Performance of advanced machine learning models in the prediction of amylose content in rice using internet of things-based colorimetric sensor

Shrinivas Deshpande[1,]*, Udaykumar Nidoni[2], Sharanagouda Hiregoudar[2], K. T. Ramappa[2], Devanand Maski[3] and Nagaraj Naik[4]

[1]ICAR-Krishi Vigyan Kendra, Kandali, Hassan 573 217, India
[2]Department of Processing and Food Engineering; [3]Department of Renewable Energy Engineering, and
[4]Pesticide Residue and Food Quality Analysis Laboratory, College of Agricultural Engineering, University of Agricultural Sciences, Raichur 584 104, India

Rice ageing is a complicated process that is difficult to examine methodically. Several physicochemical properties of rice change with age as a function of moisture content and storage temperature. Among these qualities, amylose content is the most important and numerous metrics depend on it. Several sensors, Internet of Things, Information and Communication Technology, artificial intelligence and machine learning (ML) approaches are being used in technological interventions to tackle this problem. In the present study, seven advanced ML models were evaluated to classify the different concentrations of amylose using light-intensity data obtained by the novel colorimetric amylose sensor. From the performance of the evaluated ML models, it was observed that for the light intensity dataset obtained from the sensor, higher and similar model parameters and an accuracy value of 0.77 were observed for both artificial neural network (ANN) and *k*-nearest neighbour (KNN) algorithms, followed by accuracy values of 0.75, 0.74, 0.65, 0.61 and 0.61 respectively, for the decision tree, random forest, AdaBoost, logistic regression and support vector machine algorithms. Thus ANN and KNN are promising in predicting the different classes of amylose in rice.

Keywords: Amylose content, artificial intelligence, machine learning, mathematical modelling, rice.

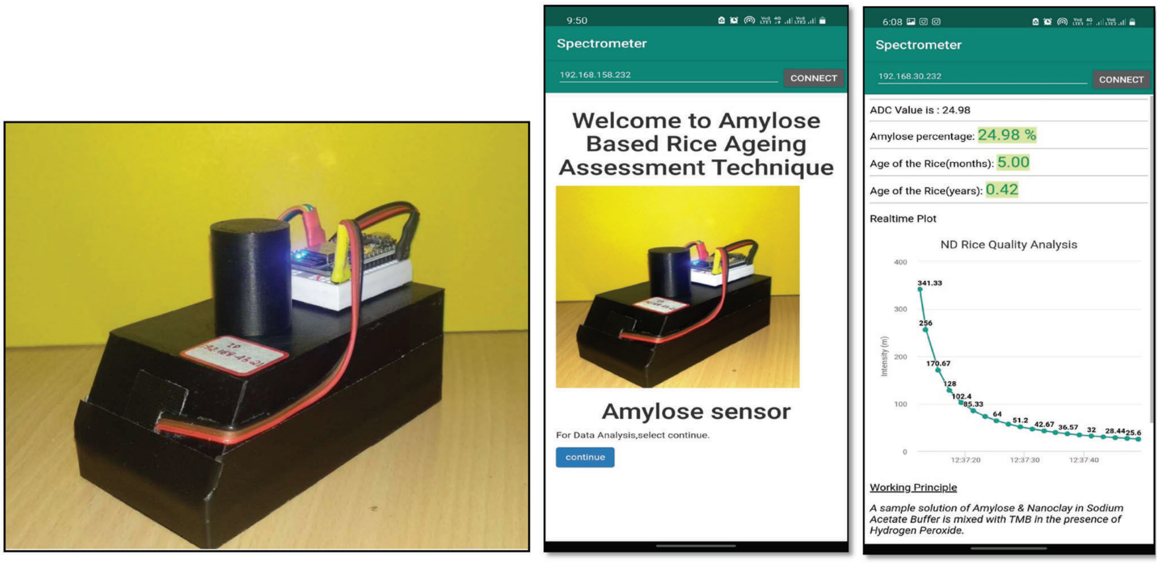
IN rice, the term ‘ageing’ basically represents biochemical changes that occur during grain storage as a function of moisture content, temperature and variety. When the freshly harvested paddy is milled, the rice gives a pasty gruel upon cooking, which the consumers least prefer. Under appropriate storage conditions, these characteristics decrease with due course of time, and the grains do not tend to adhere to one another when cooked. Ageing may also lead to a progressive increase in amylose content and changes in lipid, protein and other substances produced from enzyme activities during storage[1]. Lipids form free fatty acids and complexes with amylase enzyme along with carbonyl compounds and hydroperoxides. Hence, the ageing process could be quantified based on the amylose content present in rice. Due to these facts, aged rice is popular in Asian countries for its taste, texture and flavour compared to fresh rice. It shows higher kernel elongation, water absorption, volume expansion and less dissolved solid contents, which make the cooked grains flaky or grainy in texture[2]. Hence, old rice fetches a higher price compared to fresh rice, especially in India, while fresh rice is preferred in China, Japan and other countries[3].

The conventional method of assessing the age of rice involves the evaluation of cooking characteristics and observing the hardness of the cooked grains by pressing them on the palm. This leads to inappropriate price fixation of rice/farm produce using an unscientific method. Therefore, there is a huge demand for alternative methods and scientific devices for assessing the ageing of rice based on qualitative and quantitative measurement techniques[4].

The advances in science and technology and the high qualification of human resource have allowed sustainable growth of the world economy, resulting in the emergence of smart technological approaches. Computer applications based on sensors allow for obtaining more accurate information about any parameter. Hence, the internet of things (IoT)-based smart agriculture is more efficient than traditional approaches in solving real-time problems[5]. Furthermore, IoT-based sensors, instruments, devices or any other electronic kits could boost the qualitative analysis of agricultural produce by avoiding human interventions.

Mathematical models in food science help simulate the experimental processes and thus reduce them. In this regard, studies have been performed to provide appropriate models

*For correspondence. (e-mail: shrihd9@gmail.com)



192.168.158.232

Welcome to Amylose Based Rice Ageing Assessment Technique

Amylose sensor

For Data Analysis,select continue.

192.168.30.232

ADC Value is : 24.98

Amylose percentage: 24.98 %

Age of the Rice(months): 5.00

Age of the Rice(years): 0.42

Working Principle

A sample solution of Amylose & Nanoclay in Sodium Acetate Buffer is mixed with TMB in the presence of Hydrogen Peroxide.

Figure 1. Colorimetric amylose sensor and android mobile application.

for post-harvesting operations in order to find the best correlation between effective parameters[6].

Machine learning (ML) is one of the fastest-growing areas of computer science, with far-reaching applications. It refers to the automated detection of meaningful patterns in data. ML tools deal with call functions having the ability to learn and adapt[7]. With ever-increasing amounts of data becoming available, there is a good reason to consider that smart data analysis will become even more pervasive as a necessary tool for technological progress[6]. With this background, the objective of the present study was to evaluate the performance of different advanced ML models to accurately predict the amylose content in rice samples using intensity data obtained from the developed colorimetric amylose sensor.

Materials and methods

Description of the amylose sensor

A colorimetric amylose sensor was developed in the Department of Processing and Food Engineering, College of Agricultural Engineering, University of Agricultural Sciences, Raichur, Karnataka, India, for assessing the age of rice samples using the enzyme mimic principle exhibited by the 3,3′,5,5′-tetramethylbenzidine (TMB) in the presence of hydrophilic bentonite (nano clay) and hydrogen peroxide to attain the characteristic colour change (Figure 1). An android mobile application was also developed to rapidly estimate amylose content in the rice samples (Figure 1).

Experimentation and data acquisition

In order to evaluate the different ML models, the sensor was evaluated with standard amylose of known concentration (0.2, 0.4, 0.6, 0.8 and 1.0 mg ml^{-1}) with 100 replications. Next, 100 μl of standard amylose of known concentration was taken in a quartz cuvette, and 1 ml of sodium acetate buffer solution was added to it. Then 200 μl of nano clay, 100 μl of hydrogen peroxide and 100 μl of TMB were added to allow a change in colour for 1 min. The amylose sensor was switched on and connected to the android application by inserting the device IP address through the WiFi module. Once the WiFi connection was established with the device and mobile, a cuvette was placed in the sample compartment in the sensor and closed with a cuvette holder cap. The 652 nm LED was then turned on, and the plot as a function of time versus intensity began in the programme, with real-time values presented on the graph. The plot indicated the amylose percentage where the stationery curve continues to drop in intensity values, and the corresponding value was connected with the amylose percentage.

The sensor was operated for 900 sec for each run, and the light intensity values corresponding to each concentration of amylose were recorded as a function of time which was used for modelling. A provision was made in all the algorithms to split input data into training and testing datasets for further testing and evaluation. Since BPT 5204 is a popular rice variety in the study area, it was considered for evaluation and modelling. The selected ML models were evaluated in Anaconda-Spyder 3 (Python) and compared by means of model accuracy.

Mathematical modelling of amylose sensor data

The ML models, viz. AdaBoost, artificial neural network (ANN), *k*-nearest neighbor (KNN), decision tree, logistic regression, support vector classifier (SVC) and random forest classifier were chosen to model the intensity data for the prediction of amylose content due to their effectiveness in classifying the dataset[8]. The confusion matrices obtained for each model depicting the classification data were plotted for all the models.

AdaBoost classifier: Adaboost or adaptive boosting is an ensemble model that combines a series of low-performing classifiers to develop an improved classifier. Further, it decreases the variance with the help of the bagging approach, bias using a boosting approach or improves predictions using the stacking approach. The model assigns weightage to the trained classifier data in each iteration according to the accuracy of the prediction. Further, the accuracy of the model was studied by changing the *n*-estimator values in the programme[9].

Artificial neural network classifier: This data processing paradigm functions similarly to the biological nervous system. The input layers provide the specified independent variables, while the hidden layers process and compute the input data into usable form by assigning synaptic weightage to specific sets of data according to their unique strength. Furthermore, the chosen activation function translates the input signal of ANN node to an output signal. Furthermore, the insertion of hidden nodes between the input and output layers might improve the accuracy of ANNs[8].

k-nearest neighbor classifier: This is a non-parametric and lazy learning algorithm with no assumptions for underlying data distribution, and the model structure determined from the dataset does not follow any mathematical assumptions. In KNN, the accuracy of the model depends upon the number of nearest neighbours. To find the closest similar points, the distance between points measured, such as Euclidean distance, Hamming distance, Manhattan distance and Minkowski distance, was considered. In the present study, Euclidean distance has been taken into consideration since it is a popular and default function in KNN[8].

Decision tree classifier: A decision tree is a flowchart-like tree structure that is an easy and popular classification algorithm mainly used for classification and regression analysis. The basic working process of the decision tree algorithm starts with selecting the best attribute among the input data using attribute selection measures (ASMs) to split the dataset, then breaking it into smaller subsets as decision nodes. Once the training and testing of the model are completed, the best ASM is adopted with a higher attribute score for obtaining the decision tree as well as the confusion matrix of the dataset for interpretation[8].

Logistic regression classifier: The logistic regression model forecasts the output using maximum likelihood estimation (MLE), which is a maximizing approach that uses a sigmoid function to generate output parameters that are most likely to produce the observed data, with mean and variance as critical factors. The multinomial logistic regression model was used in this study since the target variables contained five nominal categories of amylose concentration. The sigmoid or logistic function generates an *S*-shaped curve through which any real-valued number was computed and mapped onto a value between 0 and 1 (ref. 10).

Support vector classifier: This method is commonly used for classification and regression applications. A typical SVC plot includes a hyperplane (a decision plane that divides a group of objects), support vectors (data points nearest to the hyperplane) and margins (a gap between two lines on the closest class points). The data categorization is done by creating a hyperplane in multidimensional space to distinguish distinct classes of variables. The basic idea behind SVC is to determine the optimum maximum marginal hyperplane for dividing the dataset into multiple groups. The selection of an appropriate hyperplane with the maximum possible margin between support vectors is considered the principle behind the working of the SVC[8].

Random forest classifier: This is a type of supervised ML algorithm used both for classification and regression. Random forest algorithm creates decision trees on randomly selected data samples, obtains a prediction from each tree and selects the best solution by means of voting. Initially, the entire dataset is randomly split into a number of decision trees according to the type of dataset. This group of decision trees is considered as the forest, which depends on an independent random sample. In the classification problem, each tree votes and the most popular class is chosen as the final result[11].

Results and discussion

AdaBoost classifier

The classification analysis was done with the intensity data by changing the *n*-estimator values in the range 1 to 50, and the error rate pertaining to each *n*-estimator was recorded (Figure 2 *a*). It is observed that the error rate was low when the *n*-estimator value was 5, and this value was considered for auxiliary analysis. For the optimized value of the *n*-estimator, the accuracy of the classifier was 0.65. The results are represented as a confusion matrix (Figure 2 *b*) and an accuracy table (Table 1). From Table 1, it can be seen that the average precision of the AdaBoost model is 0.59. However, the average recall and F1 scores of the model are 0.65 and 0.58 respectively. The precision values of the model in identifying the selected concentration of amylose are 1.00, 0.00, 0.48, 0.58 and 0.87 respectively, for classes 0, 1, 2, 3 and 4. The recall and F1 scores for classes 0, 1, 2, 3 and 4 be 1.00, 0.00, 0.90, 0.98 and 0.35 and 1.00, 0.00, 0.62, 0.73 and 0.50 respectively.

The confusion matrix of the classifier illustrates that the AdaBoost model efficiently identifies three classes of

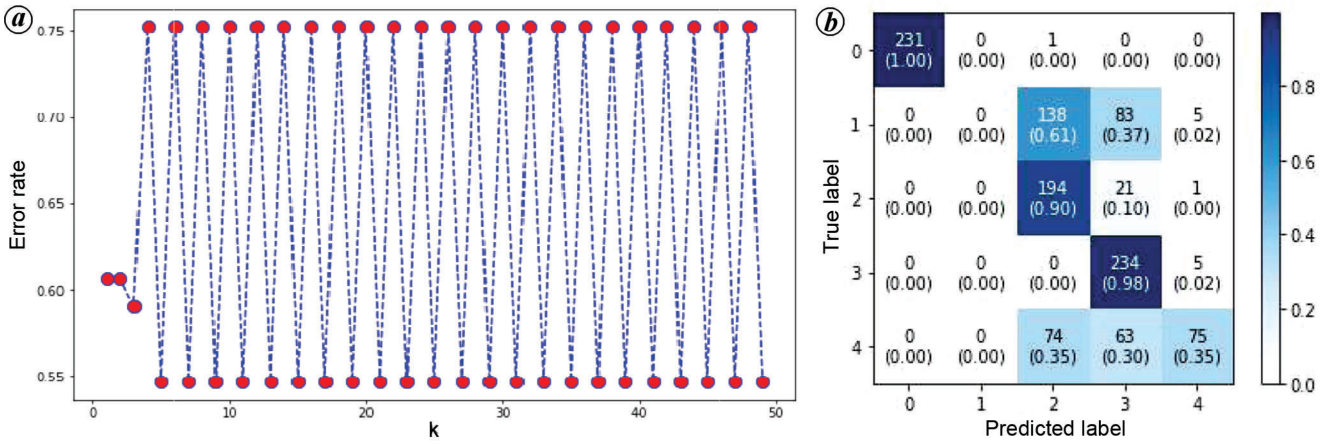


Figure 2. (***a***) Error plot and (***b***) confusion matrix plot for AdaBoost classifier.

Table 1. Evaluation results of selected machine learning algorithms for the intensity data of colorimetric amylose sensor

Model		Class							
		0 (0.2 mg ml^{-1})	1 (0.4 mg ml^{-1})	2 (0.6 mg ml^{-1})	3 (0.8 mg ml^{-1})	4 (1.0 mg ml^{-1})	Accuracy	Macro average	Weighted average
AdaBoost classifier	Precision	1.00	0.00	0.48	0.58	0.87		0.59	0.59
	Recall	1.00	0.00	0.90	0.98	0.35		0.65	0.65
	F1-score	1.00	0.00	0.62	0.73	0.50	0.65	0.57	0.58
ANN classifier	Precision	1.00	0.65	0.78	0.72	0.69		0.77	0.77
	Recall	1.00	0.55	0.82	0.71	0.78		0.77	0.77
	F1-score	1.00	0.59	0.80	0.71	0.74	0.77	0.77	0.77
KNN classifier	Precision	1.00	0.61	0.84	0.71	0.68		0.77	0.77
	Recall	1.00	0.69	0.79	0.62	0.74		0.77	0.77
	F1-score	1.00	0.65	0.81	0.66	0.71	0.77	0.77	0.77
Decision tree classifier	Precision	1.00	0.82	0.66	0.64	0.71		0.77	0.77
	Recall	1.00	0.32	0.88	0.86	0.67		0.75	0.75
	F1-score	1.00	0.46	0.75	0.73	0.69	0.75	0.73	0.73
Logistic regression classifier	Precision	1.00	0.16	0.44	0.57	0.62		0.56	0.56
	Recall	1.00	0.04	0.61	0.99	0.37		0.60	0.61
	F1-score	1.00	0.06	0.51	0.72	0.46	0.61	0.55	0.56
Support vector classifier	Precision	1.00	0.21	0.43	0.59	0.62		0.57	0.57
	Recall	1.00	0.08	0.56	0.93	0.46		0.60	0.61
	F1-score	1.00	0.11	0.49	0.72	0.53	0.61	0.57	0.57
Random forest classifier	Precision	1.00	0.75	0.66	0.61	0.79		0.76	0.76
	Recall	1.00	0.36	0.78	0.94	0.61		0.74	0.74
	F1-score	1.00	0.49	0.71	0.74	0.69	0.74	0.73	0.73

amylose concentration, i.e. classes 0, 2 and 3 respectively. However, the model misclassifies the remaining two classes of amylose concentration, viz. classes 1 and 4. This is due to the overlapping of data points into another type, and the model cannot classify the category of amylase, which can be seen in the decision boundary diagram of the model. A study reported that the existing classification algorithms are extended from traditional classification algorithms for specific data and deal with data uncertainty based on relatively ideal probability distribution and data type assumptions[12]. The unclear data were classified using ensemble models in a unique potential world-based AdaBoost technique called PwAdaBoost to improve accuracy. A similar kind of study has been conducted by various researchers[13,14]. They used the AdaBoost algorithm for early detection and diagnosis of breast cancer and seabed classification using the PSO-BP-AdaBoost algorithm with better model accuracy.

ANN classifier

The classification of amylose concentration was carried out using the ANN classifier model. The analysis was conducted with the intensity data by changing the values of the number of hidden layers and the number of neurons in each hidden layer. The error rate pertaining to each layer and neuron was recorded (Figure 3 *a*). The figure shows that the error rate is low for four hidden layers, with 300,

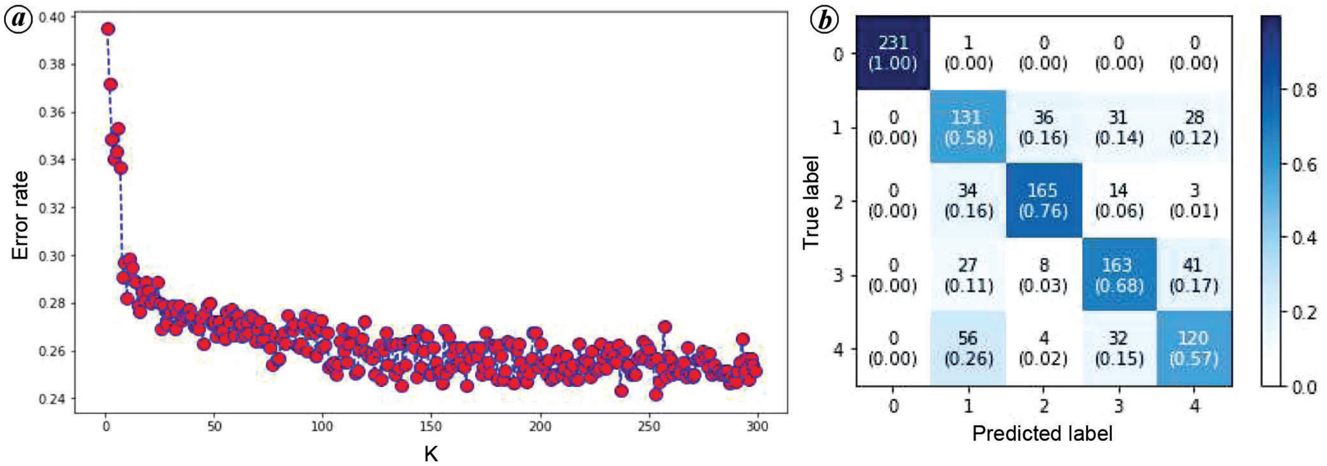


Figure 3. (***a***) Error plot and (***b***) confusion matrix plot for ANN classifier.

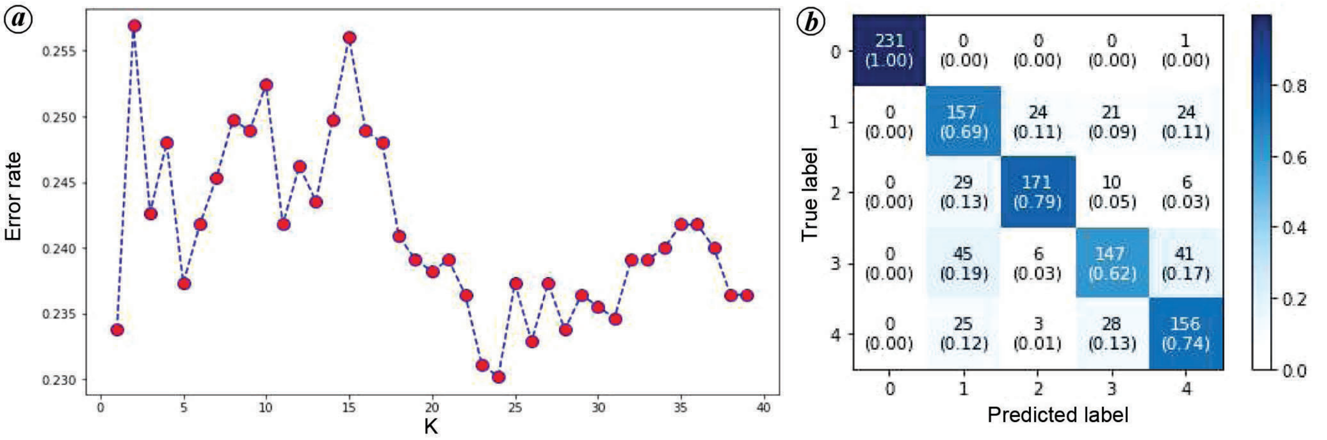


Figure 4. (***a***) Error plot and (***b***) confusion matrix plot for KNN classifier.

100, 50 and 10 neurons in each layer considered for auxiliary analysis. For the optimized value of hidden layers and neurons, the accuracy of the classifier is 0.77. The results have been represented in the form of a confusion matrix (Figure 3 *b*) and an accuracy table (Table 1). From Table 1, it can be observed that the average precision of the ANN model is 0.77. However, the average recall and F1 scores of the model are 0.77 and 0.77 respectively. Meanwhile, the precision values of the model in identifying the selected concentration of amylose are 1.00, 0.65, 0.78, 0.72 and 0.69 respectively, for classes 0, 1, 2, 3 and 4. The recall and F1 scores for classes 0, 1, 2, 3 and 4 are 1.00, 0.55, 0.82, 0.71 and 0.78, and 1.00, 0.59, 0.80, 0.71 and 0.74 respectively.

Figure 3 *b* reveals that the ANN model classifies different concentrations of amylose data with better accuracy compared to the AdaBoost model, which is seen in the confusion matrix. Though the classification efficiency is good, the model has misclassified the different concentrations due to the overlapping data, leading to a higher error rate. Various researchers have adopted deep neural networks for intrusion detection in the information systems of web-based data. Similar studies have been conducted to classify coffee bean species, fish species, EEG signals for epileptic seizures and pneumonia using neural network algorithms[15–19].

KNN classifier

The classification of amylose concentration was carried out using the KNN classifier model. The classification analysis was conducted with the intensity data by changing the *n*-neighbour values in the range 1–50, and the error rate pertaining to each *n*-neighbour was recorded (Figure 4 *a*). From the figure, it can be observed that the error rate is low when the *n*-neighbour value is 1, and this value is considered for auxiliary analysis. For the optimized *n*-neighbour value, the accuracy of the classifier is 0.77. The results are represented as confusion matrix (Figure 4 *b*) and an accuracy table (Table 1). From the table, it can be seen that the average precision of the KNN model is 0.77. However, the average recall and F1 score of the model are 0.77 and 0.77

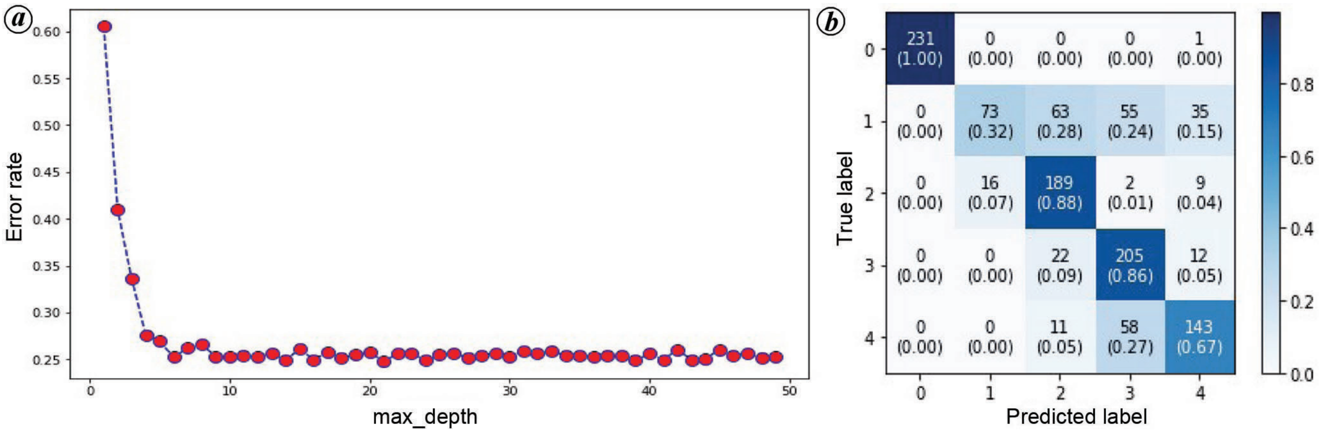


Figure 5. (***a***) Error plot and (***b***) confusion matrix plot for decision tree classifier.

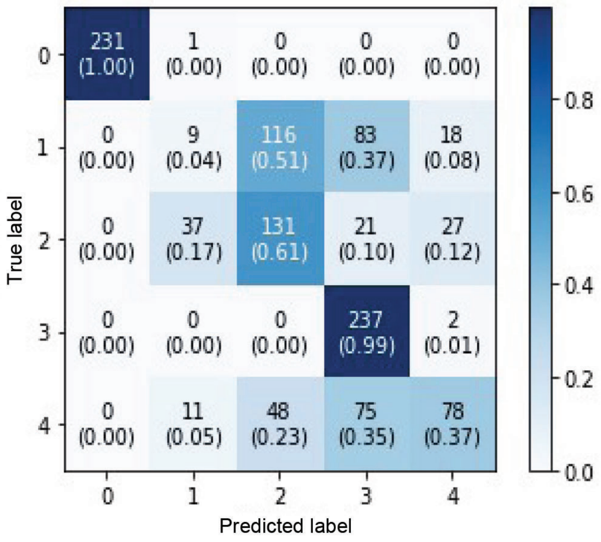


Figure 6. Confusion matrix plot for logistic regression classifier.

respectively. The precision values of the model in identifying the selected concentration of amylose are 1.00, 0.61, 0.84, 0.71 and 0.68 respectively, for classes 0, 1, 2, 3 and 4. The recall and F1 scores for classes 0, 1, 2, 3 and 4 are 1.00, 0.69, 0.79, 0.62 and 0.74 and 1.00, 0.65, 0.81, 0.66 and 0.71 respectively.

The confusion matrix depicted in Figure 4 *b* reveals that the classification accuracy of the model is good compared to the previous models. The overlapping and misclassification of amylose concentration are less, leading to better accuracy of this model. The decision boundary of the model depicts that most of the data points fall in their respective classes to achieve better accuracy of 0.77. A study reported that an automatic real-time recommendation system adopting KNN model using the Euclidean distance could classify the actual classes accurately[20]. Similar studies reveal that the KNN algorithm can also be used to classify text and textural documents into the desired classes accurately[21,22].

Decision tree classifier

The classification of amylose concentration was carried out using the decision tree classifier model. The analysis was done with the intensity data by changing the maximum depth of the tree in the range of 1–50. The error rate pertaining to each depth was recorded (Figure 5 *a*). The figure shows that the error rate is low when the maximum depth is 6, and this value is considered for all auxiliary analyses. For the optimized value of maximum depth, the accuracy of the classifier is 0.75. The results are represented in the form of a confusion matrix (Figure 5 *b*) and an accuracy table (Table 1). From the table, it can be concluded that the average precision of the decision tree model is 0.77. However, the average recall and F1 scores of the model are 0.75 and 0.73 respectively. The precision values of the model in identifying the selected concentration of amylose are 1.00, 0.82, 0.66, 0.64 and 0.71 respectively, for classes 0, 1, 2, 3 and 4. The recall and F1 scores for classes 0, 1, 2, 3 and 4 are 1.00, 0.32, 0.88, 0.86 and 0.67, and 1.00, 0.46, 0.75, 0.73 and 0.69 respectively.

Figure 5 *b* reveals that the decision tree model cannot classify the amylose content, but the classification accuracy for amylose concentration classes 0 and 3 is excellent. Except for these two, the remaining concentrations of amylose are misclassified, which might be due to the overlapping of data onto another class as well as the efficiency of the model to distinguish the different categories. Similar research demonstrated that the rules-based decision tree models in a hierarchical intrusion detection system may be classified well using the decision tree method[23–25]. Further, the model's accuracy depends on the nature and amount of the input datasets.

Logistic regression classifier

The classification of amylose concentration was carried out using logistic regression classifier model. The analysis was

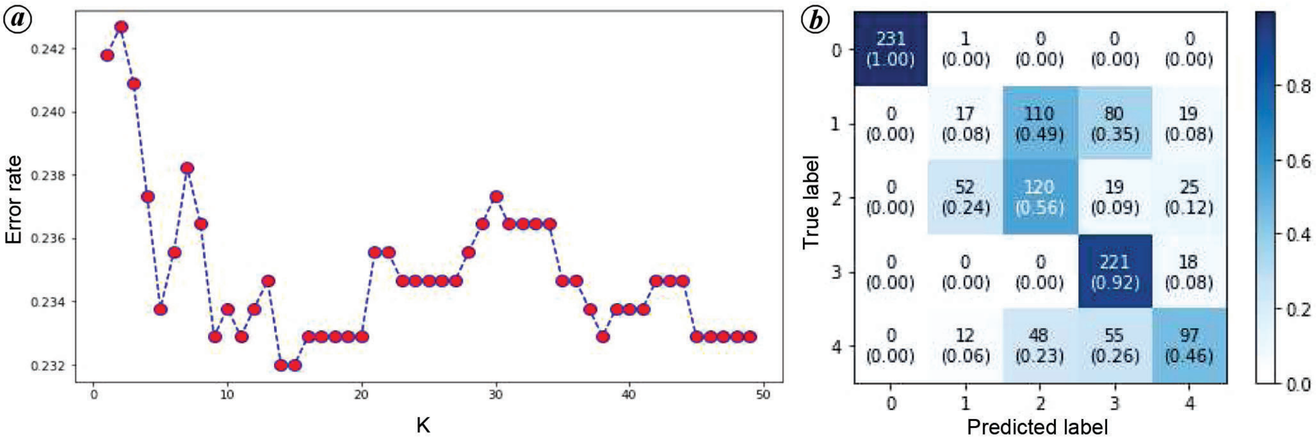


Figure 7. (***a***) Error plot and (***b***) confusion matrix plot for support vector classifier.

done with the intensity data, and the regression model was evaluated in terms of accuracy. The results depict that the accuracy of the classifier is 0.61. The results have been represented in the form of a confusion matrix (Figure 6 *a*) and an accuracy table (Table 1). From the table, it is observed that the average precision of the logistic regression model is 0.56. However, the average recall and F1 scores are 0.61 and 0.56 respectively. The precision values of the model in identifying the selected concentration of amylose are 1.00, 0.16, 0.44, 0.57 and 0.62 for classes 0, 1, 2, 3 and 4 respectively. The recall and F1 scores for classes 0, 1, 2, 3 and 4 are 1.00, 0.04, 0.61, 0.99 and 0.37, and 1.00, 0.06, 0.51, 0.72 and 0.46 respectively.

The confusion matrix of the logistic regression classifier shows that the model classifies only two concentration classes of amylose accurately, while the remaining are misclassified. This shows that the data points are displayed on several classes, which lowers the model's accuracy because the model only has a very limited capacity to identify the data on other amylose concentrations. Studies on imbalanced data classification, data classification and ML approaches for the classification of credit score data reveal that logistic regression could help classify the given input datasets[26–28].

Support vector classifier

The classification of amylose concentration was carried out using the SVC model. The analysis was done with the intensity data by changing the kernel values in the range of 1–50, and the error rate pertaining to each kernel was recorded (Figure 7 *a*). From the figure, it is observed that the error rate is low when the kernel value is 15, and this value is considered for all auxiliary analyses. For the optimized kernel value, the accuracy of the classifier is 0.61. The results are represented in the form of a confusion matrix (Figure 7 *b*) and an accuracy table (Table 1). From the table, it can be observed that the average precision of the SVC model is 0.57. However, the average recall and F1 scores are 0.61 and 0.57 respectively. The precision values of the model in identifying the selected concentration of amylose are 1.00, 0.21, 0.43, 0.59 and 0.62 for classes 0, 1, 2, 3 and 4 respectively. The recall and F1 scores for classes 0, 1, 2, 3 and 4 are 1.00, 0.08, 0.56, 0.93 and 0.46, and 1.00, 0.11, 0.49, 0.72 and 0.53 respectively.

In the plotted confusion matrix of the classifier, it is seen that only two classes of amylose concentration are classified accurately (i.e. 0 and 3) by the SVC model; however more than 50% of the data has been misclassified in the remaining three classes. The misclassified data of the remaining three classes can be seen in the decision boundary plot, which shows overlapping data on the other classes. Similar studies have indicated that the SVC model could be used efficiently to classify imbalanced data by support vector machines and diagnose liver disease in cows[29,30].

Random forest classifier

The classification of amylose concentration was carried out using a random forest classifier model. The classification analysis was done with the intensity data by changing the maximum depth values and value of the *n*-estimator in the range 1–50. The error rate pertaining to these was recorded (Figure 8 *a*). From the figure, it is observed that the error rate is low when the maximum depth value is 5 with the *n*-estimator value of 22, which was considered for all auxiliary analyses. For the optimized depth and *n*-estimator values, the accuracy of the classifier was estimated to be 0.74. The results are represented in the form of a confusion matrix (Figure 8 *b*) and an accuracy table (Table 1). The table shows that the average precision of the random forest classifier model is 0.76. However, the average recall and F1 scores are 0.74 and 0.73 respectively. The precision values of the model in identifying the selected concentration of amylose are 1.00, 0.75, 0.66, 0.61 and 0.79 respectively, for the classes 0, 1, 2, 3 and 4. The recall and F1 scores for classes 0,

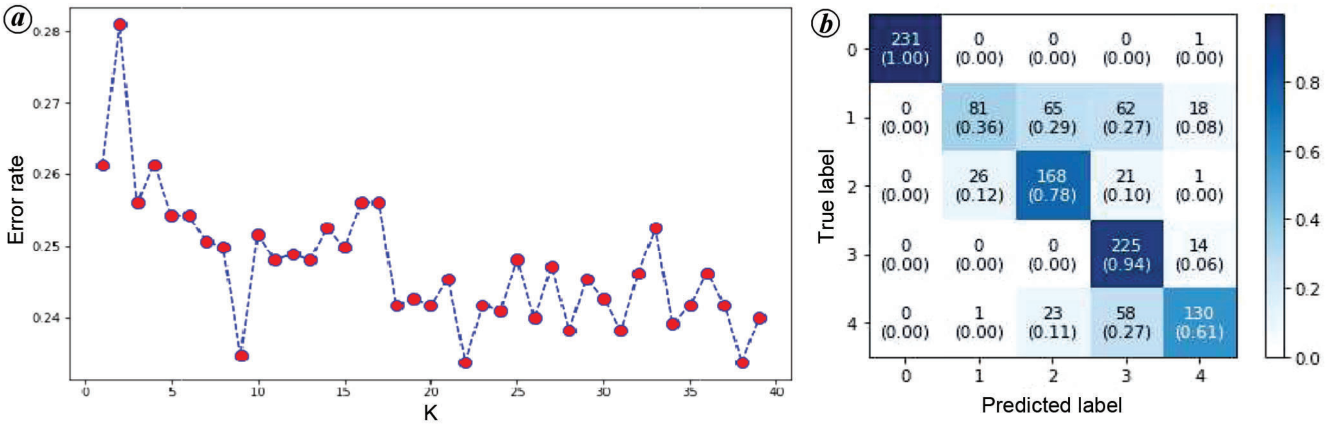


Figure 8. (*a*) Error plot and (*b*) confusion matrix plot for random forest classifier.

1, 2, 3 and 4 are 1.00, 0.36, 0.78, 0.94 and 0.61, and 1.00, 0.49, 0.71, 0.74 and 0.69 respectively.

The confusion matrix of the random classifier demonstrates that only two amylose concentration classes, i.e. 0 and 4 are classified accurately, whereas the accuracy of class 2 is comparatively good. The remaining amylose classes are misclassified. From Figure 8 *a,* it is seen that the data points of classes 1 and 4 are plotted on the surface of other classes, leading to lower classification accuracy. Random forest algorithms have also been used to classify neuro-imaging data in Alzheimer's disease, as well as big data classification in IoTs and other applications[31–33].

From the above details pertaining to the performance of the evaluated ML models, it can be observed that for the light intensity dataset obtained from the sensor, higher and similar model parameters and accuracy value of 0.77 are reported for both the ANN and KNN algorithms followed by the accuracy value of 0.75, 0.74, 0.65, 0.61 and 0.61 respectively, for the decision tree, random forest, AdaBoost, logistic regression and SVC algorithms. Though the ANN and KNN models achieve higher accuracy in classifying the different concentrations of amylose, the error rate is high because of the overlapping of the data points at a particular time period. This leads to a high error rate in classifying the amylose concentration. Further, the low model accuracy might be due to higher overlapping, as the models are unable to perform well under overlapping conditions leading to an erroneous classification. From these results, we conclude that the ANN and KNN algorithms can be used for predicting the concentration of amylose content in rice samples using the intensity dataset obtained from the sensor.

Conclusion

Thus a rapid and easy method for estimating the amylose content to assess ageing in rice has been achieved by developing a colorimetric amylose sensor. The selected ML models were evaluated using the intensity data obtained by the sensor for proper interpretation. It is concluded that with the same accuracy, the ANN and KNN models could predict amylose content more accurately and thus the ageing of rice. Since the accuracy of the ANN and KNN models are the same, the KNN algorithm is recommended to classify the amylose content due to the higher accuracy of each class. Though the model accuracy is comparatively good, it can be improved by changing the various model parameters, preprocessing the data, and using deep learning models.

1. Perez, C. M. and Juliano, B. O., Texture changes and storage of rice. *J. Texture Stud.*, 1981, **12**(1), 321–333.
2. Faruq, G., Prodhan, Z. H. and Nezhadahmadi, A., Effects of ageing on selected cooking quality parameters of rice. *Int. J. Food Prop.*, 2015, **18**(4), 922–933.
3. Zhou, Z., Robards, K., Helliwell, S. and Blanchard, C., Ageing of stored rice: changes in chemical and physical attributes. *J. Cereal Sci.*, 2001, **35**(1), 65–78.
4. Devraj, L., Natarajan, V., Ramachandran, S. V., Manicakam, L. and Saravanan, S., Accelerated aging by microwave heating and methods to distinguish aging of rice. *J. Food Process Eng.*, 2020, **43**(6), 13405–13415.
5. Popa, A. *et al.*, An intelligent IoT-based food quality monitoring approach using low-cost sensors. *Symmetry*, 2019, **11**(3), 374–391.
6. Moradi, M., Balanian, H., Taherian, A. and Mousavi Khaneghah, A., Physical and mechanical properties of three varieties of cucumber: a mathematical modeling. *J. Food Process Eng.*, 2020, **43**(2), 13323–13330.
7. Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O. and Akinjobi, J., Supervised machine learning algorithms: classification and comparison. *Int. J. Comput. Trends Technol.*, 2017, **48**(3), 128–138.
8. Celine, S., Maria Dominic, M. and Savitha Devi, M., Logistic regression for employability prediction. *Int. J. Innov. Technol. Exp. Eng.*, 2020, **9**(3), 2471–2478.
9. Anon., AdaBoost Classifier in Python, 2018; https://www.datacamp.com/community/tutorials/adaboost-classifier-python (accessed on 1 May 2021).
10. Anon., Understanding logistic regression in Python, 2019; https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python (accessed on 1 May 2021).
11. Anon., Understanding random forests classifiers in Python, 2018; https://www.datacamp.com/community/tutorials/random-forests-classifier-python (accessed on 1 May 2021).

12. Liu, H., Zhang, X. and Zhang, X., PwAdaBoost: possible world based AdaBoost algorithm for classifying uncertain data. *Knowl.-Based Syst.*, 2019, **186**, 104930.
13. Zheng, J., Lin, D., Gao, Z., Wang, S., He, M. and Fan, J., Deep learning assisted efficient AdaBoost algorithm for breast cancer detection and early diagnosis. *IEEE Access*, 2020, **8**, 96946–96954.
14. Ji, X., Yang, B. and Tang, Q., Acoustic seabed classification based on multibeam echosounder backscatter data using the PSO-BP-AdaBoost algorithm: a case study from Jiaozhou Bay, China. *IEEE J. Ocean. Eng.*, 2020, **46**(2), 509–519.
15. Kim, J., Shin, N., Jo, S. Y. and Kim, S. H., Method of intrusion detection using deep neural network. In IEEE International Conference on Big Data and Smart Computing, 2017, pp. 313–316.
16. Arboleda, E. R., Fajardo, A. C. and Medina, R. P., Classification of coffee bean species using image processing, artificial neural network and *k*-nearest neighbors. In IEEE International Conference on Innovative Research and Development, Bangkok, Thailand, 2018, pp. 1–5.
17. Siddiqui, S. A., Salman, A., Malik, M. I., Shafait, F., Mian, A., Shortis, M. R. and Harvey, E. S., Automatic fish species classification in underwater videos: exploiting pre-trained deep neural network models to compensate for limited labelled data. *ICES J. Mar. Sci.*, 2018, **75**(1), 374–389.
18. Narang, A., Batra, B., Ahuja, A., Yadav, J. and Pachauri, N., Classification of EEG signals for epileptic seizures using Levenberg–Marquardt algorithm based multilayer perceptron neural network. *J. Intell. Fuzzy Syst.*, 2018, **34**(3), 1669–1677.
19. Stephen, O., Sain, M., Maduh, U. J. and Jeong, D. U., An efficient deep learning approach to pneumonia classification in healthcare. *J. Healthcare Eng.*, 2019; https://doi.org/10.1155/2019/4180949.
20. Adeniyi, D. A., Wei, Z. and Yongquan, Y., Automated web usage data mining and recommendation system using *k*-nearest neighbor (KNN) classification method. *Appl. Comput. Informat.*, 2016, **12**(1), 90–108.
21. Shah, K., Patel, H., Sanghvi, D. and Shah, M., A comparative analysis of logistic regression, random forest and KNN models for the text classification. *Augment. Hum. Res.*, 2020, **5**(1), 1–16.
22. Moldagulova, A. and Sulaiman, R. B., Using KNN algorithm for classification of textual documents. In Eighth International Conference on Information Technology, 2017, pp. 665–671.
23. Gupta, B., Rawat, A., Jain, A., Arora, A. and Dhami, N., Analysis of various decision tree algorithms for classification in data mining. *Int. J. Comput. Appl.*, 2017, **163**(8), 15–19.
24. Ahmim, A., Maglaras, L., Ferrag, M. A., Derdour, M. and Janicke, H., A novel hierarchical intrusion detection system based on decision tree and rules-based models. In 15th International Conference on Distributed Computing in Sensor Systems, 2019, pp. 228–233.
25. Abdallah, I. *et al.*, Fault diagnosis of wind turbine structures using decision tree learning algorithms with big data. *Safety and Reliability – Safe Societies in a Changing World*, CRC Press, 2018, pp. 3053–3061.
26. Ohsaki, M., Wang, P., Matsuda, K., Katagiri, S., Watanabe, H. and Ralescu, A., Confusion matrix-based kernel logistic regression for imbalanced data classification. *IEEE Trans. Knowl. Data Eng.*, 2017, **29**(9), 1806–1819.
27. De Menezes, F. S., Liska, G. R., Cirillo, M. A. and Vivanco, M. J., Data classification with binary response through the boosting algorithm and logistic regression. *Expert Syst. Appl.*, 2017, **69**, 62–73.
28. Dumitrescu, E., Hue, S., Hurlin, C. and Tokpavi, S., Machine learning for credit scoring: improving logistic regression with non-linear decision-tree effects. *Eur. J. Oper. Res.*, 2021, **297**(3), 1178–1192.
29. Mathew, J., Pang, C. K., Luo, M. and Leong, W. H., Classification of imbalanced data by oversampling in kernel space of support vector machines. *IEEE Trans. Neural Networks Learn. Syst.*, 2017, **29**(9), 4065–4076.
30. Devikanniga, D., Ramu, A. and Haldorai, A., Efficient diagnosis of liver disease using support vector machine optimized with crows search algorithm. *EAI Endorsed Trans. Energy Web*, 2020, **7**(29), 1–10.
31. Sarica, A., Cerasa, A. and Quattrone, A., Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: a systematic review. *Front. Aging Neurosci*., 2017, **9**, 329–340.
32. Lakshmanaprabu, S. K., Shankar, K., Ilayaraja, M., Nasir, A. W., Vijayakumar, V. and Chilamkurti, N., Random forest for big data classification in the internet of things using optimal features. *Int. J. Mach. Learn. Cybern.*, 2019, **10**(10), 2609–2618.
33. Demidova, L. A., Klyueva, I. A. and Pylkin, A. N., Hybrid approach to improving the results of the SVM classification using the random forest algorithm. *Proc. Comput. Sci.*, 2019, **150**, 455–461.

ACKNOWLEDGEMENTS. We thank Karnataka Science and Technology Promotion Society (KSTePS) for the award of DST-Ph.D. fellowship for Science and Engineering Students. We also thank Dr K. Mahantshivayogayya, Scientist (Agricultural Research Station, Gangavati, University of Agricultural Sciences, Raichur) for providing different varieties of rice samples while conducting this study.

Received 3 June 2022; revised accepted 20 December 2022

doi: 10.18520/cs/v124/i6/722-730