Journal of Scientific & Industrial Research
Vol. 83, January 2024, pp. 102-113
DOI: 10.56042/jsir.v83i1.2458

Identification of Suitable Complex Machine Learning Algorithms for Amylose Content Prediction in Rice with an IoT-based Colorimetric Sensor

Shrinivas Deshpande[1]*, Udaykumar Nidoni[2], Rahul Patil[3], Sharanagouda Hiregoudar[2], Ramappa K T[2], Devanand Maski[4] & Nagaraj Naik[5]

[1]ICAR-Krishi Vigyan Kendra, Kandali, Hassan 573 217, UAS, GKVK, Bangalore, Karnataka, India

[2]Dept. of Processing and Food Engineering, [3]Dept. of Soil & Water Conservation Engineering, [4]Dept. of Renewable Energy Engineering, [5]Pesticide Residue and Food Quality Analysis Laboratory, CAE, UAS, Raichur 584 104, Karnataka, India

Received 17 August 2022; revised 19 July 2023; accepted 24 December 2023

Rice ageing is a complex phenomenon that is hard to investigate thoroughly. Many physicochemical qualities change gradually because of moisture content and storage temperature. Among these characteristics, amylose quantity is particularly essential, and most indexes rely on it. To address these challenges, various gadgets, IoT, ICT, AI and predictive technologies are frequently applied in diagnostic procedures. This study evaluated AdaBoost, Artificial neural network (ANN), k-Nearest Neighbour classifier (KNN), Decision tree, Logistic regression, Support Vector Machine (SVM), and Random forest classifiers to categorize distinct quantities of amylose using slope data gathered from the novel colorimetric amylose sensor. The random forest approach had greater coefficients and precision ratings of 0.85 for the slope dataset, followed by the decision tree, ANN, KNN, AdaBoost, logistic regression, and support vector algorithms, which had precision scores of 0.83, 0.81, 0.80, 0.29, 0.18, and 0.18, respectively, based on the efficiency of the tested learning models. The random forest model was shown to be promising in forecasting the various classes of amylose based on the data.

Keywords: Ageing of rice, Amylose sensor, IoT device, Mathematical modeling, Rice quality

Introduction

"Ageing" denotes to the biological alterations imposed on by the moisture content, temperature, and grain variety as kernels are stored. When the newly picked paddy gets cooked, it produces a crumbly gruel that people either enjoy or dislike depending on their dietary preferences. When stored properly, these characteristics deteriorate with time and the grains are less likely to clump together when cooked. Amylose levels may gradually increase with age, and alterations to lipid, protein, and other components brought on by storage-related enzyme activity may also occur.[1] Lipids yield free-fatty acids and aggregates by means of the amylase enzyme, alongside with carbonyl groups and hydro-peroxides. As a consequence, the amount of amylose in the rice might be used to gauge how quickly something is ageing. Because of such attributes, aged rice is preferred to fresh rice in Asian countries even though it has a superior taste, texture, and flavour. It produces cooked grains that are flaky or gritty because of its elevated kernel-elongation, absorption of water, expansion of volume, and poorer dissolvable solid contents.[2] As a result, fresher rice is favoured in China, Japan, and various other nations, while older rice is more prohibitively costly, particularly in India.[3]

Assessing cooking attributes and testing the toughness of cooked kernels by squeezing the kernel against the palm are two steps in the traditional method of detecting the age of rice. As a result, farmers' commodities are priced using an unscientific method. These issues make it challenging to improve the economic situation of farmers. Alternative techniques and scientific instruments based on qualitative and factual evaluation methods are in high demand for assessing rice ageing.[4]

Science and technology breakthroughs, as well as increased human capital, have allowed the world economy to flourish sustainably, culminating in the emergence of the intelligent technological strategy. Sensor-based computer programmes may offer more detailed information on any parameter. In light of these factors, agricultural technology that utilizes the Internet-of-Things (IoT) proves more cost-effective than traditional techniques for addressing genuine challenges.[5] Moreover, IoT-based sensors, instruments,

*Author for Correspondence
E-mail: shrihd9@gmail.com

devices, or other electronic packages may increase quality of agricultural production by removing the need for human input.

In food technology, mathematical models are equations that help in the simulation of scientific methods and, as a consequence, lower the increasing costs for a variety of activities. Some study has been conducted in this regard in order to propose appropriate models for post-harvesting activities in order to identify and construct the ideal link between effective factors.[6]

One of the areas of computer science that is expanding the fastest and has the broadest range of applications is machine learning. It speaks of the robotic identification of pertinent data patterns. Machine learning technologies seek to give programmes with the ability to grow and change.[7] There is reason to anticipate that sophisticated data analytics will become a crucial part of technological growth as a result of the availability of ever-increasing amounts of data.[6] In light of this, the study aimed to assess the capacity of an array of sophisticated machine-learning algorithms to reliably estimate the amylose concentration in rice specimens employing slope data collected by a newly built colourimetric amylose sensor.

Materials and Methods

An Explanation of the Novel Amylose Sensor

The enzyme mimicking property exhibited by the 3,3',5,5'-tetramethylbenzidine (TMB) in the presence of H_2O_2, hydrophilic bentonite clay (nano), and glucose was utilized to develop the color-based sensor. An Android mobile application was developed for data visualization and collection (Fig. 1) for rapid measurement of amylose content intern the ageing of rice.[8] Initially the sensor was programmed with the standard amylose at various level of concentration to establish the relationship between the amylose content and the color of the test solution and is calibrated. After calibration of the device, the testing was done with the selected rice cultivar *i.e.* BPT 5204. The developed device generated two types of data sets after testing each sample *i.e.* light intensity data and slope values. In our previous report, it was observed that the accuracy of the classification model was limited to 0.77 for the data tested with light intensity data.[8] In order to increase the accuracy of the model, the slope data set was tested with seven selected machine learning algorithms which were used in our previous study[8] to classify the actual amylose value in the test sample.

Experimentation and Data Gathering

The 652 nm LED light source was used for analyzing the sample and slope data were recorded in the android application as a graphical user interface and the data were collected as a function of time for 900 seconds for each sample. These data were subjected to normalization followed by dividing into training, testing & validation dataset in the ratio of 70:15:15. The segregated data were further evaluated with seven different machine learning algorithms. In order to improve the accuracy of the algorithm, several model parameters were tuned represented in the form of error graphs. The tuning parameter with lower error was selected for further experimentation. The confusion matrix and decision boundary plots were generated to visualize the data. Finally, the model with higher accuracy, precision and lower error

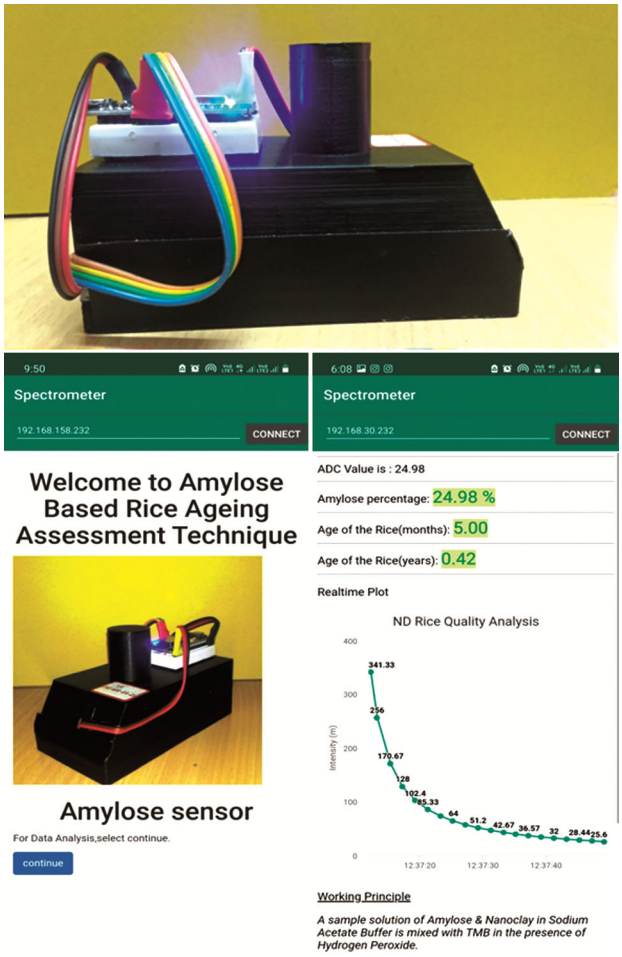

Fig. 1 — Colourimetric amylose sensor and android mobile application

rate was selected as best model to classify the slope data received from the developed sensor. The classification models selected for the study are discussed hereunder.

Mathematical Modeling

Adaboost Classifier

Adaboost, often called adaptive boosting, is a method that employs a number of weak classifiers to produce a superior classifier. Furthermore, the bagging technique (combining various learner models) lowers variance, the boosting technique (set of the least accurate classifier models) reduces bias, and the layering approach predicts better (combining multiple base classification models). Initially, the classifier will select the training subset at random and train the model repeatedly by selecting training set data based on the accuracy of prior training. The model assigns weights to the learnt classifier data depending on predictive performance in each iteration.[10]

ANN Classifier

The Artificial Neural Network (ANN) is a data processing paradigm that works in the same way as the biological neuron, namely the brain. The input layers supplied the desired independent variables (input variables), whilst the hidden layers processed and translated the input data into a readable format by assigning synaptic weights to the specific set of data based on their unique strength. The weighted synapses of all input values are calculated and delivered directly to the output layer using neural networks. Moreover, the selected activation function converts the input signal of an ANN node to an output signal. This output signal is used as an input to the next rung of the stack. Additionally, inserting hidden neurons between both the input and output layers may increase the ANNs' accuracy.[9]

KNN Classifier

K-Nearest Neighbor (KNN) is a quasi and lazy learning model that does not make explicit hypotheses concerning the dispersion of the raw data or the architecture of the model that is created from the dataset. The quantity of Nearest Neighbours affects how accurate a KNN model is. Each object votes for its own class, and the prediction is made using the class with the most votes. The Euclidean distance, Hamming distance, Manhattan distance, and Minkowski distance were utilized to find the places that were closest to each other. Since Euclidean distance is a frequently employed and preferred function in KNNs, it was taken into consideration in this study.[9]

Decision Tree Classifier

A decision tree comprises a tree architecture that looks like a roadmap and is a basic and commonly used classification process for classifying and predictive tasks. In a repetitive process called as recursive partitioning, the model learns to separate the inputs relying on feature value. The core working mechanism of the decision tree algorithm begins with picking the appropriate feature among the raw data employing Attribute Selection Measures (ASM) to split the data, preceded by partitioning the set of data into disjoint sets as decision nodes. The technique described above is repeated for each node in the model until all nodes are members of the same characteristic, there are no further attributes left, or there are no occurrences left. The decision tree and dataset's confusion matrix were created using the best ASM with the greatest attribute score once the model's training and testing phases were complete.[9]

Logistic Regression Classifier

The most common use of the linear regression technique known as logistic regression is the prediction of binaries with categorical results or predictor variable. Maximum Likelihood Estimation (MLE), a maximization strategy that employs a sigmoid function to identify the variable most likely to yield the measured values, is used by the logistic regression model to predict the output, with mean and variance serving as key determinants. Since the target variables in this experiment included 5 nominal categories of amylose concentrations, the multinomial logistic regression model was adopted. Any real-valued integer may be computed using the sigmoid or logistic function, which produces a "S"-shaped curve that converts it to a number between 0 and 1. Y anticipated becomes one if the curve reaches positive infinity, and zero if the curve reaches negative infinity. The outcome is 1 or YES if the sigmoid function output is more than 0.5; the outcome is 0 or NO if it is less than 0.5.[(11)]

Support Vector Classifier

SVM is a classifying approach that is frequently employed in applications for both regression and classification. Numerous categorical and continuous variables may be handled by the SVM. By establishing a hyperplane in higher dimensional space

to identify various classes of variables, SVM performed data classification. Ideal hyperplanes are created repeatedly using SVM and used to lower error. Finding the ideal Maximum Marginal Hyperplane (MMH) for classifying a dataset into different groups was the fundamental idea underlying SVM. The identification of an ideal hyperplane with largest practical space between support vectors was thought to be the key to how SVM worked. The hyperplane with the highest segregation or a larger separation from the closest data points was picked for data classification.[9]

Random Forest Classifier

The random forests approach creates decision trees from data points that are selected at random, gets a prognosis by each tree, and then asks for votes to determine which the great decision is. According to the kind of dataset, the entire dataset was first split into the proper amount of decision trees. The forest, based on a random sample, is the collective name for these decision trees. Each tree casts a vote in a classification problem, and the class that receives the most votes is picked as the final result. The prognosis with the best scores is then taken as the final projection after voting on each anticipated result.[12]

Results and Discussion

Adaboost Classifier

By adjusting the n-estimator constants within the region of 1 to 50, the amylose level was classified using the Adaboost classification algorithm employing the slope dataset. The error margin corresponding to each n-estimator was obtained and shown in Fig. 2(a). As seen in the picture, the n-estimator value of 18 was chosen again for ancillary evaluation since it had a relatively low level of error.

The classifier's accuracy was 0.29 for the ideal number of n-estimators, and its findings are displayed in Fig. 2(b) and in Table 1. The table illustrates that the Adaboost model's mean accuracy was 0.29. The model's mean F_1 and recall scores were 0.17 and 0.29, respectively or the classes 0, 1, 2, 3 and 4. The model's accuracy values for determining the appropriate concentration of amylose were reported to be 0.31, 1.00, 0.23, 0.00, and 0.00, respectively. Recall and F_1 indices were found to be 0.99, 0.04, 0.40, 0.00, and 0.00 and 0.48, 0.08, 0.29, 0.00, and 0.00 for groups 0, 1, 2, 3, and 4, respectively. The decision boundary was depicted well with distinct classes indicated by specific color codes, accompanied by a scatter graph of the actual data

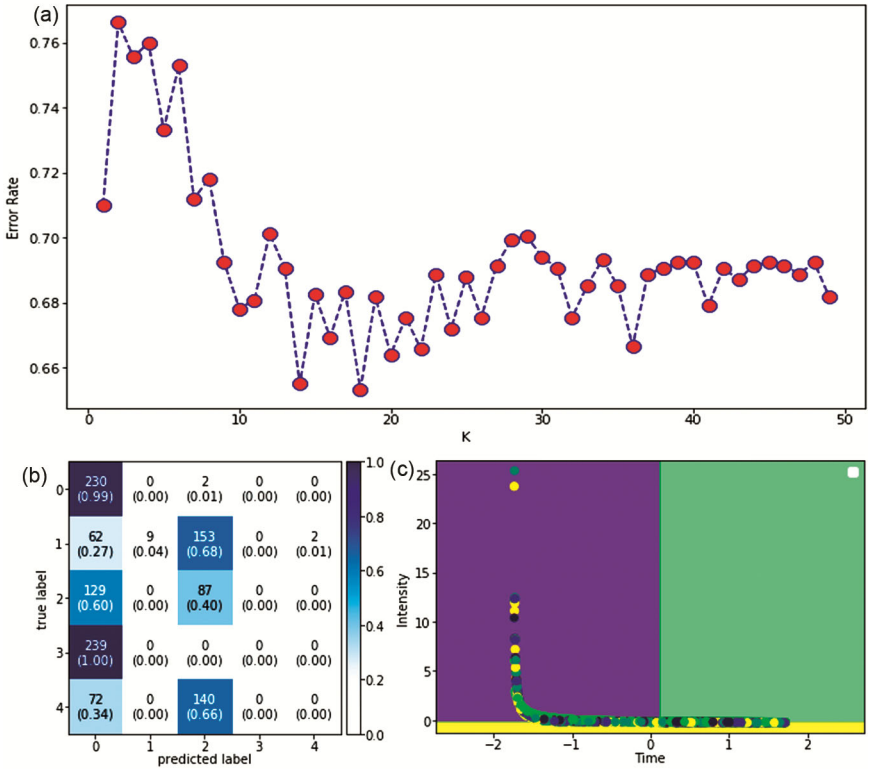

Fig. 2 — Error plot, confusion matrix and decision boundary plot obtained for AdaBoost Classifier

points for a certain level, as shown in Fig. 2 to see the data distribution pattern (c).

It was clear through the table and charts that evaluating the model using slope data resulted in a greater rate of errors. Furthermore, while estimating the required quantity of amylose in the sample, the algorithm's precision, recall, and F_1-value were lower than expected. The accuracy of amylose forecasting utilizing slope values was too minimal even if the Adaboost model was used to characterize the fuzzy figures using a novel potential world-based Adaboost approach called PwAdaboost. With the PSO-BP-Adaboost technique, it is reported data that were equivalent.[13,14] However, the PSO-BP-Adaboost strategy was shown to be erroneous in predicting the right class of amylose.

ANN Classifier

Through the evaluation process, the amount of hidden layers and the corresponding number of neurons in each layer changed in order to improve the model's precision, and the corresponding error rates were displayed Fig. 3(a). From the picture, it can be seen that the four hidden layers with 150, 50, 20 and 10 neurons each were chosen for future investigations since they had the lowest error rate values. The results were presented as a confusion matrix in Fig. 3(b) whose associated values can be found in Table 1. The effectiveness of the ANN was determined to be 0.81 for the optimal hidden layers and nodes. Again from table, it was deduced that ANN model's mean precision was 0.81. Nevertheless, the model's mean recall and F_1 score were 0.81 and 0.80, respectively.

Meanwhile, it emerged that, for classes 0, 1, 2, 3, and 4, the model's accuracy coefficients for identifying the selected level of amylose were 0.85, 0.78, 0.85, 0.93, and 0.64, accordingly. For classes 0 to 4, the recall and F_1-scores were determined to be 0.96, 0.50, 0.85, 0.98, and 0.75; and 0.90, 0.61, 0.85, 0.96, and 0.69, correspondingly. The decision boundary was depicted in the various classes indicated by distinct colors, accompanied by a scatter diagram of the recorded sample points for each level, as shown in Fig. 3(c) to see the data distribution pattern.

The aforementioned table and figures demonstrate that when the model was tested with slope data, the percentage of error decreased significantly. Furthermore, when detecting the target concentration of amylose in the sample using slope data, all model characteristics were greater. In light of the aforementioned model parameter, a greater model accuracy of 0.81 was found for slope data. The

Table 1 — Evaluation results of selected machine learning models with slope data obtained from colourimetric amylose sensor

Models		Classes (with amylase content, mg·mL^{-1})					Accuracy	Macro average	Weighted average
		0 (0.2)	1 (0.4)	2 (0.6)	3 (0.8)	4 (1.0)			
	Precision	0.31	1.00	0.23	0.00	0.00		0.31	0.31
Adaboost classifier	Recall	0.99	0.04	0.40	0.00	0.00	0.29	0.29	0.29
	F_1-score	0.48	0.08	0.29	0.00	0.00		0.17	0.17
	Precision	0.85	0.78	0.85	0.93	0.64		0.81	0.81
ANN classifier	Recall	0.96	0.50	0.85	0.98	0.75	0.81	0.81	0.81
	F_1-score	0.90	0.61	0.85	0.96	0.69		0.80	0.80
	Precision	0.84	0.59	0.90	0.98	0.77		0.81	0.82
KNN classifier	Recall	0.97	0.81	0.85	0.85	0.50	0.80	0.79	0.80
	F_1-score	0.90	0.68	0.88	0.91	0.60		0.79	0.80
	Precision	0.83	0.71	0.85	0.98	0.77		0.83	0.83
Decision tree classifier	Recall	0.93	0.72	0.83	0.93	0.72	0.83	0.83	0.83
	F_1-score	0.88	0.72	0.84	0.96	0.74		0.83	0.83
	Precision	0.20	0.18	0.17	0.14	0.17		0.17	0.17
Logistic regression classifier	Recall	0.04	0.39	0.15	0.00	0.33	0.18	0.18	0.18
	F_1-score	0.07	0.25	0.16	0.01	0.23		0.14	0.14
	Precision	0.00	0.18	0.00	0.36	0.18		0.14	0.15
Support vector classifier	Recall	0.00	0.48	0.00	0.02	0.44	0.18	0.19	0.18
	F_1-score	0.00	0.27	0.00	0.04	0.26		0.11	0.11
	Precision	0.96	0.69	0.84	1.00	0.75		0.85	0.85
Random forest classifier	Recall	0.94	0.69	0.87	0.98	0.75	0.85	0.85	0.85
	F_1-score	0.95	0.69	0.85	0.99	0.75		0.85	0.85

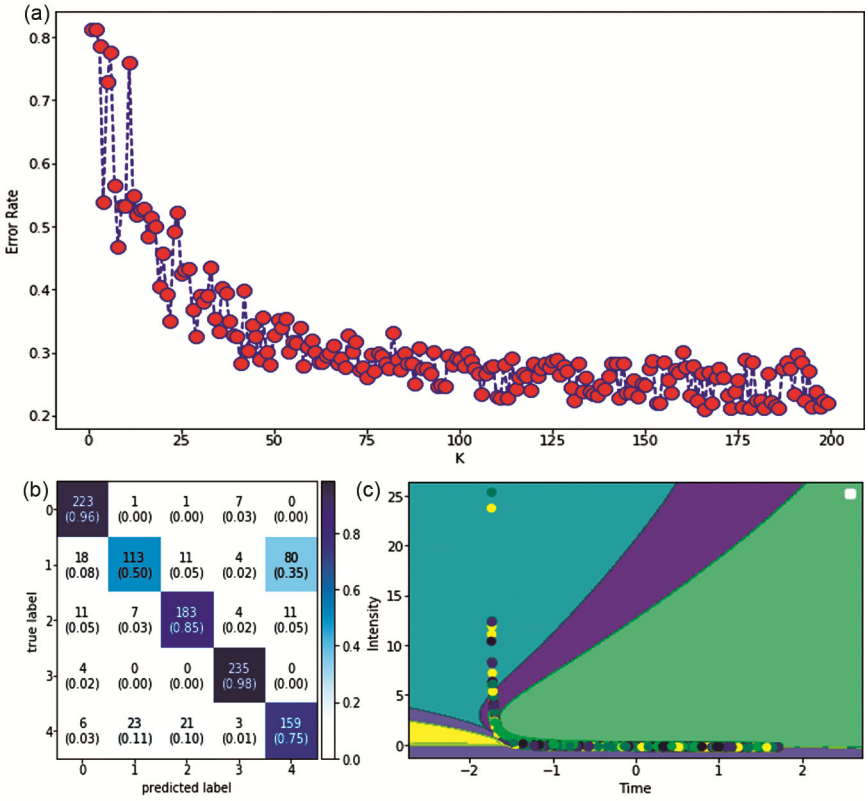

Fig. 3 — Error plot, confusion matrix and decision boundary plot obtained for ANN Classifier

amylose content in the rice sample might thus be predicted or classified more accurately using the ANN algorithm when slope-values are used for forecasting the amylose content in the rice samples. The observed results are in close agreement with results obtained for the prediction of employability using ANN.[9]

Similar studies have been carried out by various researchers[15–19] for detecting breaches in the information system of the web, categorization of coffee kinds, different kinds of fish, EEG electrical signals for seizures caused by epilepsy, and pneumonia grouping, respectively, via neural-networks with deep learning approach.

KNN Classifier

Variations in the n-neighbor values between 1 and 50 were made during the assessment process to improve the model's accuracy, and the related error values were shown Fig. 4(a). The chart shows that the n-neighbor value of 2 had the lowest error rate value, and this optimized neighbor value was chosen for the purpose of additional investigation. The results are illustrated in Fig. 4(b) as a confusion matrix, and its associated result is reported in Table 1. For the ideal estimator number, the accuracy of the KNN model was determined to be 0.80. The overall precision of KNN algorithm was computed to be 0.82 from the table. The model's average F_1 score and recall, however, were 0.80 and 0.80, accordingly. For the classes 0, 1, 2, 3, and 4, correspondingly, the accuracy values of algorithm in predicting the chosen percentage of amylose was discovered to be 0.84, 0.59, 0.90, 0.98, and 0.77. The Recall and F_1 scores were found to be 0.97, 0.81, 0.85, 0.85 and 0.50; and 0.90, 0.68, 0.88, 0.91, and 0.60, correspondingly, for classes 0, 1, 2, 3, and 4. The decision boundary was drawn with the distinct categories indicated by specific color codes, accompanied by a scatter plot of the recorded observations for a certain concentration, as shown in Fig. 4(c) to visualize the data-spreading pattern.

The aforementioned tables and figures show that although the model's accuracy with the slope data set was good, a reduced rate of error was found when the model was evaluated using slope data. Additionally, slope data allowed for a more accurate identification of the target concentration of amylose in the sample using model parameters including accuracy, recall,

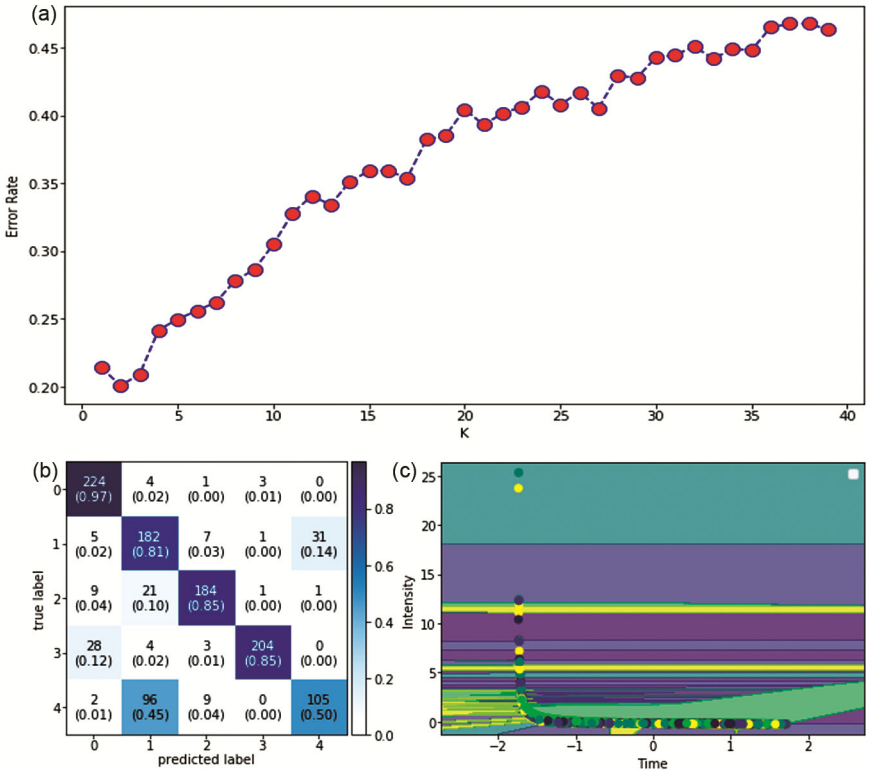

Fig. 4 — Error plot, confusion matrix and decision boundary plot obtained for KNN Classifier

and F_1 score. Given the aforementioned model parameter, it was determined that the KNN algorithm with slope data is more accurate at predicting or categorizing the amylose content in the rice sample since a higher model accuracy of 0.80 was seen in the case of slope data.

Previous research[20] found that an autonomous real-time prediction approach based on the KNN architecture and the Ecuclidean distance was effective in properly identifying the genuine groups. Similar classification work was published by several studies[21,22] demonstrated that the KNN method may also be utilized to precisely categorize text and textural items into desired classes.

Decision Tree Classifier

To categorize the level of amylose present in the sample, the decision tree classifier has been tested using the slope values acquired from the device. The tree's depth values varied from 1 to 50 and the associated error values were shown Fig. 5(a). From the graphic, it can be seen that the depth value of 14 had the lowest error rate value, and this optimized value was chosen for further investigation. The decision tree model's accuracy was reported to be 0.83 for the optimum depth parameters. The ensuing findings are shown in Fig. 5(b) as a confusion matrix, and the obtained value is shown in Table 1. From the table, it was inferred that the Decision Tree model's accuracy level was 0.83. But the model's mean recall and F_1 score were 0.83 and 0.83, accordingly. For the classes 0, 1, 2, 3, and 4, it was discovered that the model's accuracy values for detecting the specified proportion of amylose were 0.83, 0.71, 0.85, 0.98, and 0.77, respectively. Recall and F_1 values were found to be 0.93, 0.72, 0.83, 0.93, and 0.72; and 0.88, 0.72, 0.84, 0.96, and 0.74 for classes 0, 1, 2, 3, and 4, respectively. The decision boundary was drawn with the various classes indicated by distinct color codes, followed by a scatter diagram of recorded observations for a certain concentration, as shown in Fig. 5(c). in order to interpret the data distribution pattern.

It was noted that the model's reliability for the slope data set seemed excellent, and a decreased rate of errors was seen when the model was evaluated using slope data. Additionally, compared to previous machine learning algorithms employing slope data, the model parameters of accuracy, recall, and F_1 score were greater when slope data was used to determine the target concentration of amylose in the sample. The Decision Tree algorithm with slope data is more

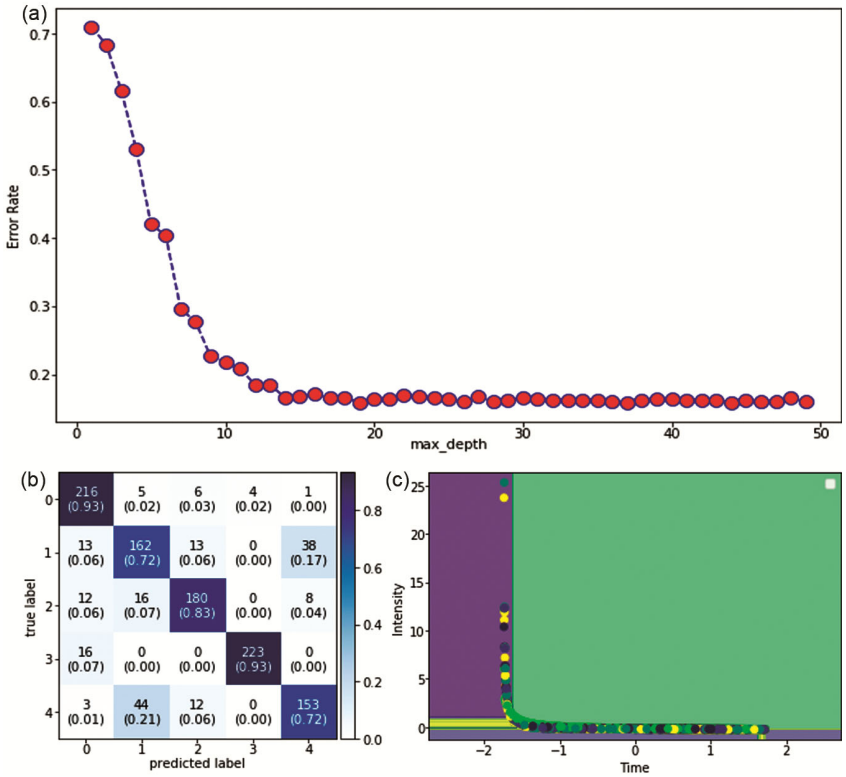

Fig. 5 — Error plot, confusion matrix and decision boundary plot obtained for Decision Tree Classifier

accurate than intensity data in predicting or classifying the amylose concentration in the rice sample within the chosen machine learning models, it was concluded. Taking into account the aforementioned model parameter, higher model accuracy of 0.83 was observed in the case of slope data.

According to earlier studies,[23–25] the decision tree technique might be effectively employed to categorize data extraction and grouping rules-based decision tree algorithms in Hierarchical-Intrusion-Detection-Systems.

Logistic Regression Classifier

To categorize the intensity of amylose in the sample, the logistic regression predictor was assessed using the slope values acquired from the device. The confusion matrix Fig. 6(a) employed to illustrate the outcomes of the logistic regression technique evaluation displays the findings, and Table 1 summarizes the values. The accuracy of the logistic regression approach was determined to be 0.18, as seen in the table. However, it was discovered that the logistic regression model's average accuracy was 0.17. The model's average F_1 score and recall were 0.14 and 0.18, accordingly. For the classes 0, 1, 2, 3, and 4, it was discovered that the model's accuracy

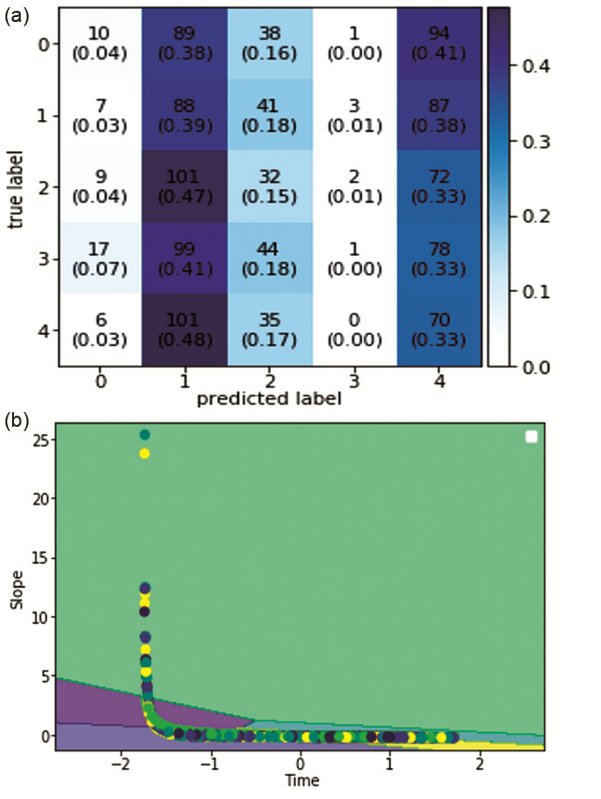

Fig. 6 — Confusion matrix and decision boundary plot obtained for Logistic Regression Classifier

values for determining the appropriate amount of amylose were 0.20, 0.18, 0.17, 0.14, and 0.17, respectively. Recall and F_1 scores were found to be 0.04, 0.39, 0.15, 0.00, and 0.33 and 0.07, 0.25, 0.16, 0.01, and 0.23 for classes 0, 1, 2, 3, and 4, respectively. The decision boundary was drawn with the various classes indicated by distinct color codes, followed by a scatter plot of the observed data points for a certain concentration, as shown in Fig. 6(b) in order to visualize the data distribution pattern.

From aforementioned table and graphs, it is marked that the model's accurateness for slope data was good, and that resulted in lower values for model when determining the appropriate amount of amylose in the samples. In light of the aforementioned model parameter, a lower model accuracy of 0.18 was observed for slope data. In order to predict or categorize the amylose concentration in the rice sample, it was determined that the logistic regression algorithm using slope data is inaccurate to predict the amylose content in the sample precisely.

The study results[26–28] for skewed information categorization, data sorting, and machine learning strategy for credit history data classification, correspondingly, indicated that logistic regression might assist in categorizing given input data sets.

Support Vector Classifier (SVC)

The Support Vector Classifier (SVC) has been checked through the sensor's slope values to improve its precision. Variations in the kernel values between 1 and 50 were used to perform evaluation, and then the related errors were shown Fig. 7(a). The figure showed that the kernel number of 40 had the minimum error rate value, and such optimized kernel value was chosen for extra investigation. The findings are displayed in Fig. 7(b) through a confusion matrix, and the obtained value was represented in Table 1. The accuracy of the SVC model was determined to be 0.18 for the optimized kernel values. It was discovered that the SVC model's average accuracy was 0.15. The model's average F_1 score and recall, however, were 0.11 and 0.18, respectively. For the classes 0, 1, 2, 3, and 4, it was discovered that the model's accuracy values for recognizing the chosen level of amylose were 0.00, 0.18, 0.00, 0.36, and 0.18, individually. Recall and F_1 scores were found to be 0.00, 0.48, 0.00, 0.02, and 0.44; and 0.00, 0.27, 0.00, 0.04, and 0.26 for classes 0, 1, 2, 3, and 4, respectively. The decision boundary was drawn with the various classes indicated by specific color codes, accompanied by a scatter plot of the observed data points for a certain concentration, as shown in Fig. 7(c) in order to visualize the data spreading array.

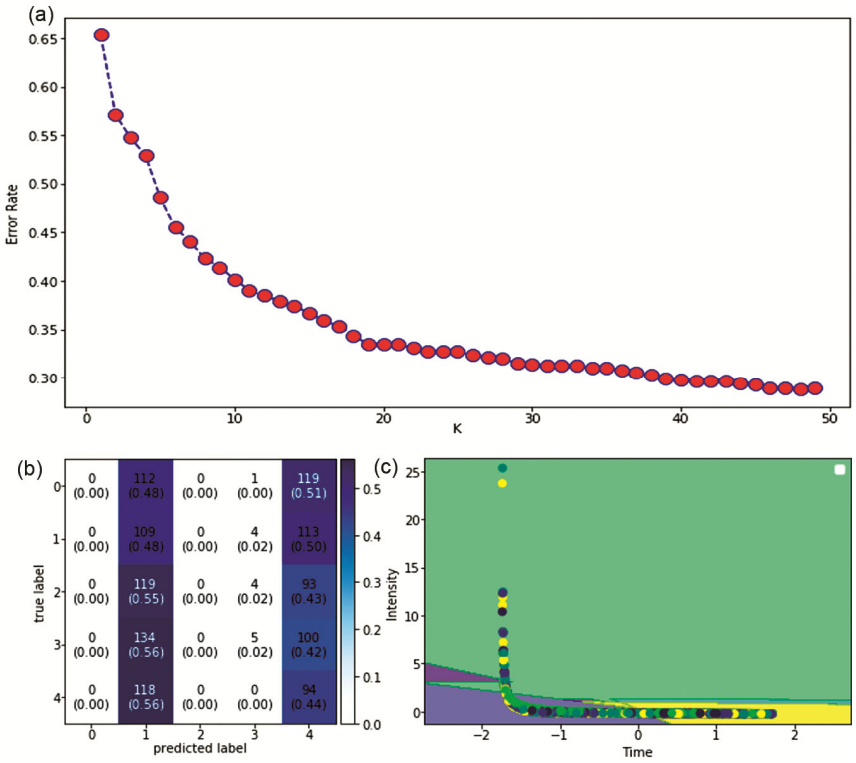

Fig. 7 — Error plot, confusion matrix and decision boundary plot obtained for Support Vector Classifier

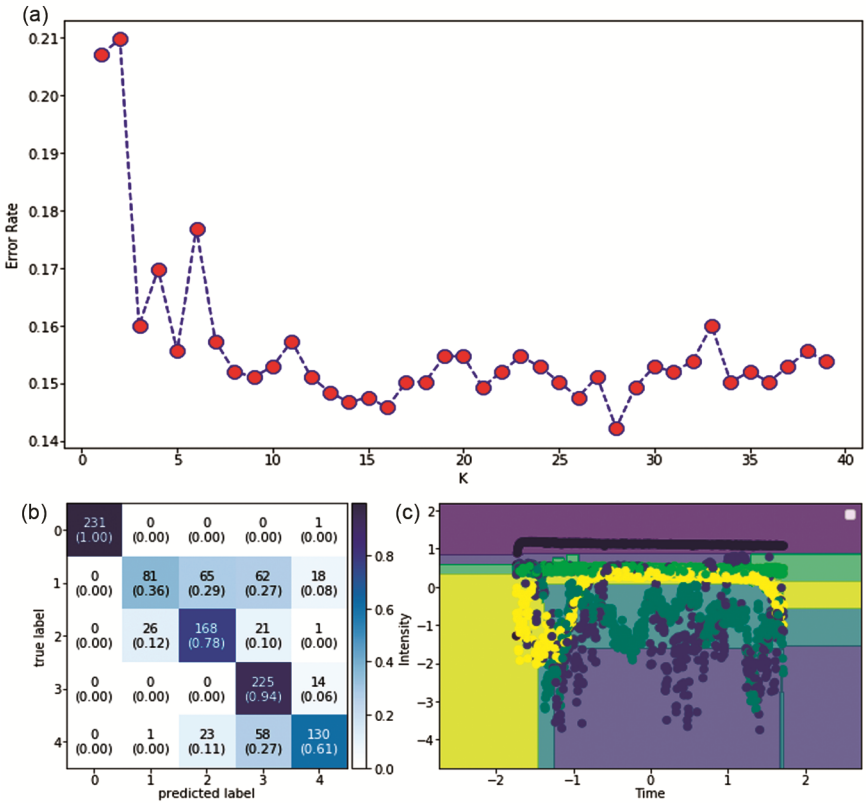

Fig. 8 — Error plot, confusion matrix and decision boundary plot obtained for Random Forest Classifier

Although the accuracy of the algorithm with the slope dataset was not acceptable, it was clear from the accompanying table and figures that a larger rate of error was there. Furthermore, using slope data to determine the target concentration of amylose in the sample resulted in reduced model parameters for accuracy, recall, and F_1 score. Given the aforementioned model parameter, slope data showed a lower model accuracy of 0.18. As a result, it was determined that the SVC algorithm using slope data is sufficiently unreliable to forecast or categorize the amylose content in the rice sample. Several researchers[29,30] did similar experiments and found that the SVC approach could be utilized effectively to categorize unbalanced data using Support Vector Machines and identified liver illness in cows.

Random Forest Classifier

By changing the maximum depth and n-estimator values from 1 to 50, the Random Forest Classifier was tested to determine which had a superior accuracy of the model, and the associated error values were shown Fig. 8(a). According to the figure, the depth and n-estimator values of 28 and 22, respectively had the lowest error rates, and these optimum values were chosen for further investigation. The findings are shown in Fig. 8(b) and the corresponding value was indicated in Table 1. The accuracy of the random forest classification algorithm was determined to be 0.85 for the optimal model parameter values. The table showed that the random forest classifier model's average accuracy was determined to be 0.85. The model's average F_1 score and recall, however, were 0.85 and 0.85, respectively. The model's accuracy was determined to be 0.96, 0.69, 0.84, 1.00, and 0.75 for the classes 0, 1, 2, 3, and 4, respectively, in identifying the chosen concentration of amylose. Recall and F_1 scores were found to be 0.94, 0.69, 0.87, 0.98, and 0.75; and 0.95, 0.69, 0.85, 0.99, and 0.75 for classes 0, 1, 2, 3, and 4. The decision boundary was drawn with the various classes indicated by specific color codes, followed by a scatter plot of the recorded observations for a certain concentration, as shown in Fig. 8(c) in order to visualize the data distribution pattern. For the slope dataset, it was noted that the model's accuracy and decreased rate error were both present. Additionally, slope data allowed for a more accurate identification of the target concentration of amylose in the sample using model parameters including accuracy,

recall, and F_1 score. It was shown that the random forest classifier method using slope data is more effective in predicting the amylose content in the rice sample by taking into account the aforementioned model parameter, which resulted in a better model accuracy of 0.85.

Relating to Table 1, it was found that the accuracy of 0.85 was noticed for the Random Forest model, preceded by accuracies of 0.83, 0.81, 0.80, 0.29, 0.18 and 0.18 for Decision-Tree, ANN, KNN, Adaboost, Logistic-Regression and Support Vector Algorithms, respectively, for the slope values acquired from the sensor. Therefore, it was determined that using the slope values received by the sensor, the Random Forest model was found to be the utmost effective in estimating amylose level in the rice sample.

As reported by various researchers[31–33], the Random Forest technique has been employed in the categorization of neural imaging information in Alzheimer's illness, large-scale data grouping in the IoTs, and web-based data classification purposes.

Conclusions

The development of a colorimetric amylose sensor resulted in a faster and simpler technique of determining the amylose concentration for the evaluation of rice ageing. The algorithms for machine learning that were selected have been put to evaluation using the slope data that the sensor had collected for correct interpretation. It was determined that the random forest model could predict amylose content more accurately and intern it was associated to rice ageing and recommended for future works. Since the accuracy of the optimized model was 0.85 for raw data received, it can be improvised by further pre–processing of data before subjecting to the modeling or ensemble modeling can be an optional. Though the other models have been tested to predict the amylose content in rice, decision tree, ANN and KNN algorithms were also performed well with on par results wherein remaining models were found inaccurate in predicting the concentration of amylose due to overlapping of data. As a result, it was advised to use a random forest approach to predict the amylose level in the rice specimen using slope data from a colorimetric amylose detector to find out how old the rice was.

Acknowledgement

The DST-Ph.D. fellowship for Science and Engineering Students was awarded by the Karnataka Science and Technology Promotion Society (KSTePS), and the contributors are thankful for their financial support in order to carry out the study efficiently. Additionally, the authors would like to express their gratitude to Dr. Mahantshivayogayya K, Scientist, AICRP on Rice, ARS, Gangavati, and University of Agricultural Sciences, Raichur, Karnataka, India for supplying specimens of several rice cultivars for the present study.

References

1 Perez C M & Juliano B O, Texture changes and storage of rice, *J Texture Stud*, **12(1)** (1981) 321–333.

2 Faruq G, Prodhan Z H & Nezhadahmadi A, Effects of ageing on selected cooking quality parameters of rice, *Int J Food Prop*, **18(4)** (2015) 922–933, doi: https://doi.org/10.1080/10942912.2014.913062.

3 Zhou Z, Robards K, Helliwell S & Blanchard C, Ageing of stored rice: Changes in chemical and physical attributes, *J Cereal Sci*, **35(1)** (2001) 65–78, doi: https://doi.org/10.1006/jcrs.2001.0418.

4 Devraj L, Natarajan V, Ramachandran S V, Manicakam L & Saravanan S, Accelerated aging by microwave heating and methods to distinguish aging of rice, *J Food Process Eng*, **43(6)** (2020) 13405–13415, doi: https://doi.org/10.1111/jfpe.13405.

5 Popa A, Hnatiuc M, Paun M, Geman O, Hemanth D J, Dorcea D, Son L H & Ghita S, An intelligent IoT-based food quality monitoring approach using low-cost sensors, *Symmetry*, **11(3)** (2019) 374–391, doi: https://doi.org/10.3390/sym11030374.

6 Moradi M, Balanian H, Taherian A & Mousavi Khaneghah A, Physical and mechanical properties of three varieties of cucumber: A mathematical modeling, *J Food Process Eng*, **43(2)** (2020) 13323–13330, doi: https://doi.org/10.1111/jfpe.13323.

7 Osisanwo F Y, Akinsola J E T, Awodele O, Hinmikaiye J O, Olakanmi O & Akinjobi J, Supervised machine learning algorithms: classification and comparison, *Int J Comput Trends Technol*, **48(3)** (2017) 128–138.

8 Deshpande S, Nidoni U, Hiregoudar S, Ramappa K T, Maski D & Naik N, Performance of advanced machine learning models in the prediction of amylose content in rice using internet of things-based colorimetric sensor, *Curr Sci* (00113891), **124(6)** (2023) 722–730, doi: 10.18520/cs/v124/i6/722-730.

9 Celine S, Maria D M & Savitha D M, Logistic regression for employability prediction, *Int J Innov Technol Explor Eng*, **9(3)** (2020) 2471–2478.

10 Anonymous (2018a), AdaBoost Classifier in Python, https://www.datacamp.com/ community/tutorials/adaboostclassifier-python, Accessed on 01-05-2021.

11 Anonymous (2019), Understanding Logistic Regression in Python, https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python, Accessed on 01-05-2021.

12 Anonymous (2018b). Understanding Random Forests Classifiers in Python, https://www.datacamp.com/community/tutorials/random-forests-classifier-python, Accessed on 01-05-2021.

13 Zheng J, Lin D, Gao Z, Wang S, He M & Fan J, Deep learning assisted efficient AdaBoost algorithm for breast cancer

detection and early diagnosis, *IEEE Access*, **8** (2020) 96946–96954, doi: https://doi.org/10.1109/ACCESS.2020.2993536.

14 Ji X, Yang B & Tang Q, Acoustic seabed classification based on multi beam echo sounder backscatter data using the PSO-BP-AdaBoost algorithm: A case study from Jiaozhou Bay, China, *IEEE J Ocean Eng*, **46(2)** (2020) 509–519, doi: https://doi.org/10.1109/JOE.2020.2989853.

15 Kim J, Shin N, Jo S Y & Kim S H, Method of intrusion detection using deep neural network, *IEEE Int Conf Big Data & Smart Comput (Big Comp)*, (2017) 313–316, doi: https://doi.org/10.1109/BIGCOMP.2017.7881684.

16 Arboleda E R, Fajardo A C & Medina R P, Classification of coffee bean species using image processing, artificial neural network and K nearest neighbors, *IEEE Int Conf Inno Res & Dev (ICIRD)*, (2018) 1–5, doi: https://doi.org/10.1109/ICIRD.2018.8376326.

17 Siddiqui S A, Salman A, Malik M I, Shafait F, Mian A, Shortis M R & 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*, **75(1)** (2018) 374–389, doi: https://doi.org/10.1093/icesjms/fsx109.

18 Narang A, Batra B, Ahuja A, Yadav J & Pachauri N, Classification of EEG signals for epileptic seizures using Levenberg-Marquardt algorithm based Multilayer Perceptron Neural Network, *J Intell Fuzzy Syst*, **34(3)** (2018) 1669–1677, doi: 10.3233/JIFS-169460.

19 Stephen O, Sain M, Maduh U J & Jeong D U, An efficient deep learning approach to pneumonia classification in healthcare, *J Healthc Eng*, **2019** (2019), doi: https://doi.org/10.1155/2019/4180949.

20 Adeniyi D A, Wei Z & Yongquan Y, Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method, *Appl Comput Inform*, **12(1)** (2016) 90–108.

21 Shah K, Patel H, Sanghvi D & Shah M, A comparative analysis of logistic regression, random forest and KNN models for the text classification, *Augment Hum Res*, **5(1)**, (2020) 1–16, doi: https://doi.org/10.1007/s41133-020-00032-0.

22 Moldagulova A & Sulaiman R B, Using KNN algorithm for classification of textual documents, *8th Int Conf Inf Tech (ICIT)*, (2017) 665–671, doi: https://doi.org/10.1109/ICITECH.2017.8079924.

23 Gupta B, Rawat A, Jain A, Arora A & Dhami N, Analysis of various decision tree algorithms for classification in data mining, *Int J Comput Appl*, **163(8)** (2017) 15–19.

24 Ahmim A, Maglaras L, Ferrag M A, Derdour M & Janicke H, A novel hierarchical intrusion detection system based on decision tree and rules-based models, *15th Int Conf Distribut Comput Sensor Syst (DCOSS)*, (2019) 228–233, doi: https://doi.org/10.1109/DCOSS.2019.00059.

25 Abdallah I, Dertimanis V, Mylonas H, Tatsis K, Chatzi E, Dervili N, Worden K & Maguire E, 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) 3053–3061, doi: https://doi.org/10.3929/ethz-b-000313962.

26 Ohsaki M, Wang P, Matsuda K, Katagiri S, Watanabe H & Ralescu A, Confusion-matrix-based kernel logistic regression for imbalanced data classification, *IEEE Trans Knowl Data Eng*, **29(9)** (2017) 1806–1819, doi: https://doi.org/10.1109/TKDE.2017.2682249.

27 De Menezes F S, Liska G R, Cirillo M A & Vivanco M J, Data classification with binary response through the boosting algorithm and logistic regression, *Expert Syst Appl*, **69** (2017) 62–73, doi: https://doi.org/10.1016/j.eswa.2016.08.014.

28 Dumitrescu E, Hue S, Hurlin C & Tokpavi S, Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects, *Eur J Oper Res*, **297(3)** (2022) 1178–1192, doi: https://doi.org/10.1016/j.ejor.2021.06.053.

29 Mathew J, Pang C K, Luo M & Leong W H, Classification of imbalanced data by oversampling in kernel space of support vector machines, *IEEE Trans Neural Netw Learn Syst*, **29(9)** (2017) 4065–4076, doi: https://doi.org/10.1109/TNNLS.2017.2751612.

30 Devikanniga D, Ramu A & Haldorai A, Efficient diagnosis of liver disease using support vector machine optimized with crows search algorithm, *EAI Endorsed Trans Energy Web*, **7(29)** (2020) 1–10, doi: https://doi.org/10.4108/eai.13-7-2018.164177.

31 Sarica A, Cerasa A & Quattrone A, Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: A systematic review, *Front Aging Neurosci*, **9** (2017) 329–340, doi: https://doi.org/10.3389/fnagi.2017.00329.

32 Lakshmanaprabu S K, Shankar K, Ilayaraja M, Nasir A W, Vijayakumar V & Chilamkurti N, Random forest for big data classification in the internet of things using optimal features, *Int J Mach Learn Cybern*, **10(10)** (2019) 2609–2618, doi: https://doi.org/10.1007/s13042-018-00916-z.

33 Demidova L A, Klyueva I A & Pylkin A N, Hybrid approach to improving the results of the SVM classification using the Random Forest algorithm, *Procedia Comput Sci*, **150** (2019) 455–461, doi: https://doi.org/10.1016/j.procs.2019.02.077.