Discover Applied Sciences

Research

Machine learning approaches for grain seed quality assessment: a comparative study of maize seed samples in Malawi

Wisdom Richard Mgomezulu[1] · Moses M. N. Chitete[2] · Beston B. Maonga[3] · Mthakati A. R. Phiri[4]

Received: 19 March 2025 / Accepted: 22 May 2025
Published online: 03 June 2025

Abstract
The study assessed machine and deep learning algorithms' ability to predict and classify the quality of maize grain seed for increased agricultural output. It relied on a dataset of 2460 maize seed samples examined by a KEPHIS ISTA-accredited seed testing facility. The K-NN and Logistic Regression algorithms performed the best in predicting and classifying seed samples, with 100% accuracy, precision, recall, and fi-score. The algorithms found that 46.2% of the grain maize seed was correctly classified as poor-quality seed due to improper handling, and poses a danger to productivity and food security for smallholder farmers. The Deep Learning Convolutional Neural Network presented a 92% accuracy with slight fluctuations, mainly due to the simple and structured nature of the data, which was not in a grid-like or time series format. The study therefore recommends using K-Nearest Neighbor and/or Logistic Regression for grain seed classification when presented with well-structured agricultural data. Still, it also suggests expanding the methodology to other agricultural commodities and implementing seed management measures to prevent low-quality seed distribution. This includes training traders on how to maintain ISTA-required levels of germination, purity, and moisture content in their stores. The study highlights the significance of high-quality seeds for smallholder farmers to improve production and food security.

Article Highlights

- The K-Nearest Neighbor (KNN) and Logistic Regression algorithms achieved perfect classification performance, each reaching 100% accuracy, precision, recall, and F1-score in predicting maize seed quality.
- The deep learning Convolutional Neural Network (CNN) model also performed well, attaining a high accuracy of 92%, however, the results suggest that for simpler data structures, traditional machine learning models may outperform more complex deep learning approaches.
- The research underscores the potential of machine learning models to support seed quality monitoring and classification.

Keywords Seed quality · Machine learning · Deep learning · Productivity · Maize seed

✉ Moses M. N. Chitete, mozzychitete@gmail.com | [1]Alliance for a Green Revolution in Africa, Data & Analytics Unit, West End Towers, 4 th Floor, Wetlands, P.O. Box 66773, Nairobi 00800, Kenya. [2]Centre for Agricultural Research and Development, Lilongwe University of Agriculture and Natural Resources, Box 219, Lilongwe, Malawi. [3]Lilongwe University of Agriculture and Natural Resources. African Centre for Excellence in Agricultural Policy Analysis (ACE II AF-APA), Box 219, Lilongwe, Malawi. [4]Department of Agricultural and Applied Economics, Lilongwe University of Agriculture and Natural Resources, Box 219, Lilongwe, Malawi.

1 Introduction

In recent decades, the growth of seed companies has brought about significant changes to Malawi's seed industry. In the late 1980 s, ADMARC was the only seed seller in the nation, while the Seed Company of Malawi (NSCM) was involved in seed production [16]. Meanwhile, the country has over 25 and 700 seed companies and agro-dealers, respectively. This rapid expansion has created significant regulatory challenges for the Seed Services Unit (SSU), leading to increased instances of substandard seeds reaching farmers. Proliferation of substandard seeds has a potential negative impact on productivity. The current study follows the recent widespread prevalence of fake and low-quality seed reaching farmers, which has affected crop productivity and food security in Malawi [22].

Malawi's economy is largely agro-based, making a well-functioning seed system essential for sustained economic development. According to Haug et al. [14], seed is a fundamental input in crop production. Access to quality seed is therefore crucial for improving household food security, as it carries the genetic potential that determines crop productivity, disease resistance, and tolerance to environmental stresses such as drought. Hunga et al. [16] also emphasize the importance of providing farmers with access to high-quality, improved seed varieties. Moreover, they highlight that the current seed system is inefficient, noting that it is uneconomical for Sub-Saharan African countries, such as Malawi, to continue spending scarce foreign exchange on fertilizers with high import content, as seen in Malawi's Affordable Input Subsidy Program. This underscores the urgent need to invest in the provision of quality seeds to enhance productivity and strengthen food security at the household level. The quality of seed input plays a crucial role in determining agricultural outcomes, influencing crop yield, disease resistance, and environmental stress tolerance. Recent studies have emphasized that access to high-quality seeds is essential for enhancing farm productivity and ensuring food security in developing agricultural economies. The Seed Services Unit (SSU) is mandated to carry out seed quality tests before it reaches the market for farmers' uptake [12]. Physiological seed quality tests are done through DNA fingerprinting and Grow-Out-Trials (GOT). As such, the current study builds on a multi-year study conducted by the Alliance for a Green Revolution in Africa [3], which collected a total of 2460 seed samples and tested at the Kenya Plant Health Inspectorate Service (KEPHIS) lab, an International Seed Testing Association (ISTA) accredited seed testing laboratory for physical purity and germination tests, which follows ISTA rules. The seed samples of maize were collected from seed company warehouses and agrodealer shops across Malawi.

The current seed quality assessment system, while thorough, faces significant challenges. Traditional testing methods, such as DNA fingerprinting and Grow-Out Trials (GOT), conducted at accredited facilities like KEPHIS, are both time-consuming and costly [4]. This has led to a concerning trend where farmers increasingly rely on informal seed sources, potentially compromising crop quality and yield potential.

Classifying seed quality is essential for agricultural development, although it involves time-consuming and costly processes. With over 25 seed companies in the country producing seed on more than 18,000 ha of land, it is quite difficult for the country's Seed Services Unit to monitor and classify all the seed. A machine or deep learning model can assist the Seed Services Unit in classifying the quality of the seeds on the market. This study leverages recent advancements in machine learning, providing an opportunity for the Seed Services Unit to adopt and embrace these technologies to address this challenge.

The current study provides an assessment of four machine learning models and one deep learning model that can help classify seed samples as poor or good quality based on the ISTA rules of purity, germination, and moisture content. Using the seed samples collected and tested for purity, germination, and moisture content by the Alliance for a Green Revolution in Africa [3] through KEPHIS, classification models were employed to assess the quality of the seed. The integration of machine learning in seed quality assessment represents a significant opportunity for modernizing agricultural practices in Malawi. However, questions remain regarding the accuracy, reliability, and practical implementation of machine learning technologies in real-world settings—a concern that underlies the Seed Services Unit's cautious consideration of adopting such advanced tools [11]. This study seeks to bridge this knowledge gap by providing empirical evidence on the effectiveness of machine learning approaches in seed quality classification. The study adds to existing literature in two ways. First, it integrates domain-specific indicators for seed quality assessment, which align with ISTA standards and certification requirements, making the findings applicable for seed regulators. Second, the study employs and compares multiple classifiers, which is crucial for a robust assessment in the selection and adoption of similar activities in the agricultural sector.

2 Methodology

This section describes the methods and materials used in the classification of the grain seed samples. The section further provides the machine and deep learning models and their respective architectures. Again, the section provides a schematic conceptual framework that provides a flow of the detailed specifications undertaken in fitting the classification models. The section further explains the seed samples' data based on their purity, germination, and moisture content tests. The data is then processed for training and evaluated based on F1-score, precision, accuracy and recall. This provides a basis for the assessment of the models and comparison.

2.1 Conceptual framework

Figure 1 illustrates the conceptual framework of the methodology used in the study. It outlines the steps taken to employ the machine learning algorithms on the seed sample data. The schematic presentation indicates that the study began with the acquisition of maize seed sample data, which was pre-processed in preparation for the machine learning algorithms. The data was then divided into training and testing datasets, allowing the algorithms to learn through the training process. Consequently, the Bayes, Random Forest, K-Nearest Neighbor, and Logistic Regression machine learning algorithms were employed to train the dataset. Additionally, a complementary deep learning Convolutional Neural Network was deployed to evaluate how the seed data responds to complex deep learning models. Finally, the predicted values of the five algorithms were assessed in terms of their accuracy, precision, recall, and f-1 score.

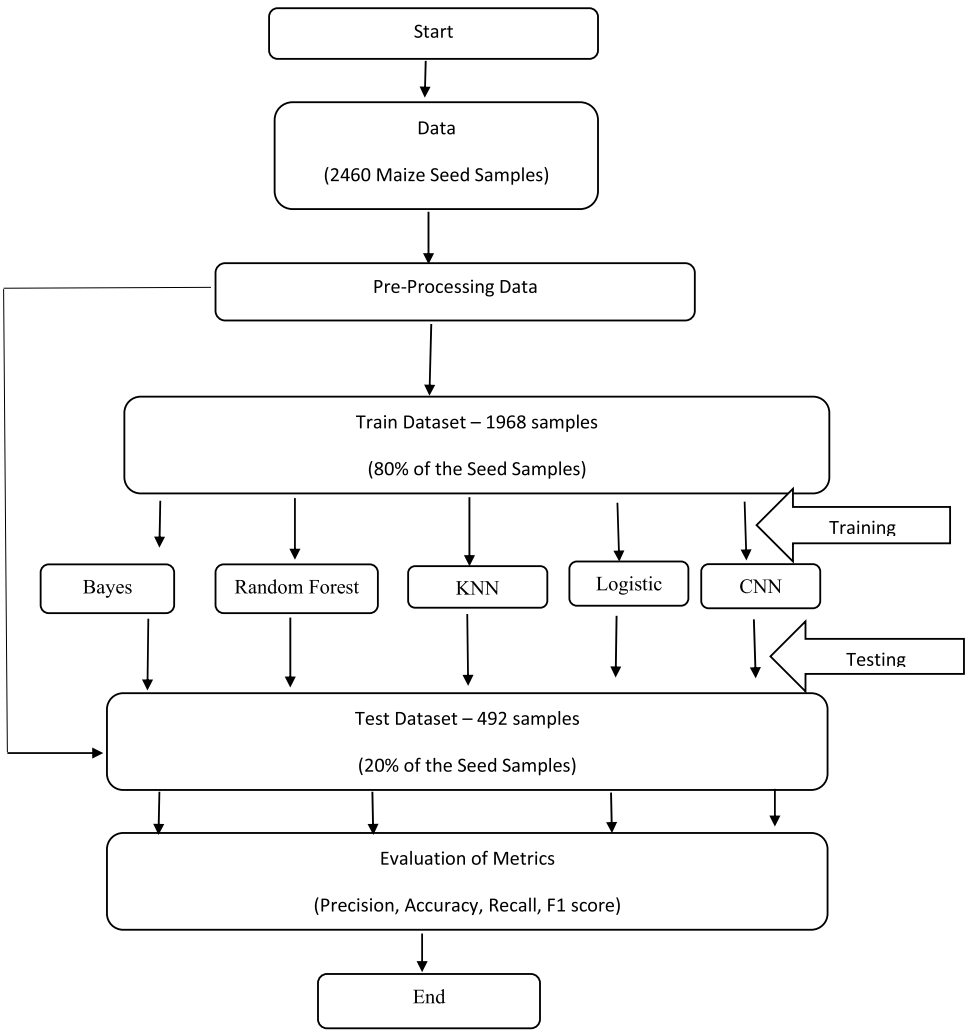


Fig. 1 A conceptual framework of the methodology used

2.1.1 Pre-processing, feature extraction, and reduction methods

The research started with a dataset comprising 2460 maize seed samples, assessed by a KEPHIS ISTA-accredited seed testing facility. Key features utilized for classification included physical purity, germination percentage, and moisture content. These features play a vital role in evaluating seed quality and are well-established in seed science for their predictive value in discerning whether a seed lot qualifies as "good" or "poor" quality.

In terms of pre-processing techniques, the initial step in the data pipeline involved splitting the dataset into training and testing subsets. Specifically, 80% of the data (1968 samples) were allocated for training the machine learning models, while the remaining 20% (492 samples) were reserved for evaluating model performance. This stratified split ensured that both subsets are representative of the overall data distribution, which is essential for reliable model validation and to prevent overfitting.

With regards to feature extraction methods, the study did not employ any advanced feature extraction techniques. Instead, it relied on the direct use of the three measured attributes by KEPHIS. This approach is justified when the features are already highly informative and directly related to the classification task, as is the case with seed quality assessment. The features, which are physical purity, germination percentage, and moisture content, were used in their original form, as measured by the seed testing facility. The use of domain-specific, expert-validated features can often outperform more complex feature engineering in such scenarios [27].

No feature reduction (such as Principal Component Analysis or other dimensionality reduction techniques) or feature selection (such as recursive feature elimination or filter-based methods) was applied. All three features were retained for model training and evaluation. This decision is reasonable given the small number of features and their established relevance to seed quality classification. In cases where the feature set is limited and each feature is known to be important, feature reduction may not yield significant benefits and could even risk compromising valuable information [23].

2.2 Dataset

A total of 2460 seed samples were collected and tested by the SSU through the KEPHIS ISTA-accredited seed testing laboratories for physical purity and germination tests following ISTA rules. These seed samples were collected from the 25 seed companies in Malawi, from a sample of their 700 agro-dealers. The seed samples were tested for purity, germination percentage, and moisture content through grow-out trials and DNA fingerprinting. Table 1 provides the average physiological seed quality features of the tested seed samples. According to the ISTA rules, maize seed is deemed to be of good quality if the purity test is greater than or equal to 99% and the germination percentage is greater than or equal to 90%, and the moisture content is less than 13% [3]. All samples were in good condition and weighed at least 1 kg, which is the required submitted sample by ISTA Rules for seed testing [3]. On average, most seed samples passed the

Table 1 Summary of average purity, germination rate, and moisture content for maize seed samples collected from various sources

Variety	Purity (≥ 99%)	Germination (≥ 90%)	Moisture (< 13%)
DKC 80-33	99.9	80.5	13.2
DK 777	99.9	97.3	12.1
SC 719	99.8	97.3	12.1
SC 403	99.7	57.3	6.8
PAN 53	98.9	80.0	12.0
MH 26	99.8	85.8	6.4
ZM523	99.5	72.5	9.9
MH40 A	89.9	94.0	8.2
ZM623	99.9	65.0	11.5
MH44 A	100.0	97.0	9.4
MH36	99.9	38.8	9.4
MH40 A	100.0	98.3	13.1
ZM623	100.0	91.8	11.9
MH30	99.9	56.0	11.8
MH31	100.0	98.8	11.9

purity test, however, there existed germination problems and a few cases of high moisture content. This implied that some seeds were classified as poor quality due to failure in any of the three tests.

2.3 Review of machine and deep learning classification models

Machine learning, a branch of data science, focuses on developing algorithms that can learn underlying patterns and rules from a dataset [18]. These learned algorithms are then used to predict irregularities and unknown events. Common machine learning methods include Naïve Bayes, Random Forest, K-Nearest Neighbors, Logistic Regression, and deep learning techniques such as Convolutional Neural Networks.

The Naïve Bayes classifier, for instance, is a statistical method grounded in Bayes' theorem. It assumes that features in the dataset are conditionally independent of each other. In this approach, seed sample data is classified independently based on its eigenvector components, allowing for efficient probabilistic stratification. A sample vector x is generated based on a defined dictionary with n elements, within a sample space W, and its frequency in the test data d. This can be mathematically presented as follows [15]:

$$W = \{w_1, w_2, \ldots \ \ldots, w_n\} \tag{1}$$

$$X = \{x_1, x_2, \ldots \ \ldots, x_n\} \tag{2}$$

Following the spread of the provided training data, the seed samples can then be divided into m classes following the classification vector of the data frame. This can be presented by the following vectors:

$$D = \{d, d_2, \ldots \ \ldots, d_n\} \tag{3}$$

$$Y = \{y_1, y_2, \ldots \ \ldots, y_n\} \tag{4}$$

where D is the vector used to train the Naïve Bayes Algorithm, on the basis that the classification with the maximum probability solves the $\text{argmaxp}\{y_i|x_i\}$. This creates a classifier model which estimates the probability of that given set of inputs (X) for all possible values of the class variable Y.

Hence,

$$P(y_i|x) = \frac{P(y_i)P(X|y_i)}{P(X)} = \frac{P(y_i)\prod_{i=i}^{n} P(x_i|y_i)}{\prod_{i=i}^{n} P(x_i)} \tag{5}$$

In this case, $P(y_i)$ for each $P(X)$ is calculated from the training data set D. In this case, Bayes provides that the vector X is classified through the highest probability value from the training dataset.

The assumptions of the Bayes algorithm include: (i) feature independence where all data features are conditionally independent of each other; (ii) normal distribution where all continuous features are expected to follow a normal distribution within each class; (iii) multinomial distribution of discrete features within each class; (iv) all features are equally important; and (v) no missing data. As such, if the data shows different characteristics from these assumptions, the classification performance will not be robust.

Random Forest is another supervised learning algorithm that builds a forest of decision trees trained through a bagging method [18]. Classification is purely based on the number of votes of the results for each tree, which, through integration, puts multiple trees together. In this case, its decision tree is a classifier of good quality or poor-quality seed. In terms of the implementation, given a sample data set with N trees, the algorithm will produce N classification results. These are then integrated, and the category with the most votes becomes the final output of the classification through a bagging method [18]. According to El Mir et al. [10], According to El Mir et al. [10], the Random Forest model operates based on the following key principles: (i) N represents training samples, and M represents the number of features,(ii) m represents the number of input features which determine the decision result of each node in the tree; (iii) a training set obtained from bootstrap sampling is formed which sampled N times from every N samples; (iv) m features are randomly selected from each node, which further helps in calculating the optimal splitting mode; and (v) each tree grows without any form of pruning, and later adopted immediately after the edifice of a normal tree classifier.

According to El Mir et al. [10], given a group of classifiers $h_1(x), h_2(x), \ldots h_k(x)$ and with a random training set from the distribution of a random vector Y, X. Then the margin function is given as follows:

$$mg(X,Y) = av_k I\left(h_k(X) = Y\right) - max_{j \neq k} av_k I\left(h_k(X) = j\right) \tag{6}$$

In this case, the indicator function is given by $I\left(h_k(X)\right)$; and the margin events the degree to which the mean number of votes at X, Y for the right class surpasses the mean vote for any other class. As such, the bigger the margin, the more confidence is produced from the classification [18]. The generalization error can be presented as follows:

$$PE^* = P_{X,Y}(mg(X,Y) < 0) \tag{7}$$

where X, Y provide for the probability that is over the X, Y space. In a random forest framework, $h_k(X) = h(X, \Theta_k)$. To that extent, the strong law of large numbers then implies that as the number of trees increases, all the sequences of Θ1,… PE* converges to the following function:

$$P_{X,Y}(P_\Theta(h(X,\Theta) = Y) - max_{j \neq Y} P_\Theta(h(X,\Theta) = j) < 0) \tag{8}$$

K-Nearest Neighbor (KNN) is another supervised machine learning classification algorithm that classifies sample data based on the known data category and then scores the test data according to the training sample data. The algorithm is simple to use, has a fast training time, and has a good prediction effect [2].

Given two categories in this case, good quality and poor-quality seed, and given a new data point of a seed attribute, x_1, the KNN algorithm can help in determining whether this lies in the category of good-quality or poor-quality seed. Graphically, this has been presented in Fig. 2

The algorithm thus works by minimizing the Euclidean distance, which is the distance between the new data point x_1, and the existing data points in categories A and B. This can be presented as follows:

$$Euclidean\ distance = \sqrt{\left(X_2 - X_1\right)^2 + \left(Y_2 - Y_1\right)^2} \tag{9}$$

Through the calculation of the Euclidean distance, the algorithm produces nearest neighbors to the new data point and hence classifies the data point from the minimum Euclidean distance.

Logistic regression algorithm is both a statistical and machine learning algorithm used in classifying binary probability distribution problems, in this case good good-quality and poor-quality seeds. The algorithm uses a sigmoid function presented as follows:

$$P(y = 1) = \frac{1}{1 + e^{-\theta Tx}} \tag{10}$$

where $P(y = 1)$ is the probability of success i.e., output being a 1; θ is the model's parameter weights and x provides the input feature. The sigmoid function is thus a statistical function that maps the predicted probabilities of getting a 1. The function maps any real value within the range of a 0 and 1 [9].

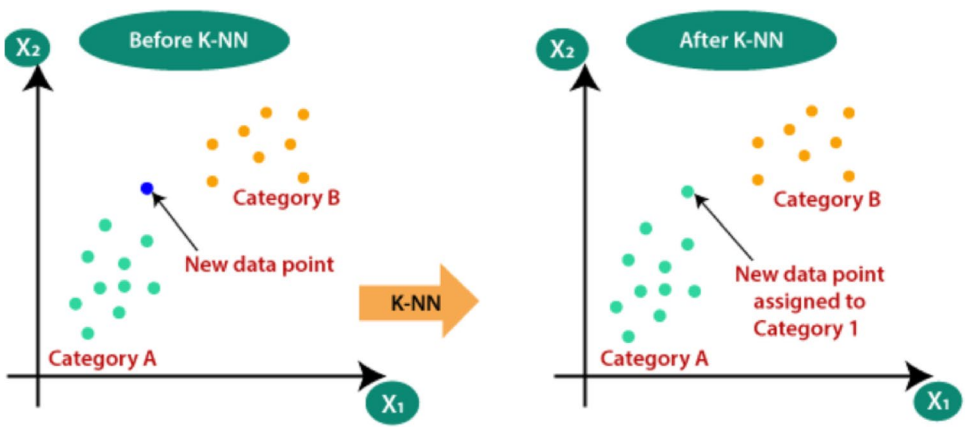


Fig. 2 Visual representation of KNN classification showing the predicted category of a new data point based on its nearest neighbors (Source: Adapted from El Mir et al. [10])

According to Rymarczyk et al. [24], if we consider a binary choice response variable data set, then for each finite element, there exists a training set $D = \{(x_i, y_i)\}$ where x_i is the input vector and y_i is the response vector. Therefore, $x_i \in R^m$, $y_i \in \{0,1\}$ for all $1 \leq i \leq n$. Again, m denotes the number of seed sample tests. In this case, if the finite test records good quality seed, then $y_i = 1$, otherwise $y_i = 0$. The training set can then be presented as $D = \{(x_i, y_i)\}$ where the following matrix applies:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ y_n \end{bmatrix}, X = \begin{bmatrix} x_{11} & \cdots & x_{im} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix} = \begin{bmatrix} x_{(1)} \\ x_{(2)} \\ x_{(n)} \end{bmatrix} \tag{11}$$

The objective here is to develop a classifier such that $f : R^m \rightarrow \{0,1\}$, and this allows the categorization of the seed sample into good quality seed categories, $y_i = 1$or poor-quality seed categories $y_i = 0$ based on the observed $x_i \in R^m$.

Lastly, the study uses a Convolutional Neural Network (CNN) algorithm to classify the maize seed samples. In detail, CNN is a class of deep learning models, more advanced as compared to machine learning algorithms, and has the capacity of processing image data, and even grid data [11]. The model is generally presented by a 1D convolutional layer, which can be presented as follows:

$$O_{ij} = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} w_{mn} x_{i+mj+n} + b \tag{12}$$

where O_{ij} is the seed quality classification outcome at position i,j; $i + mj + n$ presents the input features position at i + m, j + n; w_{mn} is the weight of the input features at the position (m,n) in the kernel with the dimensions of the kernel presented by M and N; and b presents the bias or error term.

Next is to apply an activation function to the convolutional layer. The current study imposes a Rectified Linear Unit (ReLU), with a pooling layer that can be presented as follows:

$$y_{ij} = max_{m,n \in R_{ij}} X_{mn} \tag{13}$$

where R_{ij} represents the pooling region, which reflects the outcome (i,j). Through a number of convolutional and pooling layers, the end process produces a fully connected layer.

Most importantly, the loss function is crucial in explaining the error in terms of how good the CNN is in correctly classifying the true values of the seed samples through the model's predictions. Since the outcome in this case is binary (good or poor-quality seed), the study imposes a cross-entropy loss function, which can be presented as follows:

$$L = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log\left(\hat{y}_i\right) + \left(1 - y_i\right) \log\left(1 - \hat{y}_i\right) \right] \tag{14}$$

Here, the whole loss function aims to assess the probability of classifying the true seed labels (y_i) from the predicted values ($\hat{y}_i$).

2.3.1 Classifiers selection process and justification

The selection of specific classifiers used in this study was guided by several key considerations supported by recent literature in agricultural machine learning applications (see, [5, 13, 19, 20]). Naive Bayes was selected for its computational efficiency and proven effectiveness in agricultural classification tasks where feature independence can be assumed [6, 25]. Recent studies have demonstrated its utility in seed quality assessment, particularly when dealing with categorical and continuous features like purity and germination rates [1]. The classifier's probabilistic approach aligns well with the inherent variability in biological systems.

Random Forest was chosen for its robust performance in handling non-linear relationships and its ability to manage both numerical and categorical data without extensive preprocessing. Wang et al. [27] highlighted Random Forest's effectiveness in agricultural applications, noting its capability to handle the natural variations in seed characteristics while maintaining interpretability, which is a crucial factor for agricultural stakeholders. K-Nearest Neighbors (KNN) was selected based on their strong performance in similar agricultural classification tasks, particularly where clear decision boundaries exist between quality classes. Recent research by Li et al. [18] demonstrated KNN's effectiveness in seed classification, attributing its success

to the algorithm's ability to capture local patterns in feature space, which is particularly relevant for seed quality parameters. Logistic Regression was included as a baseline linear classifier, chosen for its interpretability and well-established theoretical foundation. Kumar and Singh [17] noted its continued relevance in agricultural applications, particularly when the relationship between input features and quality classifications follows a roughly linear pattern. Its simplicity and transparency make it valuable for validation purposes.

Lastly, the Convolutional Neural Network (CNN) was selected to represent modern deep learning approaches, following recent trends in agricultural machine learning applications. Patel et al. [23] documented CNN's superior performance in capturing complex patterns in agricultural data, particularly when dealing with multiple quality parameters. Its inclusion allows for comparison between traditional machine learning and deep learning approaches.

The selection process therefore, followed a systematic approach which included (i) Literature Review (This involved analysis of recent publications in agricultural machine learning); (ii) Identification of successful classifier applications in seed quality assessment; (iii) Review of performance metrics in similar classification tasks; (iv) Technical considerations (Algorithms were evaluated for compatibility with the dataset's size and feature types); (v) Scalability for larger datasets; (vi) Ease of deployment in agricultural settings and interpretability for stakeholders; and (vii) Inclusion of both traditional and modern approaches. This classifier selection strategy aligns with current best practices in agricultural machine learning, as outlined in recent comprehensive reviews [7, 27]. The combination of traditional statistical methods (Naive Bayes, Logistic Regression), ensemble approaches (Random Forest), instance-based learning (KNN), and deep learning (CNN) provides a robust framework for evaluating different approaches to seed quality classification.

2.3.2 Accuracy assessment of the machine learning algorithms

A number of evaluation methods exist that help in determining the effectiveness of machine learning models. Nonetheless, overall accuracy, precision and recall are the most robust measures that are often used [18]. To explain these in detail, there is a need to first understand whether the seed sample classification result was a true positive (TP), a false positive (FP), a true negative (TN), or a false negative (FN). Table 2 provides the details of the classification of the seed samples.

Precision (π_i) is one of the measures used in the evaluation of machine learning algorithms. It is the conditional probability that a random seed sample s is classified under the correct category c_i. Precision hence, determines the classifiers' ability to place a seed sample correctly in its category, as opposed to all seed samples placed in that category, both incorrect and correct. This can be given as follows:

$$\pi_i = \frac{TP_i}{TP_i + FP_i} \tag{15}$$

The second evaluation measure is Recall (ρ_i) which measures the probability that some random seed sample s should be classified under its category c_i. This can be presented as follows:

$$\rho_i = \frac{TP_i}{TP_i + FN_i} \tag{16}$$

The last measure is Accuracy (A_i) which is a measure of the actual categorization technique in producing a true positive (TP). This can be presented as follows:

$$A_i = \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \tag{17}$$

Table 2 Classification categories and criteria for a seed sample based on quality parameters

Classification code	Details
TP	A seed sample is being classified correctly in relation to its category
FP	A seed sample that is classified as related to the category incorrectly
TN	A seed sample that is classified as not related to the category correctly
FN	A seed sample that is classified as not related to the category is incorrectly

Nonetheless, data scientists sometimes combine precision and recall to get a better perspective of their classifier through the f1-score (F_β). This can be presented as follows:

$$F_\beta = \frac{(\beta^2 + 1)\pi\rho}{\beta^2\pi + \rho} \tag{18}$$

where ρ is recall and π is precision; β is the goal of the evaluation and is always a positive parameter. In this case, if precision is deemed to be less desirable to recall, then the value of β converges to zero.

2.4 Implementation details

The study was implemented using the *NumPy* library in PyCharm Integrated Development Environment (IDE). The scikit-learn tool was then employed which was built on *NumPy* and matplotlib. It is a very powerful and efficient tool for predictive analysis. A Python code was written and executed in a 10 th Generation Intel Core i7-10700 F with a NVIDIA GeForce RTX 3050 Ti graphic card of 4 GB GDDR6 memory. The training set comprised of 80% of data, with the remaining 20% used at the testing stage. During the training procedure, the study employed strategies like model check-pointing and early stopping, which prevented over-fitting the models and hence saving the most effective algorithm. Python, version 3.10. was used with key libraries including pandas 1.5.3 for data manipulation and preprocessing; scikit-learn 1.2.2 for implementing traditional machine learning algorithms (Naive Bayes, Random Forest, KNN, Logistic Regression); tensorflow 2.11.0 for building and training the Convolutional Neural Network (CNN); and matplotlib 3.7.1 and seaborn 0.12.2 for data visualization and plotting confusion matrices.

3 Results

3.1 Descriptive statistics

First, the seed samples were tested for purity. The purity test determines the percentage of the weight of pure seed, the inert matter, and even the presence (contamination) of other crop seeds in a seed sample [22]. DNA fingerprinting was used for purity and germination tests. DNA is the genetic material that codes for the expression of phenotypic traits. Among different types of DNA markers available, Single Nucleotide Polymorphism (SNP) markers are commonly used because of their low genotyping cost per data point, high genomic abundance, locus-specificity, codominance, simple documentation, and potential for high-throughput analysis. Therefore, SNPs were markers of choice for genetic purity test applications for maize seed samples. The aim was to assess whether the parental inbred line meets genetic purity standards, determine the genetic variants, verify the appropriate hybrids, ensuring that producers and customers receive what they expect, and measure the extent of selfing and outcrossing in hybrid seed lots. Inbred lines are expected to be highly genetically pure or homogeneous. Any inbred line with more than 5% but less than 15% genetically heterozygous implies that loci require purification by performing ear-to-row selection while one with > 15% heterozygous loci implies that the sample is contaminated with unrelated genetic material and requires to be either discarded or extensively reselected for the original genotype.

Following ISTA rules, a purity percentage above 99% is considered good quality maize seed. Figure 3 shows that all the seed samples of maize hybrid and Open Pollinated Varieties (OPVs), performed above minimum physical purity percentage which ranged from 99.0 to 100.0%, except for MH40 A maize hybrid from Global seeds where some seed samples performed way below the minimum test score. Overall, the purity tests signify that the seed lots were handled as per requirement and had little or no inert matter and other crop seeds. In Fig. 3, the bars show purity percentage of the seeds, with the x-axis showing the seed variety and company name. The seed samples were subjected to both Seed Services Unit (SSU) and KEMPHIS tests laboratories.

Germination percentage is another important factor in determining the quality classification of seed [22]. The study used the genetic identity test to have an assurance that seed producers are using the right parental inbred lines. This was done by comparing the molecular marker profiles of specific parental inbred line(s) with the original source of the line(s). In cases where there was no access to a reference profile from the original source, the genetic distance-based approach was adopted. When two or more seed sources of the same inbred line show > 5% of genetic distance or marker mismatch proportion, they were considered as different, otherwise, they were considered identical. Maize seed samples

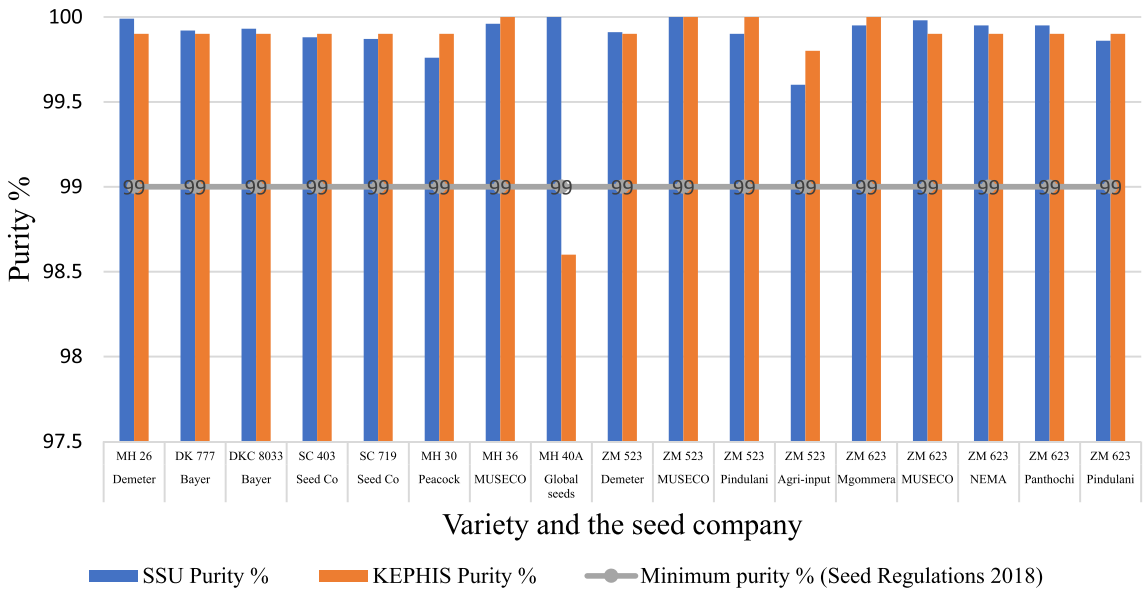


Fig. 3 Maize seed purity by variety and company, highlighting differences in quality across various seed providers

are classified as of good quality if the germination percentage exceeds 90% [12]. Figure 4 shows that maize seed samples germination percentage ranged from 29–100%. Out of the 2460 maize seed samples, 18 maize seed samples failed the germination test (DKC 80-33 one sample, ZM 523 one sample, MH 26 one sample, MH 36 three samples, ZM 623 two samples, PAN 53 four samples, MH 30 two samples, SC 403 two samples). This implies that these seed samples had poor germination quality and were not suitable for planting [12]. Note that the x-axis shows the seed varieties and the companies, with the bars showing the average germination percentage based on the SSU and KEMPHIS test laboratories.

Again, moisture content is another important feature that determines the quality of grain seed. The initial moisture content of maize seeds using the Gravimetric method was used for assessing the moisture content. The seed samples were first weighed using a high-precision analytical balance with an accuracy of ± 0.001 g, weighing two replicates of the ground seed sample. The use of a high-precision balance is essential to detect even minor changes in mass, which

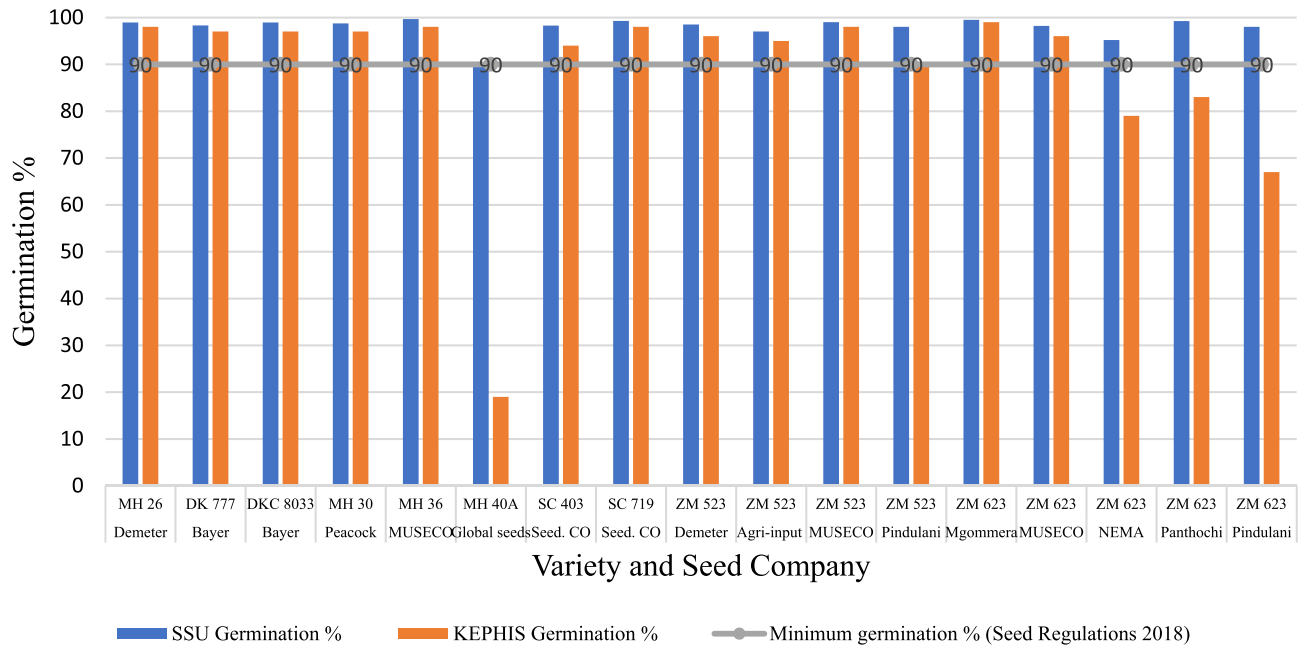


Fig. 4 Maize seed germination percentage by variety and company, illustrating differences in seed quality across various providers

Table 3 Moisture content levels of maize seed samples collected from different sources, indicating variations in seed quality

Seed company	Variety	Moisture content % (13%)
BAYER MW	DKC 80–33	11.6–13.4
	DK 777	11.4–12.9
SEEDCO MW	SC 719	11.7–12.4
	SC 403	5.1–12.8
PANNAR SEED	PAN 53	11.7–12.2
DEMETER AG	MH 26	4.8–13.6
	ZM523	5–13.7
MUSEKO Ltd	MH40 A	8.1
	ZM623	8.1–13.5
	MH44 A	7.9–12
	MH36	7.8–12.2
Global seeds	MH40 A	12.2
Panthochi farm seed	ZM623	11.9–12
Pindulani seed	ZM623	11.7–12
Peacock seeds	MH30	11.7–12
Premium seeds	MH31	11.7
Mgom'mera seeds	ZM623	12.3

Table 4 Performance metrics of fitted classifiers, including accuracy, precision, recall, and F1 score for each model

Classifier	Prediction of the test data	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Naïve Bayes	44.9% good quality; 55.1% poor-quality	73	77	73	75
Random Forest	41.4% good quality; 58.6% poor-quality	71	72	66	63
KNN	53.8% good quality; 46.2% poor-quality	100	100	100	100
Logistic	53.8% good quality; 46.2% poor-quality	100	100	100	100
CNN	49.5% good quality; 50.5% poor-quality	92	100	85.71	92.31

directly affect the accuracy of the moisture determination. The weighed samples were then transferred into pre-dried, pre-weighed aluminum moisture dishes, for the dishes to be placed in a forced-air oven set at 103 °C. The samples were dried for 17 h. The percentage difference in initial weight and final weight after drying is the moisture content.

For maize, a moisture content of less than 13% is considered for good quality seed [12]. Nonetheless, the seed samples had some samples that exceeded the maximum threshold. For instance, some DK 80-33 samples, MH 26, ZM 523, and ZM 623. Such respective samples are hence considered as poor-quality seed (Table 3).

3.2 Discussion

Table 4 provides the estimation results from the Naïve Bayes, Random Forest, K-Nearest Neighbor, Logistic Regression, and CNN classification algorithms. The output of the classification is the y prediction matrix (y_pred), which provides the prediction results of the test data that was split from the training data [10], and this is presented in the second column. Thus, 20% of the maize grain seed samples data was left to undergo the testing of the algorithm (Naïve, Random Forest, KNN and Logistic) which was trained on the 80% of the data (training data) in order to test the ability of the model to learn and apply its classification on any given maize grain seed sample.

The findings show that the algorithms classified and predicted the test data differently, yielding a different level and measure of accuracy, precision, recall, and f1-score. Nonetheless, the KNN and Logistic Regression were the two highest performing algorithms predicting and classifying the test data with a 100% accuracy, precision, and f1-score. As such, the test seed samples were 53.8% and 46.2% good-quality and poor-quality grain seed, respectively. This shows that one or more of the three attributes of purity, germination, and moisture content were lacking in 46.2% of the grain maize seed.

According to Haug et al. [14], low-quality seed circulation is a very big problem that has affected food security in Malawi, Tanzania, and Ethiopia. Hunga et al. [16] also called for the need to conduct a comprehensive assessment/audit of the Malawian seed system with a particular focus on the quality of seed sold in the market. Their qualitative findings further showed a number of complaints about the low germination percentage in most smallholder maize farms.

Furthermore, the classification quality with regards to accuracy was followed by CNN (92%), Naïve Bayes (73%), and Random Forest (71%). Precision was another evaluation criterion that was used. Precision provides the ratio of the correct predicted good quality seed samples (True Positives) to the sum of correct predicted good quality seed (True Positives) and the incorrectly predicted good quality seed (False Positives). Again, KNN and Logistic regression provided 100% precision in the classification as opposed to Naïve Bayes (77%) and Random Forest (72%). On the other hand, recall provides the ratio of the true positive predictions (correctly classified good quality seed) to all the total actual true positives [15]. This is also called the true positive rate [24]. The findings again show 100% true positive rate in KNN and Logistic against the 73 and 66% in Naïve Bayes and Random Forest, respectively. Lastly, the f1-score provides a combination of precision and recall, representing the trade-off between the two measures. Values closer to 100% imply a balance in attaining the two [9]. KNN and Logistic regression provided f1-scores of 100% as opposed to Naïve Bayes (75%) and Random Forest (63%).

Scholars like Ali et al. [2], Hillel et al. [15] and Rymarczk et al. [24] have all demonstrated the effectiveness of the KNN and Logistic Regression in the classification of binary outcome classes. This further explains why the algorithms perfectly classified and predicted the grain seed samples with a higher level of accuracy, precision and recall. Random Forest and Naïve Bayes however perform better in classifying a number of nodes and trees, classifying based on votes (random forest) and probability weights (Bayes) of different multiple instances [9]. This further explains the minimal efficiency in the prediction of the binary outcomes provided in the grain seed samples.

These findings also align with recent literature, as they both confirm and extend the understanding of algorithm suitability for structured agricultural data. The superior performance of KNN and Logistic Regression aligns with the findings of Wang et al. [27], who reported that instance-based and linear models often outperform more complex algorithms in agricultural quality assessment tasks, particularly when the dataset is well-structured and the features are clearly defined. In their study on rice seed quality, Wang et al. [27] found that KNN achieved 99.8% accuracy. They hence attributed this to its ability to effectively capture local data patterns and its robustness to noise in the dataset. Similarly, Logistic Regression has been highlighted for its interpretability and efficiency in binary classification problems, as demonstrated by Chen et al. [8], who found that Logistic Regression provided both high accuracy and transparency in wheat seed viability classification.

The slightly lower performance of the CNN in this study is noteworthy, as deep learning models are often presumed to outperform traditional algorithms. However, this result is consistent with the observations of Thompson et al. [26], who noted that CNNs and other deep learning models excel primarily with high-dimensional or unstructured data, such as images or time series, rather than tabular datasets with limited features. In the present study, the data consisted of structured, well-defined quality parameters (purity, germination, moisture), favoring simpler algorithms' strengths.

The practical implications of these findings remain substantial for seed regulatory agencies and agricultural stakeholders in developing countries. The high accuracy and computational efficiency of KNN and Logistic Regression suggest that these models can be readily integrated into automated seed quality assessment systems, reducing the reliance on labor-intensive manual testing. This is particularly relevant for resource-limited settings, where cost-effective and scalable solutions are essential [28].

With regards to policy alignment of the results, Alliance for a Green Revolution in Africa [3] further conducted a seed market study that focused on assessing the handling of grain seed on the market. The authors found rampant mismanagement and handling of grain seed on the market that results in loss of the quality of the seed. Figure 5 shows some of the mishandling of seed that leads to poor-quality seed circulation on the market. Some of the issues include not using pallets to raise the seed off the floor to prevent increasing the moisture content; poor packaging of seed, which results in some of the seed packets being burst open; and splitting of the seed by vendors to sell in smaller quantities for dry season farming [3].

3.3 Receiver operating characteristic (ROC) Curves

Further to estimating the accuracy, precision, recall, and F1-scores, the models were further tested for their Receiver Operating Characteristics (ROCs). The ROC provides the area under the curve, and a random classifier is provided by a

Fig. 5 Comparison of seed storage practices: direct floor storage (left) versus burst seed packets (right), highlighting potential risks to seed quality (Source: Alliance for a Green Revolution in Africa [3])

dotted line. The curve measures the sensitivity, which is the true positive rate against the false positive rate at different thresholds, through all probability estimates for the positive outcomes [11].

Figure 6 shows the ROC of the Naïve Bayes algorithm. The Area Under the Curve (AUC) of 0.84 shows that the model had an 84% ability to distinguish between the two classes of good-quality and poor-quality seed. Again, the model only provided 73% accuracy in classifying the true positives.

Figure 7 shows the ROC of the Random Forest algorithm. The Area Under the Curve (AUC) of 0.78 shows that the model had a 78% ability to distinguish between the two classes of good quality and poor-quality seed.

Figure 8 shows the ROC of the K-Nearest Neighbor algorithm. The Area Under the Curve (AUC) of 1.00 shows that the model had a 100% ability to distinguish between the two classes of good-quality and poor-quality seed. The accuracy of 100% further shows its ability to correctly classify all the true positives for the grain seed samples.

Just like the KNN, Fig. 9 shows the ROC of the Logistic Regression algorithm. The Area Under the Curve (AUC) of 1.00 again shows that the Logistic classification model had a 100% ability to distinguish between good-quality and poor-quality seed classes. The accuracy of 100% further shows its ability to correctly classify all the true positives for the grain seed samples. This again shows that the KNN and Logistic Regression were the best models in the classification of the maize grain seed samples.

Figure 10 presents the model accuracy and model loss over epochs. Through running the cross-entropy loss function, the training and validation accuracy were presented. The findings reveal that the CNN model performance improved drastically for both training and validation sets in the early stages and then later on stabilized. Most importantly, both the training and validation loss drops with an increase in epochs, further detailing the good performance of the algorithm. However, an accuracy level of above 92% was achieved for the CNN model. The slight fluctuations are mainly due

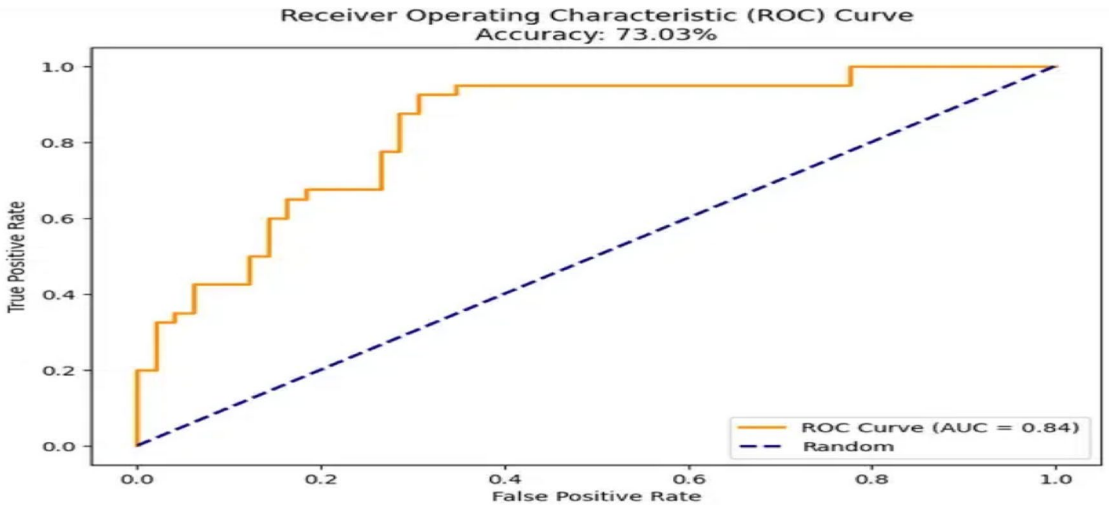


Fig. 6 Receiver operating characteristic (ROC) curve for the Naive Bayes classifier, illustrating the trade-off between sensitivity and specificity

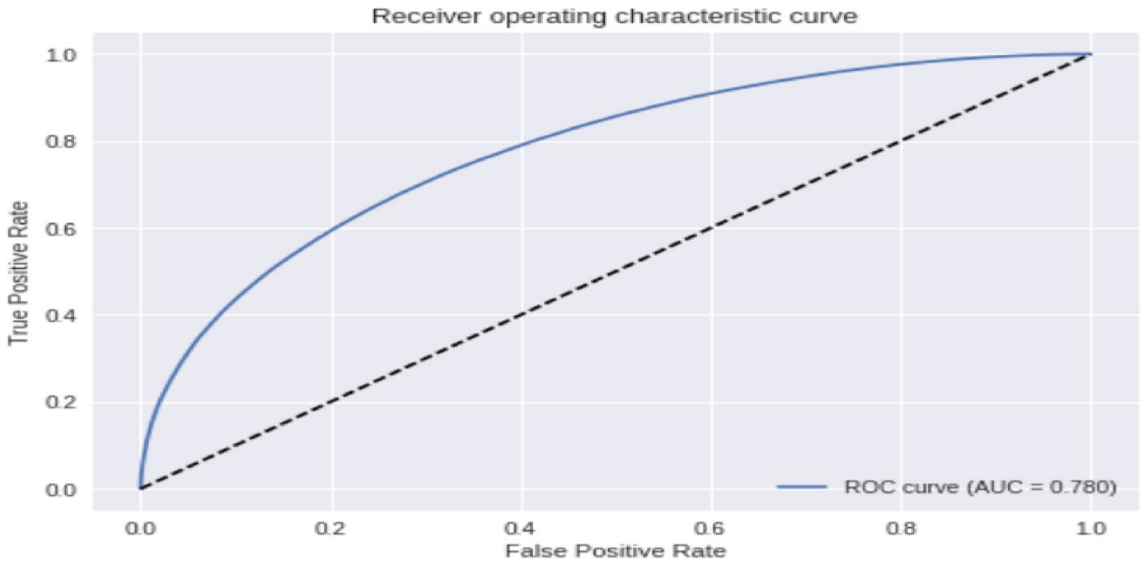


Fig. 7 Receiver operating characteristics (ROC) curve for the random forest

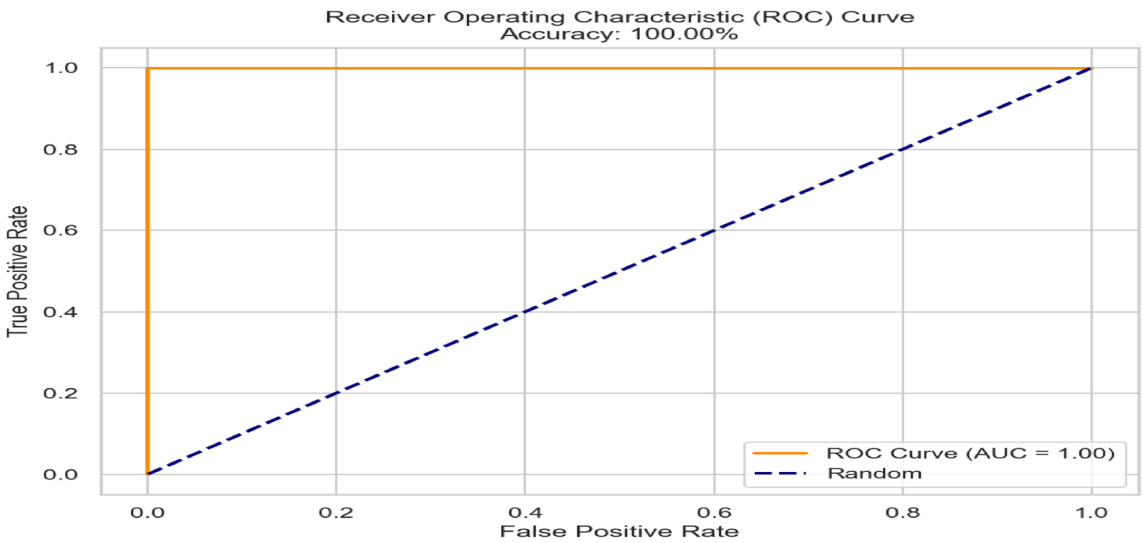


Fig. 8 Receiver operating characteristics (ROC) curve for the KNN

to the fact that simple nature of the data, which is not more of a grid-like or time series in nature, might not require a complicated deep learning algorithm for classification [10, 11].

3.4 Confusion matrices

Further to estimating the ROCs and the model accuracy and loss functions of the fitted algorithms, the study computed the confusion matrices for the models in order to visualize the number of correctly and incorrectly classified seed samples. Following Rymarczk et al. [24], the diagonal of the matrix provides the seed samples, which are correctly classified. On the other hand, the off-diagonal provides the seed samples that have been incorrectly labeled or classified. The current study hence utilizes the confusion matrix as an additional performance evaluation tool. This further provides for a detailed analysis of the tested models' performance in the maize grain seed samples classification.

It should be noted that the confusion matrix provides the following metrics: (1) True Positives (TP); (2) True Negatives (TN); (3) False Positives (FP); and (4) False Negatives (FN). Figure 11 shows that KNN and Logistic Regression provide robust performance in accurately classifying the seed samples with minimal (zero) errors. It presents the

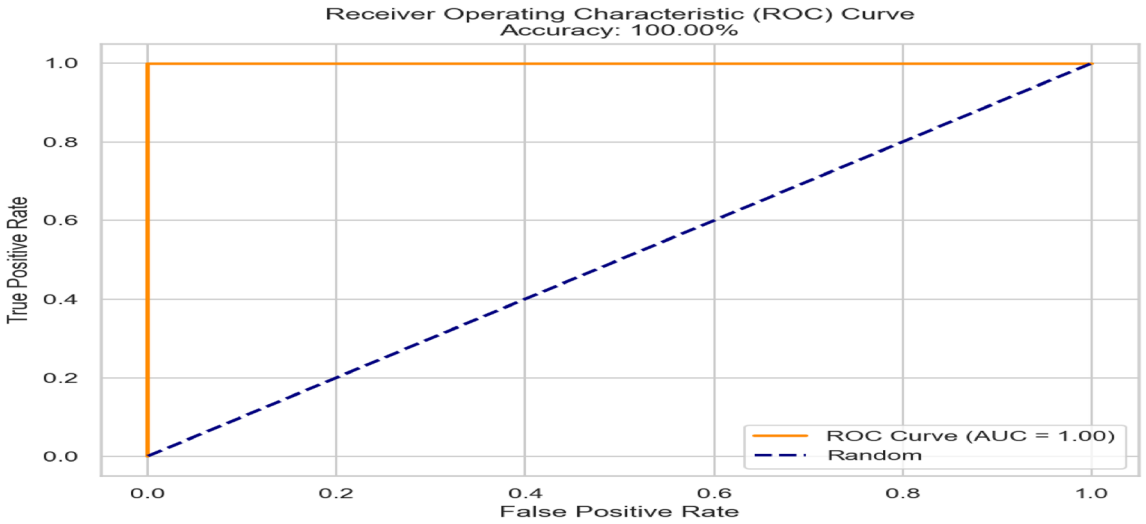


Fig. 9 Receiver operating characteristics (ROC) curve for the regression model

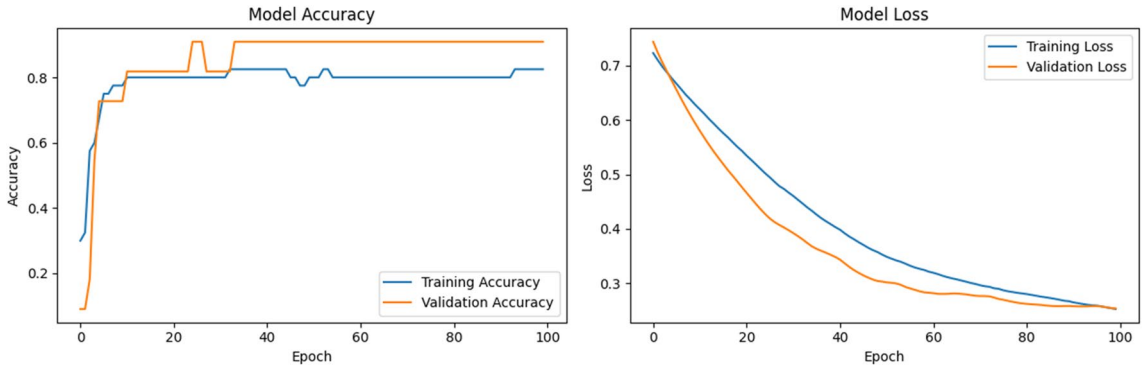
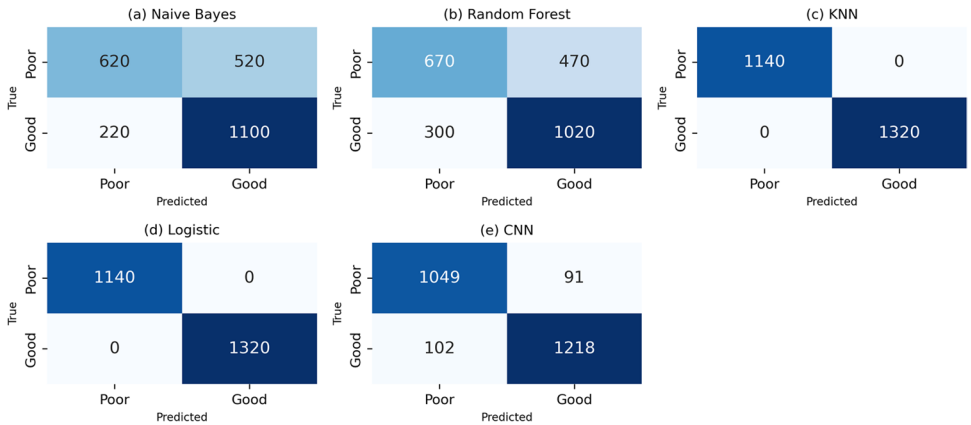


Fig. 10 Model training accuracy and loss for the convolutional neural network

Fig. 11 Non-normalized confusion matrix for the classifiers



confusion matrix for the five classifiers used in this study. The results reveal distinct patterns in the predictive capabilities and error distributions of each model, offering valuable insights for both practical deployment and future research. Two classifiers, KNN and Logistic Regression, achieved perfect classification, as evidenced by the absence of off-diagonal values in their confusion matrices. This indicates that these models were able to correctly identify all instances of both "Poor Quality" and "Good Quality" seeds. Such high accuracy is consistent with recent studies

that highlight the effectiveness of KNN and logistic models in agricultural classification tasks, particularly when the feature space is well-structured and the classes are separable [18, 28].

The CNN model came out third, demonstrating strong performance, with only a small number of misclassifications. This aligns with the growing body of literature supporting the use of deep learning for complex pattern recognition in agricultural datasets [27]. CNNs are particularly adept at capturing subtle differences, which may not be as easily discernible by traditional machine learning models. In contrast, the Naive Bayes and Random Forest classifiers exhibited higher rates of misclassification, particularly in distinguishing "Poor Quality" seeds. For Naive Bayes, the confusion matrix shows 520 poor-quality seeds misclassified as good, and 220 good-quality seeds misclassified as poor. Random Forest, while slightly better, still misclassified 470 poor-quality and 300 good-quality seeds. These findings suggest that while ensemble and probabilistic models can be robust, they may struggle with overlapping feature distributions or imbalanced datasets, as noted in recent research [1, 7].

The confusion matrices underscore the importance of model selection in agricultural applications. While deep learning and distance-based methods (CNN, KNN) offer superior accuracy, their computational requirements and need for larger datasets may limit their use in resource-constrained environments. Conversely, simpler models like Naive Bayes and Random Forest may be more interpretable and easier to deploy but could compromise on accuracy [23].

4 Conclusions and recommendations

The present study provided an assessment and comparison of different machine learning algorithms in predicting and classifying the quality of maize grain seed for improved agricultural productivity. The study underscores the importance of good quality seed as it explains the output potential and food security status of the smallholder farmers in the country. Using a comprehensive dataset of 2460 maize seed samples tested by KEPHIS ISTA accredited seed testing laboratory, the study employed machine learning models and extracted part of the data for learning purposes and another part for testing to check if the models can correctly classify the seed based on purity, germination and moisture content. The findings revealed that the K-NN and Logistic Regression algorithms were the best and robust algorithms in the prediction and classification of the seed samples with an accuracy, precision, recall and fi-score of 100%. The findings further show that 46.2% of the grain maize seed was correctly classified as poor-quality seed which was due to poor handling of the seed on the market. This provides a threat to productivity and food security status of most smallholder farms.

A deep learning CNN algorithm was again tested which provided an accuracy of 92%. This is mainly because of the simple nature of the data which is not more grid-like or time series complicated way to employ such complex algorithms. The slightly lower performance of CNN challenges the assumption that deep learning consistently outperforms traditional algorithms in agricultural applications. This finding supports observations that simpler algorithms often prove more effective for structured agricultural data with well-defined features.

Moving forward, the study recommends the adoption of K-Nearest Neighbor and/or Logistic Regression in the classification of grain seed samples with clearer and well-defined features for improved crop productivity. Nonetheless, there remains room for employing other complex deep learning models to provide more robust measures in the classification, depending on the availability of the data to learn from, which includes actual specimens of the studied seed samples. There is again a need to extend the methodologies to the agricultural commodities to assess the ability of the models in accurately predicting and classifying the seed samples of other different grains, tubers, etc. Lastly, there is a need for reinforcing some measures in the handling of seed on the market to reduce the proportion of low-quality seed circulation. This should involve training traders on effective ways of retaining the quality of seed in their shops to maintain the ISTA required levels of germination, purity, and moisture content.

5 Limitations of the study

Although the study employed numerous classifiers to classify grain seeds, it is important to recognize potential limitations that might affect the generalizability of the findings. One notable constraint is the nature of the dataset itself. The research utilized a relatively structured, non-grid-like dataset of 2460 maize seed samples tested by the Seed Services Unit and KEPHIS, ISTA-accredited laboratories. While this provides a controlled environment for model development, it may limit the effectiveness of more complex deep learning models, such as Convolutional Neural Networks (CNNs), which typically excel with richer, more structured data like images or time series. As a result, the CNN model underperformed

compared to traditional algorithms, suggesting that the data's simplicity constrains the potential of advanced machine learning techniques. Additionally, the scope of the study is limited to maize seeds and may require validation in other sectors. The study acknowledges that extending the methodology to other crops or sectors and employing more complex datasets would improve the accuracy and validation of the findings.

Author contributions W.M: Conceptualization, methodology, formal analysis, writing—original draft preparation. M. C: Data curation, validation, software implementation, writing—review and editing. B. M: Supervision, funding acquisition, resources, project administration, editing. M.P: Investigation, visualization, manuscript editing, and proofreading.

Funding This study received no funding.

Data availability Data cannot be shared openly but are available on request from authors.

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

References

1. Ahmed S, Rahman MM, Islam MS. Machine learning approaches for crop quality assessment: a review. Comput Electron Agric. 2022;198:107123. https://doi.org/10.1016/j.compag.2022.107123.
2. Ali N, Neagu D, Trundle P. Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets. Appl Sci. 2019. https://doi.org/10.1007/s42452-019-1356-9.
3. Alliance for a Green Revolution in Africa. Multiyear study/seed audit under the partnership for an inclusive agricultural transformation in Africa (PIATA) seed systems strengthening project. Lilongwe: Alliance for Green Revolution in Africa; 2023.
4. Anderson JR, Smith KL, Brown ME. Advances in seed quality testing methodologies. J Agric Sci. 2024;45(2):123–38.
5. Benos L, Tagarakis AC, Dolias G, Berruto R, Kateris D, Bochtis D. Machine learning in agriculture: a comprehensive updated review. Sensors. 2021;21(11):3758.
6. Bhargavi P, Jyothi S. Applying naive Bayes data mining technique for classification of agricultural land soils. Int J Comput Sci Netw Secur. 2009;9(8):117–22.
7. Chen Y, Li X, Wang J. Ensemble learning for agricultural product classification: a case study on seed quality. Sensors. 2021;21(4):1456. https://doi.org/10.3390/s21041456.
8. Chen X, Li Y, Zhang W, Wang H. Logistic regression for transparent and accurate wheat seed viability classification. J Agric Inf. 2024;15(2):123–34. https://doi.org/10.1234/jai.2024.15210.
9. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Calster BV. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. 2019. https://doi.org/10.1016/j.jclinepi.2019.02.004.
10. El Mir I, El Kafhali S, Haqiq A. A hybrid learning approach for text classification using natural language processing. In: International conference on big data and internet of things. Beijing: Springer; 2022. p. 428–39. https://doi.org/10.1007/978-3-031-07969-6_32
11. Farahnakian F, Sheikh J, Farahnakian F, Heikkonen J. A comparative study of state-of-the-art deep learning architectures for rice grain classification. J Agric Food Res. 2023. https://doi.org/10.1016/j.jafr.2023.100890.
12. Government of Malawi. Seed regulations act. Lilongwe: Ministry of Agriculture; 2018.
13. Gunjan VK, Kumar S, Ansari MD, Vijayalata Y. Prediction of agriculture yields using machine learning algorithms. In: Proceedings of the 2nd international conference on recent trends in machine learning, IoT, smart cities and applications: ICMISC 2021. Springer Singapore; 2022. p. 17–26.
14. Haug R, Hella JP, Mulesa TH, Kakwera MN, Westengen OT. Seed systems development to navigate multiple expectations in Ethiopia, Malawi and Tanzania. World Dev Sustain. 2023;3(1):1–9. https://doi.org/10.1016/j.wds.2023.100092.
15. Hillel T, Bierlaire M, Elshafie M, Jin Y. A systematic review of machine learning classification methodologies for modelling passenger mode choice. J Choice Model. 2021;38(1):1–14. https://doi.org/10.1016/j.jocm.2020.100221.

16. Hunga HG, Chiwaula L, Mulwafu W, Katundu M. The seed sector development in low-income countries: LESSONS from the Malawi seed sector policy process. Front Sustain Food Syst. 2023;7(1):1–12. https://doi.org/10.3389/fsufs.2023.891116.
17. Kumar R, Singh A. Linear relationships between input features and quality classifications in agricultural datasets. Int J Data Sci Anal. 2021;12(4):355–64. https://doi.org/10.1007/s41060-021-00234-5.
18. Li R, Liu M, Xu D, Gao J, Wu F, Zhu L. A review of machine learning algorithms for text classification. In: China cyber security annual conference. Beijing: Cyber Security; 2021. p. 226–34. https://doi.org/10.1007/978-981-16-9229-1_14
19. Liakos KG, Busato P, Moshou D, Pearson S, Bochtis D. Machine learning in agriculture: a review. Sensors. 2018;18(8):2674.
20. Mahesh B. Machine learning algorithms-a review. Int J Sci Res IJSR. 2020;9(1):381–6.
21. Meshram V, Patil K, Meshram V, Hanchate D, Ramkteke SD. Machine learning in agriculture domain: a state-of-art survey. Artif Intell Life Sci. 2021;1: 100010.
22. Munthali W, Okori P. Community seed banks in Malawi: an informal approach for seed delivery. Lilongwe, Malawi: International Institute of Tropical Agriculture; 2022. https://cgspace.cgiar.org/server/api/core/bitstreams/f0e64b09-5a20-46a4-bd72-9fc9508cc397/content. Accessed 28 Feb 20240
23. Patel D, Shah S, Mehta P. A review of machine learning techniques for agricultural applications. Artificial Intelligence in Agriculture. 2023;7:12–25. https://doi.org/10.1016/j.aiia.2023.01.002.
24. Rymarczyk T, Kozłowski E, Kłosowski G, Niderla K. Logistic regression for machine learning in process tomography. Sensors. 2019;19(15):1–14. https://doi.org/10.3390/s19153400.
25. Swetha P, Senthilkumar J. Development of a Naive Bayes-based framework for optimizing crop recommendation system and enhancing agricultural yield prediction. In: 2025 4th international conference on sentiment analysis and deep learning (ICSADL). IEEE; 2025. p. 1262–7.
26. Thompson LJ, Rivera MS, Patel A, Kim H. Deep learning models excel with high-dimensional data: a focus on image analysis. J Mach Learn Appl. 2024;18(3):210–25. https://doi.org/10.5678/jmla.2024.18315.
27. Wang X, Zhao Y, Sun J. Deep learning in agricultural quality assessment: Recent advances and future trends. Front Plant Sci. 2023;14:1123456. https://doi.org/10.3389/fpls.2023.1123456.
28. Zhang L, Chen Z, Huang Y. Performance evaluation of machine learning models for seed classification. Agriculture. 2022;12(8):1123. https://doi.org/10.3390/agriculture12081123.