damages, defect, pest infestation, and seed health. Researchers have worked together to create quick, precise, and affordable spectral systems that will be used in the seed and grain industries as a result of this amazing capabilities. Modern multispectral imaging and pertinent multivariate chemometric analysis have been effectively utilized, not only for food safety and quality control purposes but also in order to address significant research difficulties related to seed science and technology. This is made possible by the ability to acquire three-dimensional information across an array of the electromagnetic spectrum [10].

Based on the high value of prediction accuracy for soybean viability, which is near 100%, the research of FT-NIR spectral analysis using the PLS-DA approach that employs all factors or the selected variables showed good performance. This study make use of both the viable and synthetically aged soybean seeds. To differentiate between viable and non-viable soybean seeds, the FT-NIR spectra of soybean seeds were gathered and examined using PLS-DA. Additionally, the PLS-DA and the variable significance in the projection (VIP) approach for variable selection were both used [11].

An effective method for evaluating the physiological quality of hybrid maize seeds and forecasting the seed vigour of the samples is the combination of the attenuated total reflection Fourier transform infrared (ATR-FTIR) and chemometrics. Additionally, this work offers a theoretical framework for maize producers to optimize their genetics in order to achieve higher physiological seed quality. The rapid ageing of high-vigor seeds demonstrates the link between the chemicals and seed vigour by causing minor changes in biochemical composition during stress. Low-vigor seeds are more susceptible to stress, and this sensitivity is associated with lower lipid and protein content and higher levels of amino acids, carbohydrates, and phosphorus. High-vigor seeds have higher peaks for amino acids and phosphorus compounds [12].

With whole-seed samples from 240 individual plants, several kinds of seed samples and their use on individual plants were demonstrated. A straightforward and affordable technique, ZX-50 NIR analysis of seed protein and oil content, was used in this investigation. Research and soy breeding would benefit from using this technique. Using a variety of cultivars with different seed sizes and a broad range of variations in the objective qualities, an appropriate bias formulation should be created in advance in order to produce predictions that can be trusted, according to the study’s findings. It is especially important when the materials being studied include seeds that are different from those used by the manufacturer to create the original product [13].

The vitality of rice seeds was assessed using a variety of pre-processing techniques and classification methods. Three pre-processing techniques were employed to reduce the impact of discrepancies in the spectrum data brought on by elements such as random noise, light scattering, and sample roughness. The regions-of-interest (ROI) of the hyperspectral picture was used to extract the spectral data. While lowering processing costs, the set pair analysis (SPA) method was utilized to find the best wavelengths for seed vitality. In comparison to the whole wavelength classification model, the chosen wavelengths 9, 8, 11, and 6 for raw data, Savitzky–Golay (S-G), SG-D1, and multiplicative scattering correction (MSC) pre-processed data, respectively-may greatly lessen the data processing load. In order to determine the viability of seeds from three distinct years and to discriminate between non-viable and viable seeds from three different seeds, multivariate models were developed using these ideal wavelengths. Using PLS-DA, least-squares support-vector machines (LS-SVM), and extreme learning machine (ELM) with selected wavelengths and SG pre-processing, the classification accuracy for the seed vitality of three distinct years was found to be 87.5%, 93.33%, and 93.67%, respectively. With the use of raw data, chosen wavelengths, and the LS-SVM model, it was possible to distinguish between non-viable and viable seeds from various years with a classification accuracy of 94.38%. In this study, the most effective technique for differentiating between non-viable and viable seeds was created [14].

Using chemometric techniques and hyperspectral imaging technologies, five wheat seed variants were deployed and tested in an effort to categorize them. Two exploratory techniques for categorization analysis, principal component analysis (PCA) and linear discriminant analysis (LDA), were looked into during the procedure. The latter one in particular showed a higher degree of differentiating capacity and the ability to properly identify wheat grains. The results of this study demonstrate that the performance of three models based on various pre-treatment techniques did not significantly improve when compared to models based on raw spectra data, showing that pre-processing techniques were not always effective in achieving classification recognition. The classification outcome of the RF feature extraction method is likewise satisfactory, demonstrating the viability of feature variable selection. The model’s efficiency on feature wavelengths also implies that it has higher promise for building a portable multispectral sensor to quickly and non-destructively distinguish wheat seeds. The results show that wheat grain categorization performance still has to be enhanced. The creation of the pseudo-color visualization map, which also helps with the construction of an online and extensive detecting system to check the purity of various wheat types, allows users to easily and intuitively perceive the specific variance of each sample. The properties of both spectrum and spatial information, along with high throughput, all point to the possibility that HSI with chemometrics might detect and identify large-scale grain seeds in the current seed business [15].

This work focuses on a short-wave infrared (SWIR) hyperspectral imaging system that was tuned and used for the differentiation of soybean seed viability using NIR. The tool's capacity to perform bulk measurements and its simplicity of integration with an automated seed separation procedure make it helpful for non-destructive viability measurement. In this work, a kernel-based image processing algorithm was employed to classify the entire seed as viable or nonviable rather than recognizing specific pixels of hyperspectral images. The experimental outcomes of this work demonstrate that the PLS-DA-VIP model developed using just a few wavebands can accurately (>95% accuracy) forecast the viability of soybean seeds [16].

An experimental comparison of some of the major factors determining the speed and accuracy of modern object detectors is implemented. According to the findings of this, practitioners will be able to select an acceptable strategy when deploying object detection in the real world. This includes some novel strategies for increasing speed without reducing accuracy, such as using notably fewer proposals than is conventional for faster R-CNN [17].

The advantages of big data in the restricted data realm can be obtained by artificially inflating datasets using the techniques described in this article [18]. In this work, several augmentations that can be broadly categorized as either a data warping or an oversampling approach have been presented. These search algorithms using data bending and oversampling techniques have a lot of potential. Deep neural networks' (DNN) layered design offers a variety of possibilities for data augmentation. The DisturbLabel approach is even utilized in the output layer, but the majority of the augmentations do work in the input layer and some are generated from hidden layer representations. This is unable to eliminate every bias seen in a short dataset. Access to vast data typically makes overfitting less of a problem. This work enhances data to prevent overfitting by transforming small datasets into large data-like properties.

The classification of three stages of plant growth as well as soil on various accessions of two species of red clover and alfalfa were addressed using a full image processing and machine learning pipeline. To take into account previous knowledge about the sequence in which the various phases of growth take place, several methodologies were contrasted. Their suggested convolutional neural network-long short-term memory (CNN-LSTM) model had the best classification performance on these sorts of pictures, achieving 90% detection accuracy with the aid of a denoising method that included the ontological order during the development phases. The models used in this work were developed and evaluated using a variety of genotypes from two species of red clover and alfalfa. The findings that have been shown indicate that the trained model is resilient on some genotypes, but they do not imply that the model is robust on other genotypes or other species. To strengthen the robustness of models, one might either add additional real data from other genotypes or employ data augmentation to artificially increase the data variability in the training database based on potential priors on the predicted morphological plasticity of the species. These findings expand the method to several species relevant to agriculture and offer a library of trained networks. Following the general methods described in this work to evaluate the deep learning models, these additions might be evaluated with ease. According to another potential route of the research, plants cannot be monitored at night using traditional standard red, green and blue (RGB) photos, and certain missing occurrences might cause the assessment of the seedlings' developmental stages to change. Nighttime occurrences might be seen using low-cost lidar cameras. Additionally, to estimate the time for the potentially missing information, Bayesian techniques like Gaussian processes might be utilized [19].

In an effort to increase crop yield, a predictive model for identifying seed classes using machine learning algorithms is being developed. A machine learning technique will be utilized in this study to produce learning models that can forecast the future, build realistic simulations, find patterns in data, and categorize incoming data. For intricate interactions between inputs and outputs, an artificial neural network (ANN) is used to model the data and uncover patterns. This tries to comprehend the machine-learning method using neural networks and construct models that foretell machine-learning seed classes. The created model is tested on a seed dataset, and seed classes are anticipated using the developed model. Finally, the model built is used to identify and rank the parameters affecting seed classifications [20].

Image analysis has a wide range of applications in the evaluation of several physiological and morphological traits of the seed with a more thorough perception. It relies on the gathering of numerical information, such as the color, size, and shape of seeds and seedlings, from a taken image, and then the subsequent processing of that information using the appropriate computer software. The positive aspects of image analysis over more traditional approaches include quick assessments, cost-effectiveness, automated nature, and a user-friendly working environment. As it is faster, more accurate, and offers close examination of seeds and sprouting seedlings, image analysis systems have been demonstrated to be extremely useful methods for seed-related studies [21].

According to the topic of the research and the data's availability, the author highlights the chosen publications' utilization of various attributes is taken into consideration. Every article examines yield prediction using machine learning; however, the characteristics vary. Scale, geological location, and crop also vary among the experiments. The dataset’s availability and the research's objectives will influence the characteristics that are chosen. Additionally, this study notes that models with more characteristics did not necessarily have the highest yield prediction performance. Models with various numbers of features should be evaluated to determine which one performs the best. Various research has employed a variety of algorithms. The findings indicate that it is impossible to pinpoint the optimal model, but they do demonstrate that some machine-learning models are employed more frequently than others. The most popular models are gradient-boosting trees, neural networks, linear regression, and random forest. In the majority of the research, different machine-learning models were tested to see which one made the best predictions. Since neural networks are the most often used method, they also intended to ascertain how much deep learning algorithms were utilized for agricultural yield prediction. They extracted and synthesized the used methods after finding 30 articles that used deep learning. According to this analysis, the most popular deep learning algorithms are the CNN, LSTM, and DNN algorithms [22].

To categorize weedy plants among the seeds, a study on seed classification was conducted. The seed datasets were entered, along with the seed image. In the preprocessing stage, the seed picture was used as input. Through the use of the ID3 algorithm, a seed’s features were compared to samples of other seeds to identify unwanted seeds. One of the causes of unsuccessful crop production is the choosing of improperly farmed soil. The data set of samples in this study provides information about crop growth in soil, which aids in choosing the right soil for seeds [23].

The interactive machine-learning approach for classifying soybean seeds based on their appearance has been shown to be very accurate. Using this method, damaged seeds are efficiently identified, and seedling vigour ratings are assigned to each seedling. Based on data produced by the Ilastik software, the usage of LDA, RF, and SVM algorithms is advised for categorizing soybean seeds and seedlings. Low physiological quality is seen in soybean seeds with alterations in the deterioration of chlorophyll, fungal staining, and mechanical damage [24].

To classify defects in maize seedlings, CNN and transfer learning are used. The effectiveness of CNNs in categorizing seed defects is demonstrated through experiments. In evaluating maize seed defects, CNNs outperformed machine learning techniques, and the accuracy of the model grew as the network’s depth rose. One of the metrics for evaluating the seed’s quality is the flaw in its outward look. Although CNN may be used with multispectral or hyperspectral pictures, in this study they have only used it for RGB images. The use of multispectral pictures improves the model’s generalizability and applicability by enabling recognition of both the phenotypic traits of seeds and distinct types [25].

Research that demonstrates the capacity and promise of machine vision with the well-trained multilayer neural network classifiers for the recognition of uneven rice grain samples’ forms, sizes and varietal types that are cultivated in the various agro-environmental zones across the nation is reviewed. They employed Weka classification techniques, including the function, Bayes, meta, and lazy approaches, to categorize the seeds. They employed the classifiers logistics, sequential minimal optimization (SMO), naive Bayes updateable, multilayer perceptron (MLP), naive Bayes, Bayes Net, and classifier multi-class from these approaches. The categorization of seeds, according to this study, may be done using three distinct cross-validation folds, namely 10-fold, 5-fold, and 2-fold, as well as a training set approach. After analyzing the data, researchers attempt to deduce that, with the exception of the MLP classifier, which yields the greatest accuracy score of 97.6% using 5-fold cross-validation, the overall performance measurements decline as they lower the fold value. They employed the training set approach and K-Fold cross-validations to gauge performance. Of all the Weka classifiers, MLP has the best performance, with a 97.6% accuracy rate using 5-fold cross-validation. When using the training set approach, MLP gets the best accuracy value, which is 99.5%, while logistics gives the second-highest accuracy score, which is 98.6%. Finally, they discovered that the training set approach performs the classification process with more accuracy than cross-validation. According to the results of this study, the unsupervised artificial neural network performs better than the supervised artificial neural networks, which only achieve 73% accuracy [26].

The study focuses on the selection of pepper seeds using a single predictor for each of two variables (a* and single-kernel weight), which have both benefits and drawbacks in that they can increase germination rates while lowering selection rates. One characteristic from each model could not concurrently meet two crucial requirements: attaining a high selection rate while boosting germination percentage. The purpose of this study was to build a model based on binary logistic regression to predict whether a seed will germinate. It also employed the binary logistic regression network classifier to identify the best selection strategy. The germination of pepper seeds was more accurately predicted by MLP and binary logistic regression models than by single-feature models. The multilayer perceptron neural network, with 15 characteristics selected as covariates, was shown to be the best model through comparisons of all models. The model stability was 99.4%, the selection rate reached 90%, and the germination rate increased from the initial 59.3% to 79.1% [27].

Consideration is given to a demonstration that uses 3D points from terrestrial Lidar scanning and deep learning to segment each corn using regional growth methodologies. In order to train the faster R-CNN to recognize stems, a total of 10,784 compressed pictures from 337 distinct samples of maize were employed. To evaluate the stem identification capability, three locations at the same development stage with various planting densities were utilized. Further mapping into 3D points was done for these tested stems. The findings demonstrated how effective the faster R-CNN based technique is in identifying stem anchors in 2D views derived from 3D Lidar pictures. Individual maize with recognized stem seed sites may be precisely segmented using the regional growth approach. Despite some false positive and false negative mistakes, the better accuracy with r, p, and F of more than 90% can greatly decrease the burden to obtain 100% accurate results using just manual approaches. Overall, the segmented height of maize showed a strong correlation to the manually measured value, proving our approach can accurately estimate the height of individual corn. The regional growth technique algorithm did not have parameters; therefore, the Faster R-CNN was able to detect the seed sites, making the suggested method non-parametric and usable in various field scenarios [28].

The researchers work on a detailed study of crop prediction. Numerous machine-learning approaches have been used to estimate agricultural productivity, according to literary works. Additionally, root mean square error and other performance measures for machine learning algorithms are investigated. The influence of big data approaches on agricultural production prediction will be investigated in addition to machine learning algorithms for prediction. For the same, a conceptual strategy is recommended. The suggested strategy is being used [29].

A framework was created to assess several limits to determine the dirt's quality. Depending on the result, the crops were suggested based on the data gathered during the mining process. The approach uses supervised machine learning more precisely and effectively to propose appropriate harvests. The ranchers are left to pick the crop to plant, but the system maintains track of the correct harvests based on the soil [30].

From the literature survey, it is evident that there is not much work related to the prediction of seed quality, and the results achieved are not found to be satisfactory. There are not many readily available datasets. So, in our research work, we made created our own datasets consisting of Bitter gourd, Brinjal, and Calabash seeds divided into three categories premium, standard, and regular. In our research work, we make an attempt to analyze and predict the quality of the seeds using machine learning techniques.

3. PROPOSED DESIGN AND ANALYSIS

The suggested model is founded on the design framework depicted in Figure 1. The experimental and analytical work completed in this project has been categorized into five stages namely: data preparation, image processing, building CNN, training CNN, and analysis. Each of the stage is explained in detail below.

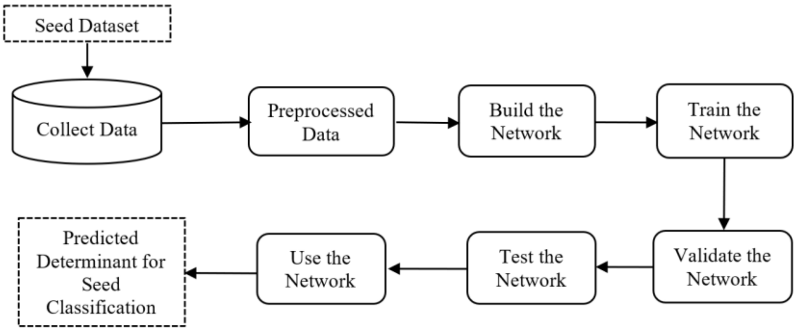

Figure 1. Proposed framework for predicting the seed quality

3.1. Stage 1: data preparation

To generate our dataset, we have used three varieties of seeds i.e., bitter gourd seeds (*Momordica Charantia*), Calabash seeds (*Lagenaria Siceraria*), and Brinjal seeds (*Solanum Melongena*). As shown in Figure 2, we have included three qualities of each variety i.e., premium quality, standard quality, and regular quality of seeds in the dataset, and categorized them into 9 classes based on the variety and quality they exhibit. The images of seeds have been clicked with a high-quality camera and generated a high-quality dataset with sharp detailing. The dataset has been captured with all the possible orientations of the seeds so that the model learns all the different positions of the seeds in the cluster.

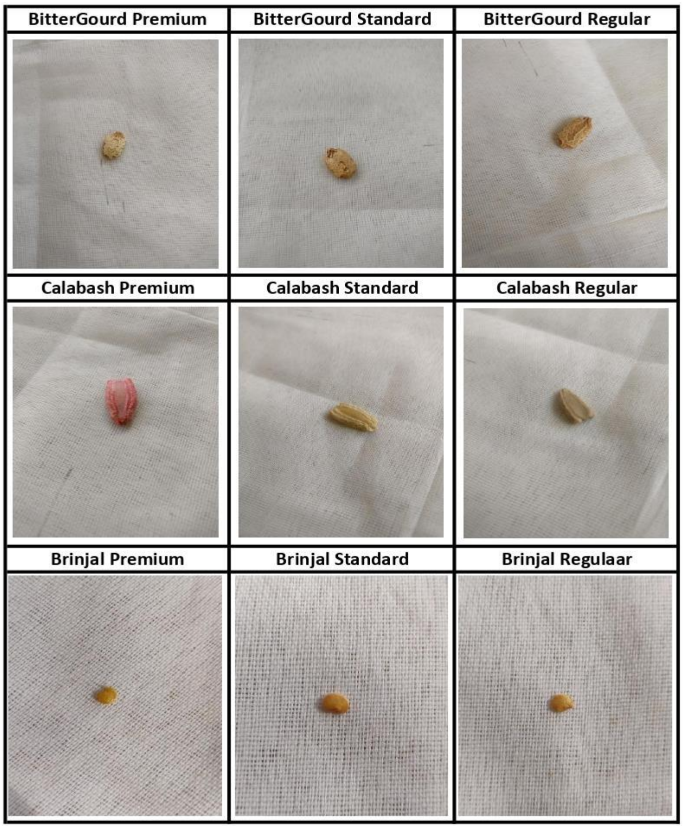

Figure 2. Dataset preview

The dataset comprises 288 distinct images with 198 images in training data, 36 images in invalidation data, and 54 images in testing data respectively. Training data consists of 22 images per class consuming 68% of the dataset, validation data consists of 4 images per class consuming 13% of the dataset, and testing data consist of 6 images per class consuming 19% of the dataset.

3.2. Stage 2: image preprocessing

The dataset has to be cleaned up before being loaded into the CNN. The library function offered by Keras was used to expand the dataset, and all of the photos were processed to a consistent 200*200 resolution. Beyond this resolution, the input form will give more precision but at the expense of speed. This resolution keeps the distinctive features of seed discrimination, making the machine simple to use. There are several parameters and characteristics that have been allocated, including rescaled, rotation range, width shift range, height shift range, horizontal flip, vertical flip, shear range, zoom range, and fill mode. The CNN may now be loaded with all of the photos because they have consistent properties following the dataset preparation. The dataset is divided into three sections, as shown in Figure 3.

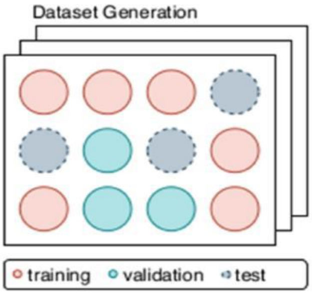

Figure 3. Dataset distribution

3.3. Stage 3: building CNN

To Build a CNN without any loss of seed quality and seed variety information, we made use of the available CNN model as the base model. Layers and activation functions have been modified with appropriate parameters in order to achieve the desired model. The framework of the CNN model and its attributes is shown in Table 1. It is best to adopt a sequential approach for a simple stack of layers where each layer has precisely one input tensor and one output tensor.

Table 1. CNN model framework

	Details	Remarks
Number of convolution layer	3	Feature detectors: 32, 64, 64 Input shape: 200*200 Activation function: ReLU
Number of pooling layers	3	Use of Max-Pooling
Number of dense layers	2	Feature detectors: 512, 9 Activation function: ReLU, Softmax
Number of neurons in 1st dense layer	512	Selected with multiple experiments
Number of neurons in 2nd dense layer	9	Since, there are 9 classes in dataset
Flatten layer	-	Before the Hidden Layer
Optimizer	RMSprop	-
Loss function	Categorical Cross Entropy	-
Evaluation metric	Accuracy	Training accuracy and loss
Output class	9	BitterGourd_premium, BitterGourd_standard, BitterGourd_regular Calabash_premium, Calabash _standard, Calabash _regular Brinjal_premium, Brinjal _standard, Brinjal _regular

3.3.1. Convolution and pooling layer

This model is made up of a 2D matrix with 3 convolutional layers, each followed by a max-pooling layer. In this layer, feature detectors scan the image extensively and quickly, looking for specific features. To achieve spatial invariance, the pooling function decreases the resolution of the feature maps. To remove the dominant features from the feature map, the max-pooling function is used. In the corresponding 3 convolutional layers, I have used 32, 64, and 64 feature detectors (nodes) and initialized these layers to accept images of 200×200 resolution and depth of three RGB color channels with rectified linear unit (ReLU) as their activation function. The 3×3 matrix's feature detector slides over the image and searches for features in it diligently. In max-pooling layers, a 2×2 pool size matrix rolls over the function map with a stride of 2 and collects the full values, which are the image's respective major characteristics.

3.3.2. ReLU activation function

ReLU rectifies the linearities in the non-linear images. ReLU boosts up the training process to converge to accurate predictions. The negative values are pruned by the rectifier function and only the non-linear details are preserved. When the input is positive, it will directly give the output as positive otherwise zero.

3.3.3. Flatten layer

After passing through the convolution and max-pooling layers, the images are in a 2D pooled function map. However, the images should be loaded into a series of artificial neurons, with the input represented as a 1D single vector representing the dominant features of the input image. This is accomplished by flattening all of the feature maps that have been pooled. Finally, the flattened array of input images is ready to be loaded into the dense layer.

3.3.4. Dense layer

This model consists of 2 dense layers. The first dense layer is assigned to one fully connected layer with 512 hidden neurons (nodes) with ReLU as their activation function. The second dense layer (output layer) consists of 9 nodes as there are 9 classes in the dataset with softmax as their activation function. The flattened data is passed through the number of hidden neurons and the predictive probability is passed to the final output layer.

3.3.5. Softmax activation function

The activation function known as softmax scales numbers into probabilities. A softmax produces a vector containing probability for each potential result. For all conceivable outcomes or classes, the probabilities in the vector add up to one. A softmax activation function with an output in the range of 0-1 is used to activate the output layer. In our model, this function divides the input picture into 9 separate output classes. The softmax function with any input *x* is calculated by using (1):

$$f(X_i) = \frac{exp(x_i)}{\sum_j \; exp(x_j)} \tag{1}$$

where, X is vector of input to softmax function *f* and *exp(x*$_i$*)* is the exponential function employed on x_i.

3.4. Stage 4: compiling and training CNN

The compilation of the CNN model is based on the type of loss used, specifying the optimizer, and evaluating the metrics. The training of the CNN model is based on a training dataset validated by a validation dataset with appropriate parameters. Training accuracy depends on how well the model is trained. The more would be the number of epochs the higher would be the training accuracy i.e., the better would be the trained CNN model. Accuracy is inversely proportional to loss. To determine the perfectly trained model the accuracy should be equivalent to constant after quiet epochs.

3.4.1. Loss function

Every iteration of the training and validation process results in a loss. The loss cannot be overlooked because it is a critical entity in determining the uncertainty between the expected and actual values. To measure the errors and loss in the training of our multi-class classification approach, we have assigned this model the “categorical cross-entropy” function. The error is measured and then propagated backward across the network, updating the previous weights as shown in (2). The optimizer selects a learning rate, propagates errors, and updates the weights in each neuron. This contributes to the training model’s accuracy. Categorical cross-entropy (CCE) loss is defined as (2):

$$ECC = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}(\, p_{ic}\log(y_{ic})\,) \tag{2}$$

where, *N* is number of training data, *y* is output, and p_{ic} is binary indicator function that validates the accuracy of i^{th} training pattern to c^{th} category.

3.4.2. Optimizer

RMSprop optimizer, which is a component of this model, limits the oscillations in the vertical plane. so that we may speed up learning and the algorithm can converge more quickly in the horizontal direction. This optimizer normalizes the gradient using a moving average of squared gradients, which balances the step size by reducing steps for big gradients to prevent explosions and increasing steps for small gradients to prevent disappearing. It is optimized with 10 raises to the power minus four units i.e., RMSprop (*lr=1e-4*).

3.4.3. Evaluation metrics

A predictive model’s effectiveness is measured by an evaluation metric. A model is generally trained on a dataset, then used to predict values on a holdout dataset that was not utilized during training. The predictions are then compared to the predicted values in the holdout dataset. To enhance the functionality of the system, the CNN training should be assessed using a certain measure. To evaluate the CNN model, this model is given an accuracy measure.

3.5. Stage 5: predictive analytics on CNN

In this project, the testing dataset will undergo predictive analytics by evaluating the maximum predictive score when passed through the trained CNN model to predict the correct variety and quality of individual seeds in the testing dataset and represent the output images with predicted quality and variety. Later, these prediction values will undergo summation per category to find the average of all classes defined as the prediction accuracy of the algorithm. The link between the predictor variables and the independent variables is statistically modeled using a linear method. For the forecasting of objects, predictive analytics uses (3) of linear regression. That is:

$$Y = C_1X_1 + \; C_2X_2 + \; C_3X_3 + - - - - - - - - \; + \; C_NX_N \tag{3}$$

where, Y is dependent variable, X is independent variable, C is coefficients. These are basically the weights assigned to the features, based on their importance.

The implementation is divided into four stages as shown below:

a) Dataset implementation: this stage deals with the creation, labelling, and formatting of data.
b) Data pre-processing: this stage deals with data cleaning methods and processes with unique abilities and characteristics.
c) CNN model: this stage deals with the construction, optimization, and deployment of the model onto the training dataset and used a validation dataset to validate the model.
d) Prediction: this stage deals with the accurate prediction of the algorithm with the help of predictive analytics. The prediction accuracy process is carried out manually by calculating the average predictive score of images.

4. RESULTS AND DISCUSSION

The model has been trained successfully after running 100 epochs with 30 steps per epoch. Finally, the training accuracy we obtained for the CNN Model was 97%. On the basis of the training model, predictive analytics has been employed to forecast the predictive score of the seed pictures in the testing dataset. The class position index number is used to forecast the type and grade of seeds after each picture has been assessed with a maximum predictive value under the chosen category.

Seeds with the correct prediction of class and quality are denoted as 1 i.e., correct prediction=1; seeds with the correct prediction of class but the incorrect prediction of quality are denoted as 0.5 i.e., partially correct prediction=0.5; seeds with the incorrect prediction of class and quality are denoted as 0 i.e., incorrect prediction=0. The prediction accuracy which we achieved has been calculated manually through mathematical operations and is equivalent to 64% percent as shown in Table 2. The accuracy achieved by some of the existing methods and the proposed method is shown in the Table 3 and the corresponding chart is shown in Figure 4 and it is clear that the proposed method performs better than some of the existing methods.

Table 2. Prediction accuracy

Category	Seed classes		
	BitterGourd premium	BitterGourd standard	BitterGourd regular
Predictive score	0	0	0.5
	1	0	0.5
	1	0.5	1
	0	1	0
	0.5	1	1
	1	1	1
Average	0.58	0.58	0.67
	Calabash premium	Calabash standard	Calabash regular
Predictive score	1	0	1
	1	0	1
	1	0	0
	0	0.5	1
	0.5	0.5	0.5
	0	0.5	1
Average	0.58	0.25	0.75
	Brinjal premium	Brinjal standard	Brinjal regular
	0	1	0.5
	1	1	0.5
	1	1	0.5
	1	0.5	0.5
	1	1	1
	1	1	0.5
Average	0.83	0.92	0.58
Prediction accuracy	64%		

# Quality prediction prerequisite
# Correct prediction-1; Wrong prediction-0; Partially correct prediction-0.5

Table 3. Comparison of existing methods with proposed methods

Techniques used	Accuracy achieved
FT-NIR spectroscopy	61%
X-ray imaging methods	68%
PLS-DA	87.5%
LS-SVM	93.33%
ELM	93.67%
CNN-LSTM	90%
Proposed method	97%

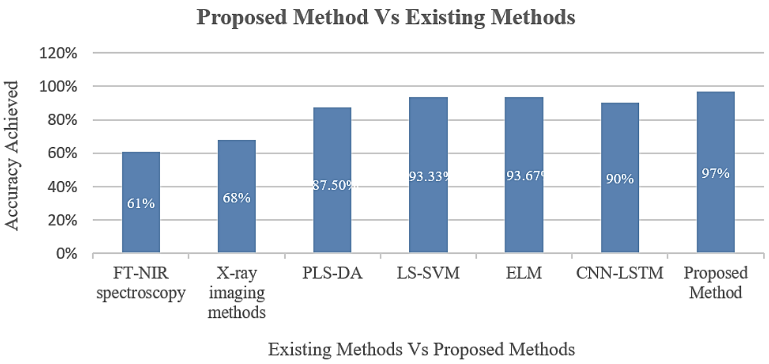

Figure 4. Comparison of existing methods with proposed methods

5. CONCLUSION

As compared to earlier research in this field, which has an accuracy rate of 90% and cumulative prediction accuracy of 68%, this study has an accuracy rate of over 97% and a prediction accuracy of 64%, as detailed in the critical assessment of the literature. The suggested technique made use of CNN architectures to distinguish between different types of seeds with accuracy and to recognize individual seeds with extreme accuracy. For the training CNN model, the models reach above 97% accuracy, while for the testing dataset, they obtain a prediction accuracy of 64%. In comparison to conventional and manual techniques, our model can assist to accelerate the seed health prediction system with reduced error rates and improved performance for bigger experimental data. As a result, we have developed an algorithm that is ideal for accurately predicting seed quality.

ACKNOWLEDGEMENTS

We would like to thank CHRIST Deemed to be University for supporting us in completing this research work.

REFERENCES

[1] A. Kockelmann and U. Meyer, “Seed production and quality,” in *Sugar Beet*, Oxford, UK: Blackwell Publishing Ltd, pp. 89–113.

[2] R. A. George, *Agricultural seed production*. Cabi, 2011.

[3] A. F. Kelly, *Seed production of agricultural crops*. Longman Scientific & Technical, 1988.

[4] Z. Guo *et al.*, “Quantitative detection of apple watercore and soluble solids content by near infrared transmittance spectroscopy,” *Journal of Food Engineering*, vol. 279, Aug. 2020, doi: 10.1016/j.jfoodeng.2020.109955.

[5] H. El-Mesery, H. Mao, and A. Abomohra, “Applications of non-destructive technologies for agricultural and food products quality inspection,” *Sensors*, vol. 19, no. 4, Feb. 2019, doi: 10.3390/s19040846.

[6] T. King *et al.*, “Food safety for food security: Relationship between global megatrends and developments in food safety,” *Trends in Food Science and Technology*, vol. 68, pp. 160–175, Oct. 2017, doi: 10.1016/j.tifs.2017.08.014.

[7] D. K. Ray, N. D. Mueller, P. C. West, and J. A. Foley, “Yield trends are insufficient to double global crop production by 2050,” *PLoS ONE*, vol. 8, no. 6, Jun. 2013, doi: 10.1371/journal.pone.0066428.

[8] J. M. Filho, “Seed vigor testing: an overview of the past, present and future perspective,” *Scientia Agricola*, vol. 72, no. 4, pp. 363–374, Aug. 2015, doi: 10.1590/0103-9016-2015-0007.

[9] A. D. de Medeiros *et al.*, “Machine learning for seed quality classification: an advanced approach using merger data from FT-NIR spectroscopy and x-ray imaging,” *Sensors*, vol. 20, no. 15, Aug. 2020, doi: 10.3390/s20154319.

[10] G. ElMasry, N. Mandour, S. Al-Rejaie, E. Belin, and D. Rousseau, “Recent applications of multispectral imaging in seed phenotyping and quality monitoring-an overview,” *Sensors*, vol. 19, no. 5, Mar. 2019, doi: 10.3390/s19051090.

[11] D. Kusumaningrum, H. Lee, S. Lohumi, C. Mo, M. S. Kim, and B.-K. Cho, “Non-destructive technique for determining the viability of soybean (Glycine max) seeds using FT-NIR spectroscopy,” *Journal of the Science of Food and Agriculture*, vol. 98, no. 5, pp. 1734–1742, Mar. 2018, doi: 10.1002/jsfa.8646.

[12] G. C. Andrade, C. M. Medeiros Coelho, and V. G. Uarrota, “Modelling the vigour of maize seeds submitted to artificial accelerated ageing based on ATR-FTIR data and chemometric tools (PCA, HCA and PLS-DA),” *Heliyon*, vol. 6, no. 2, Feb. 2020, doi: 10.1016/j.heliyon.2020.e03477.

[13] G.-L. Jiang, “Comparison and application of non-destructive NIR evaluations of seed protein and oil content in soybean breeding,” *Agronomy*, vol. 10, no. 1, Jan. 2020, doi: 10.3390/agronomy10010077.

[14] X. He, X. Feng, D. Sun, F. Liu, Y. Bao, and Y. He, “Rapid and nondestructive measurement of rice seed vitality of different years using near-infrared hyperspectral imaging,” *Molecules*, vol. 24, no. 12, Jun. 2019, doi: 10.3390/molecules24122227.

[15] Y. Bao, C. Mi, N. Wu, F. Liu, and Y. He, “Rapid classification of wheat grain varieties using hyperspectral imaging and chemometrics,” *Applied Sciences*, vol. 9, no. 19, Oct. 2019, doi: 10.3390/app9194119.

[16] I. Baek *et al.*, “Rapid measurement of soybean seed viability using kernel-based multispectral image analysis,” *Sensors*, vol. 19, no. 2, Jan. 2019, doi: 10.3390/s19020271.

[17] J. Huang *et al.*, “Speed/accuracy trade-offs for modern convolutional object detectors,” in *2017 IEEE Conference on Computer*

Vision and Pattern Recognition (CVPR), Jul. 2017, pp. 3296–3297, doi: 10.1109/CVPR.2017.351.

[18] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, Dec. 2019, doi: 10.1186/s40537-019-0197-0.

[19] S. Samiei, P. Rasti, J. Ly Vu, J. Buitink, and D. Rousseau, "Deep learning-based detection of seedling development," *Plant Methods*, vol. 16, no. 1, Dec. 2020, doi: 10.1186/s13007-020-00647-9.

[20] T. Tujo, G. Kumar, D. Yitagesu, and B. Girma, "A predictive model to predict seed classes using machine learning," *International journal of engineering research and technology (IJERT)*, vol. 6, pp. 334–344, 2019.

[21] Hemender, S. Sharma, V. Mor, Jitender, and A. Bhuker, "Image analysis: a modern approach to seed quality testing," *Current Journal of Applied Science and Technology*, vol. 27, no. 1, pp. 1–11, Apr. 2018, doi: 10.9734/CJAST/2018/40945.

[22] T. van Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning: A systematic literature review," *Computers and Electronics in Agriculture*, vol. 177, Oct. 2020, doi: 10.1016/j.compag.2020.105709.

[23] N. Nandhini and J. G. Shankar, "Prediction of crop growth using machine learning based on seed," *Ictact journal on soft computing*, vol. 11, no. 1, 2020.

[24] A. D. de Medeiros, N. P. Capobiango, J. M. da Silva, L. J. da Silva, C. B. da Silva, and D. C. F. dos Santos Dias, "Interactive machine learning for soybean seed and seedling quality classification," *Scientific Reports*, vol. 10, no. 1, Jul. 2020, doi: 10.1038/s41598-020-68273-y.

[25] S. Huang, X. Fan, L. Sun, Y. Shen, and X. Suo, "Research on classification method of maize seed defect based on machine vision," *Journal of Sensors*, pp. 1–9, Nov. 2019, doi: 10.1155/2019/2716975.

[26] R. H. Ajaz and L. Hussain, "Seed classification using machine learning techniques," *Seed*, vol. 2, no. 5, pp. 1098–1102, 2015.

[27] K.-L. Tu, L.-J. Li, L.-M. Yang, J.-H. Wang, and Q. Sun, "Selection for high quality pepper seeds by machine vision and classifiers," *Journal of Integrative Agriculture*, vol. 17, no. 9, pp. 1999–2006, Sep. 2018, doi: 10.1016/S2095-3119(18)62031-3.

[28] S. Jin *et al.*, "Deep learning: individual maize segmentation from terrestrial lidar data using faster R-CNN and regional growth algorithms," *Frontiers in Plant Science*, vol. 9, Jun. 2018, doi: 10.3389/fpls.2018.00866.

[29] K. Palanivel and C. Surianarayanan, "An approach for prediction of crop yield using machine learning and big data techniques," *International Journal of Computer Engineering and Technology*, vol. 10, no. 3, Jun. 2019, doi: 10.34218/IJCET.10.3.2019.013.

[30] G. Suresh, A. S. Kumar, S. Lekashri, R. Manikandan, and C. Head, "Efficient crop yield recommendation system using machine learning for digital farming," *International Journal of Modern Agriculture*, vol. 10, no. 1, pp. 906–914, 2021.

BIOGRAPHIES OF AUTHORS

Raghavendra Srinivasaiah is currently working as Associate Professor in the Department of Computer Science and Engineering at CHRIST Deemed to be University, Bangalore. He completed his Ph.D. degree in Computer Science and Engineering from VTU, Belgaum, India in 2017 and has more than 18 years of teaching experience. His interests include data mining, artificial intelligence and big data. He can be contacted at email: raghav.trg@gmail.com.

Meenakshi is currently working in RNSIT Bengaluru which is affiliated to VTU Belagavi in the department of CSE, she secured VTU 8th Rank, she completed masters from VTU, she has 4 years of teaching experience and her area of interest are big data analytics, data mining. She can be contacted at email: meenakshib437@gmail.com.

Ravikumar Hodikehosahally Channegowda completed his Ph.D. from VTU, Belagavi in 2021. He has done his masters in VLSI design and embedded systems from VTU Extension Centre, PESCE, Mandya. His areas of interest are image processing, machine learning, pattern recognition and multimedia concepts. He is currently working as Assistant Professor at Dayananda Sagar Academy of Technology and Management, Bengaluru. He can be contacted at email: raviec40@gmail.com.

Santosh Kumar Jankatti is currently working as Associate Professor in the Department of Computer Science and Engineering at Dayananda Sagar University, Bangalore. He completed his Ph.D. degree in Computer Science and Engineering from VTU, Belgaum, India in 2022 and has more than 11 years of teaching experience and 3 years of IT Industry experience. His interests include data mining, artificial intelligence and big data. He can be contacted at email: sjankatti@gmail.com.