Original Article

# RICE VARIETIES CLASSIFICATION USING MACHINE LEARNING ALGORITHMS

**Pranshu Saxena[1], Kanu Priya[2], Sachin Goel[3], Puneet Kumar Aggarwal[4], Amit Sinha[5], Parita Jain[6]**

[1]Department of Information Technology, ABES Engineering College, Ghaziabad, India.
pranshusaxena@gmail.com
[2]National Institute of Fashion Technology (NIFT).
kanu.priya@nift.ac.in
[3]Department of Information Technology, ABES Engineering College, Ghaziabad, India.
s.goel@abes.ac.in
[4]Department of Information Technology, ABES Engineering College, Ghaziabad, India.
puneetaggarwal7@gmail.com
[5]Department of Information Technology, ABES Engineering College, Ghaziabad, India.
amit.sinha@abes.ac.in
[6]Department of Computer Science and Engineering, KIET group of institutions, Ghaziabad, Delhi-NCR.
paritajain23@gmail.com

**DOI:** 10.47750/pnr.2022.13.S07.479

## Abstract

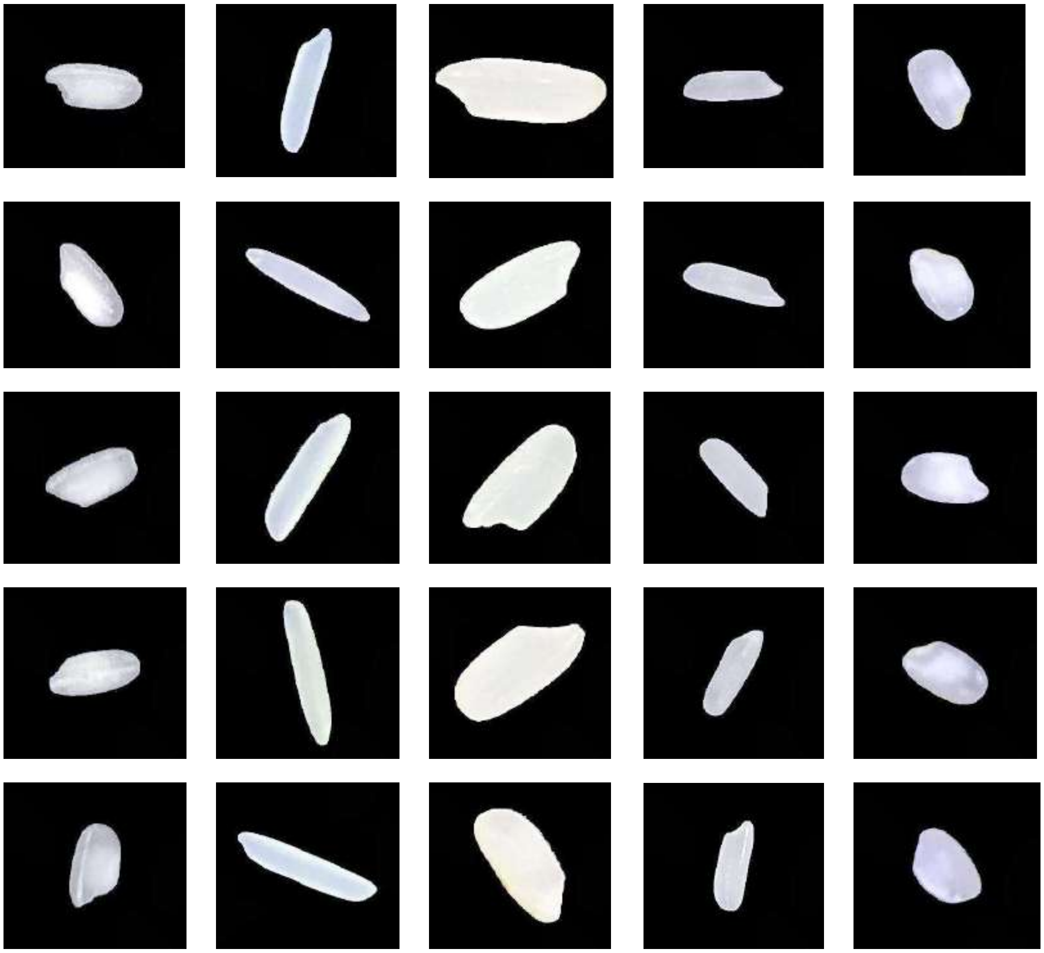
Rice is one of the most important crops in the world. About one-half of the world's population is wholly dependent upon rice as food. The rice plant height is about 1.2 meters and is an annual grass. In this paper, the data set contain 5 variety of rice that is growing all over the world. The dataset contains a total of 75000 samples, of which 15000 are from each class. The data set contains 107 features from which the best 20 features are selected using Random Forest Classifier. Performance metrics such as accuracy, precision, recall, and f1 score have been compared with and without the feature selection method. Most popular machine learning algorithms, namely logistic regression, decision tree, support vector machine classifier, random forest classifier, perceptron, K-nearest neighbors' classifier, and Gaussian naïve Bayes classifier, have been trained on 70% training - 30% testing data and 80% training - 20% testing data. Experimental results show very promising results. In random forest classification, accuracy is 99.85%, while the decision tree classifies the rice sample with 99.68% accuracy.

**Keywords:** Machine Learning, Classifiers, Rice Production classification, Feature selection, Preprocessing.

## Introduction

Rice is one of the most important food crops worldwide. Rice can grow on various soils with PH ranging from 5.0 to 9.5 (free of water logging). The temperature range 16-20 °C is good for rice production. The rainfall range 100-200 cm is also good for rice production. The historical way to cultivate the rice is by flooding the field while setting the young seeding. Applications of machine learning in agriculture are interesting because they offer non-destructive evaluation and are less expensive than manual procedures. When compared to manual procedures, machine learning technologies have advantages. Manual evaluation or classification of grains can be time-consuming and expensive because the human element is so important. The evaluation procedure may be different when using manual methods because it is only as good as the evaluation specialists' experience. Additionally, when an assessment is done on a big scale, quick decision-making through manual methods may be challenging. Fig. 1 explaining the 5 different varieties of rice is taken is this research.

Fig. 1: Rice Varieties used in this research (a) Arborio (b) Basmati (c) ipsala (d) Jasmine (e) karacadag



(a) (b) (c) (d) (e)

Five different varieties of rice that are grown worldwide are included in the data set for this article. A total of 75000 samples—15,000 from each class—make up the dataset. Detailed discussion of database, which is having 12 morphological features, 4 shape-based features, and 90 color-based feature available open source and discussed in details in [21]. This data set contains 22 missing features values, so before actual classification, firstly data is preprocessed and scaled them up to a limit [0 1]. Later on, the Random Forest (RF) Classifier is used to choose the top 20 features from a total of 107 features in the data set. Accuracy, precision, recall, and f1 score performance metrics have been compared with and without the feature selection method. Logistic Regression (LR), Decision Trees (DT), Support Vector Classifiers (SVC), random forest classifiers, perceptrons, K-nearest neighbours classifiers (k-NN), and Gaussian Naive Bayes Classifiers (NB) are used to classify rice variety into respective class. Performance of all the classifiers is compared with previous research discussed in the literature.

A machine learning model can be divided into three parts regression, classification, and clustering. Regression and classification are supervised learning techniques, and clustering is an unsupervised learning technique. The regression model can predict a continuous value, while the classification model will be used to predict a discrete value. Here in this paper, the classification of rice varieties is presented with different classification models. The rice classification dataset contains 75000 images classified into 5 classes using artificial neural networks [1]. In this paper, the author measures the performance of different feature sets. The author has achieved approximately 99% accuracy for 5-class classification. The study by Sun et al. had 1700 rice data containing 2 varieties of rice, classified using a support vector machine algorithm, and achieved the success of 98.5%. They found that image processing technology is effective for 2 to five connected kernels [2]. Lin et al., in their study, had 7399 rice data contain have 3 varieties of rice. The work is carried out by a convolutional neural network and achieved an accuracy of 95.5%. Their goal is to create an automatic intelligent detection system [3].

The study was carried out by Liu et al. on 200 pieces of data, and the data set contained three varieties of rice; the algorithm used on this is a convolutional neural network and achieved the success of 88.07%. This work highlights an improved hidden layer by optimizing with the help of particle swarm optimization (PSO) [4]. Dubey et al. use of artificial neural network having 45 morphological features. In this study, they show an increase in the number of feature result in an increase in success rate. They achieved a success of 88% [5]. An independent study by Ou. Yang et al. on the data in which there are five different rice seeds. This classification is carried out by an image processing technique, backpropagation classification (BP-ANN), and achieved an accuracy of 86.85% [6]. Abirami et al., in their study, use the neural network pattern recognition technique. The data set contains only one variety of rice: Basmati rice, and by the neural network pattern recognition technique, the accuracy is 98.7% [7]. The study by Sethy & Chatterjee has a data set containing six rice varieties. The study is to determine rice varieties' geometric and texture features. They use the Muti class support vector machine (M-SVM) algorithm; by this, they achieved the success of 92% [8].

Chen et al., in their study, support vector machine for classification in used and achieved the accuracy of 99.3% they use morphological features on an image of rad Indica rice. They develop a machine vision system to study calcareous defects in rice grain [9]. Koklu & Ozkan perform image classification of dry bean varieties. Classification is done by KNN, SVM, decision tree (DT), etc. SVM achieves the best accuracy is 93.13%. Work is done using the shape and morphological features of 7 varieties of dry beans [10, 22]. The study by Baykan et al. used 9 morphological features of the grains they get the grey level average. The data set used for this study contains 5 varieties of grains and achieved the success of 82.65% [11].

Babalik et al. classify wheat based on their study's 9 geometric and 3 color properties. The data set was used to contain the 5 varieties of wheat and used a multiclass support vector machine. They achieve the success of 92.02% [12]. In the study of Zapotoczny et al., 5 different barely species are classified using the image classification technique. They used 74 morphological features. They used linear discriminant analysis as well as nonlinear discriminant analysis [13]. In their study, Kaur and Singh segregate the rice by which their features are extracted. They use multiclass support vector machines and achieve the success of 86% [14]. Pazoki et al. classify the 5 different rice varieties with the help of 4 shape factors, 24 color factors, and 11 morphological factor features. Their study used an artificial neural network and achieved success of 99.46% [15].

This paper is organized as follows, In the second section of the paper, dataset is described followed by proposed methodology is discussed. In the third section, the experimental results are obtained and experimental results are discussed. In the last section comparative analysis with previous literature and conclusion is placed.

## Material and Methodology

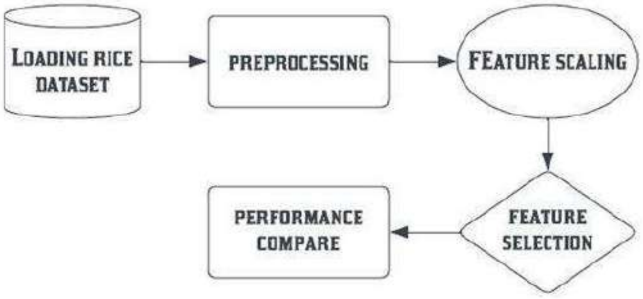
### Dataset description

The study analyzed data sets from five rice varieties that are often grown in Turkey: Arborio, Basmati, Ipsala, Jasmine, and Karacadag. 75,000 images of rice grains, 15,000 from each kind, make up the first image dataset (fig. 1). The size of the image containing each grain of rice in the RGB images in this dataset is $250 \times 250$ pixels. Additionally, a second feature dataset with a total of 106 features was created using these images to extract 12 morphological, 4 shape, and 90 colour features from each rice grain [21]. The rice types employed in the investigation and the result are tested on various classifiers.

### Methodology

Proposed research work in this paper includes various steps such as loading of data, preprocessing of the data, feature scaling using standard scaling method, feature selection using random forest classifier, and finally, comparing the performance and execution time of various machines learning algorithms. Fig. 2 shows the detailed methodology of this paper.

Fig. 2. Proposed Methodology



**Loading the dataset**

To perform classification, a rice dataset is used. It contains a total of 75000 samples having 106 features divided into 5 classes. Table 1 shows the data set characteristics used in the paper.

Table 1. Class distribution of Rice data

| Class Name | Samples | Morphological features | Shape features | Color features | Total |
|---|---|---|---|---|---|
| Basmati | 15000 | 12 | 4 | 90 | 106 |
| Arborio | 15000 | 12 | 4 | 90 | 106 |
| Jasmine | 15000 | 12 | 4 | 90 | 106 |
| Ipsala | 15000 | 12 | 4 | 90 | 106 |
| Karacadag | 15000 | 12 | 4 | 90 | 106 |
| Total | 75000 | 12 | 4 | 90 | 106 |

**Preprocessing**

This paper's proposed research has a data set containing 22 missing values. The missing values are replaced by the mean of that feature (in which missing values are present). Table 2 shows the features that contain the missing values, the total number of missing values in that feature, and the mean of the respective features. The next step in preprocessing in this research work is label encoding. Label encoding is the technique in which categorical data is converted into numerical data. This paper converts the rice varieties Arborio, Basmati, Ipsala, Jasmine, Karacadag into 0, 1, 2, 3, 4 respectively.

Table 2. Inputted Method and Value

| Feature name | Missing values counts for each feature | Replaced by the mean value of each feature |
|---|---|---|

| | | |
|---|---|---|
| skewB | 6 | 0.52 |
| kurtosisB | 6 | 4.73 |
| skewCb | 3 | -0.47 |
| skewCr | 2 | 1.73 |
| kurtosisCb | 3 | 5.12 |
| kurtosisCr | 2 | 67.69 |

**Feature Scaling**

The data used in this research paper contain some features with high-frequency values, and other feature values are in the range $0\ to\ 1$. By using feature scaling, all features' values are normalized to a certain range. We use standard scaling imported from sklearn library [16] to accomplish this task. The formula used for standard scaling is:

$$X_{std} = \frac{x - \mu}{\sigma} \qquad (1)$$

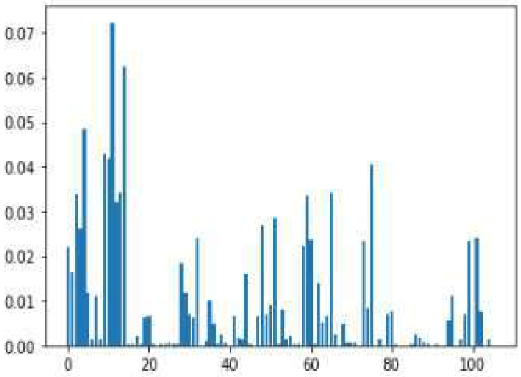
In standard scaling, there is no certain range. This formula $x$ is the data set, $\mu$ is the mean value, and $\sigma$ is the standard deviation.

**Feature Selection**

As our research paper contains a large number of features (106), due to this time for training, the model takes more and more memory space. We select the 20 best features using the Random Forest to reduce the execution time for training the model. Random forest calculates the importance (fig. 3) of each feature by increasing the pureness of the leaves (the importance of each feature is how pure each bucket is).

By reducing the number of features to 20, the time of execution for model training reduces drastically. The execution time for 106 features by support vector machine is approximately $367\ sec$, and for 20 features, it is approximately $250\ sec$.

Fig. 3. Important features (random forest classifier)



**Machine Learning Classifiers**

A machine learning model can be divided into three parts regression, classification, and clustering. Regression and classification are supervised learning techniques, and clustering is an unsupervised learning technique. The regression model can predict a continuous value, while the classification model will be used to predict a discrete value. Here in this paper, the classification of rice varieties classification model has been used. Classification is the grouping of objects into preset categories. This research paper uses various classification algorithms such as logistic regression, support vector machine, Gaussian naïve Bayes, random forest, and decision Tree.

The proposed research work in this paper checks accuracy, precision, f1 score, and recall for both 80% training data – 20% testing data and for 70% training data – 30% testing data. The model's accuracy can be calculated as [17,18] in equations 2 and 3.

$$Accuracy = \frac{No. of\ correct\ samples\ classified}{Total\ number\ of\ samples} \qquad (2)$$

OR

$$Accuracy\ = \frac{TP + TN}{Total\ number\ of\ samples} \qquad (3)$$

TP: True positive (samples in positive class and classified as positive).

TN: True negative (samples in negative class and classified as negative).

Precision and recall are the base of the F1 score.

Precision is the correct prediction ratio out of total predictions [19].

$$Precision = \frac{TP}{(TP + TN)} \qquad (4)$$

$$Recall = \frac{TP}{(TP - FN)} \qquad (5)$$

FN: False Negative (samples in negative class but classified as positive). Recall used an individual machine learning metric.

F1 score is the improvement of two simpler performance metrics. It is used where we have imbalanced data. The F1 score looks at the number of prediction errors the model makes and the type of errors made [20][23].

The F1 Score formula is:

$$F1_{Score} = 2 * \frac{Precision * Recall}{Precision + Recall} \qquad (6)$$

## Results and Discussion

This research is implemented using python language in a personal computer with the following specification 4GB RAM and i5-3210M CPU @ 2.50GHz. In this research paper, we compare the scores of 80% training data – 20% testing data (case 1) with 70% training data – 30% testing data (case 2) for both category 106 features and 20 features. The author finds in comparison of 106 features that the accuracy by Logistic regression is approximately the same for both the cases. The accuracy of the support vector classifier in 80% of training data is 90%, and in 70% of the training, data is 91%. (fig. 5) The accuracy of the k-nearest classifier is approximately the same in both cases. The accuracy by perceptron is around 58% in 80% of training data and 65% in 70% of training data. The accuracy in a random forest classifier is approximately the same in both cases. The accuracy by decision tree and Gaussian naïve Bayes are approximately the same in both cases. By analysis, the author finds that the random forest classifier and decision tree have the best accuracy score in both cases; the score by the random forest classifier is 99.85%, and the score by the decision tree is 99.68%. All the details and results are shown in table 3.

Fig. 4: Time comparison of each algorithm

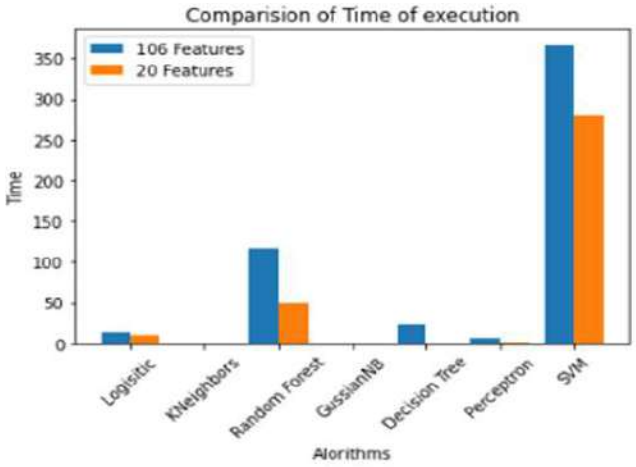


Table 3: Score Comparison of 80-20 and 70-30 for 106 features.

| **Scores of 80/20 for 106 features** | | | | | **Scores of 70/30 for 106 features** | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Sr. No. | **Model** | **Accuracy** | **F1 Score** | **Precision** | **Recall** | **Accuracy** | **F1 Score** | **Precision** | **Recall** |
| **1** | **Logistic Regression** | 77.43 | 76.87 | 77.67 | 77.53 | 77.49 | 76.89 | 77.60 | 77.55 |
| **2** | **SVC** | 90.87 | 90.79 | 90.85 | 90.85 | 91.07 | 91.02 | 91.06 | 91.09 |
| **3** | **K-Neighbors Classifier** | 92.27 | 92.20 | 92.22 | 92.25 | 92.56 | 92.52 | 92.52 | 92.58 |
| 4 | **Perceptron** | 58.19 | 53.16 | 72.09 | 58.01 | 65.13 | 60.63 | 81.11 | 65.10 |
| 5 | **Random Forest Classifier** | 99.88 | 99.88 | 99.88 | 99.88 | 99.85 | 99.85 | 99.85 | 99.85 |
| 6 | **Decision Tree Classifier** | 99.68 | 99.68 | 99.68 | 99.68 | 99.55 | 99.55 | 99.55 | 99.55 |
| 7 | **Gaussian NB** | 76.19 | 76.12 | 78.55 | 76.25 | 76.67 | 76.63 | 78.77 | 76.68 |

Below performance graph of various classifiers are placed for both data sets (80/20 & 70/30). The fig. signifies that RF classifier and DT classifier are superior in terms of all the discuss classifier in this paper on Rice dataset.

Fig. 5. Accuracy analysis of different classifiers used on 106 features (a) 80-20 training/testing ratio (b) 70-30 training/testing ratio

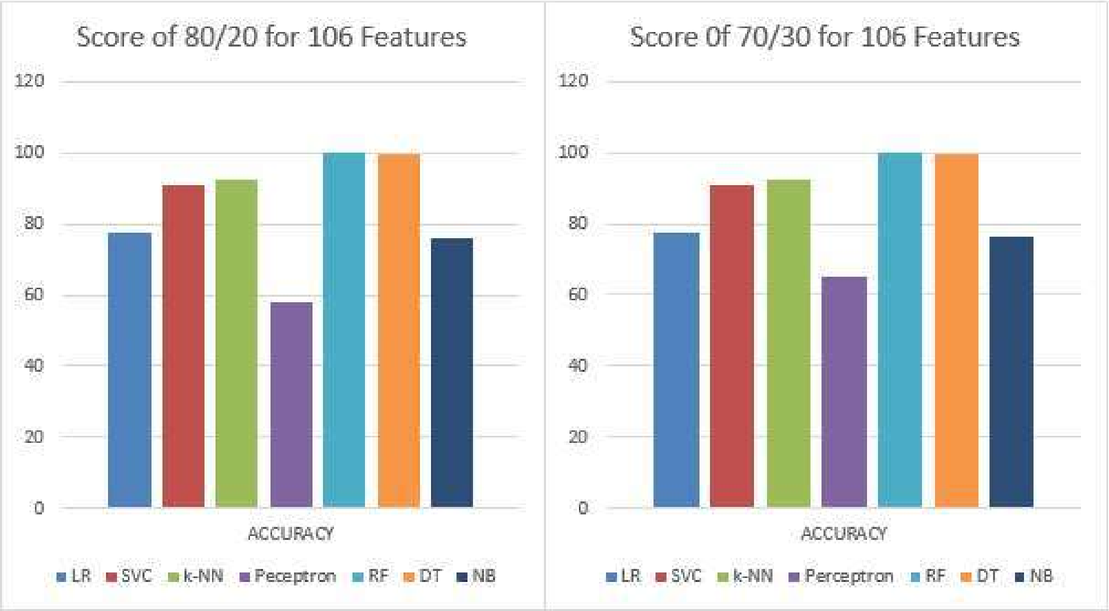
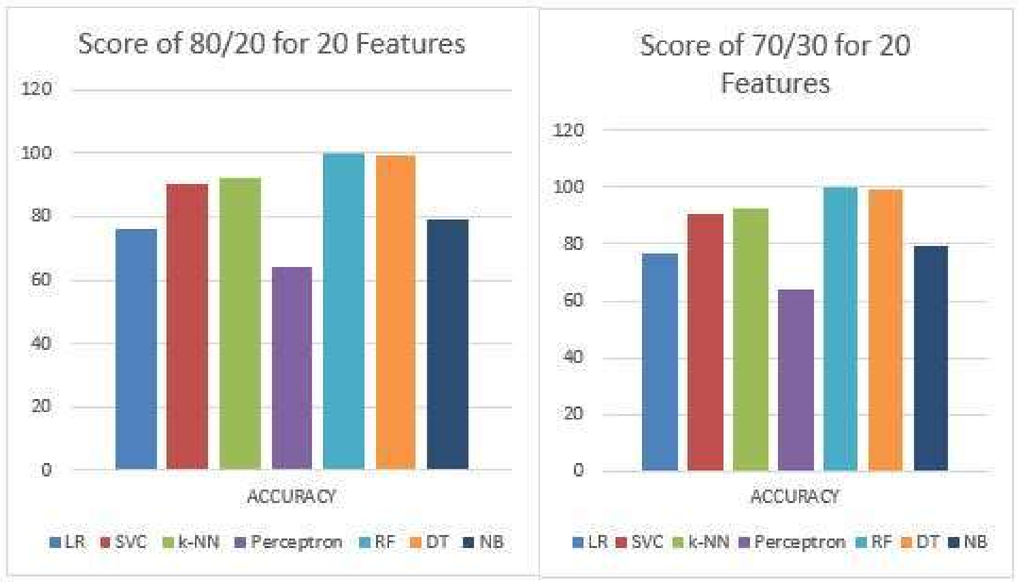


Table 4: Score Comparison of 80-20 and 70-30 for 20 features

| Scores of 80/20 for 20 features | | | | | | Scores of 70/30 for 20 features | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Sr. No.** | **Model** | **Accuracy** | **F1 Score** | **Precision** | **Recall** | **Accuracy** | **F1 Score** | **Precision** | **Recall** |
| 1 | **Logistic Regression** | 76.4 | 75.50 | 76.23 | 76.35 | 76.29 | 75.39 | 76.08 | 76.10 |
| 2 | **SVC** | 90.5 | 90.42 | 90.47 | 90.49 | 89.93 | 89.82 | 89.84 | 89.87 |
| 3 | **K-Neighbors Classifier** | 92.4 | 92.33 | 92.34 | 92.38 | 92.11 | 92.00 | 92.01 | 92.07 |
| 4 | **Perceptron** | 64.41 | 62.65 | 72.86 | 64.17 | 67.94 | 64.32 | 78.18 | 67.59 |
| 5 | **Random Forest Classifier** | 99.70 | 99.70 | 99.70 | 99.70 | 99.77 | 99.77 | 99.77 | 99.77 |
| 6 | **Decision Tree Classifier** | 99.42 | 99.42 | 99.42 | 99.42 | 99.42 | 99.42 | 99.42 | 99.42 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 7 | **Gaussian NB** | 79.31 | 79.40 | 81.55 | 79.33 | 79.02 | 79.00 | 81.14 | 78.88 |

Fig. 6. Accuracy analysis of different classifiers used on 20 features (a) 80-20 training/testing ratio (b) 70-30 training/testing ratio



The accuracy of decision tree and random forest classification is best in both the cases, and from fig. 6, it can conclude that the time of training the model for random forest classifier and decision tree is less than the time taken by the support vector machine (fig. 4) so, we conclude that random forest classifier and decision tree classifier are best in all aspects.

## Conclusion

Finally, classification results using the LR, SVC, k-NN, Perceptron, RF, DT, and NB algorithms are shown. 75,000 images of rice grains were utilized. Selected features for the data set that was inputted to different classifiers. Rice varieties such as Arborio, Basmati, Ipsala, Jasmine, and Karacadag were provided as categorization outputs. From table, 3&4 the author can conclude that the random forest classifier and decision tree classifier are performing best in both cases having 106 features and 20 features. The accuracy of the random forest classifier and decision tree classifier is more than 99%, and the time is taken to train the model in a random forest classifier and decision tree classifier is much less than the support vector machine. Table 5 shows comparative results showing the nobility of proposed algorithms.

Table 5: Performance analysis of this study and studies placed in literature on Rice data set.

| **References** | **Data Pieces** | **Class** | **Classifiers** | **Accuracy** |
|---|---|---|---|---|
| [1] | 75,000 | 5 | ANN | 99.87% |
| [2] | 17,00 | 2 | SVM | 98.5% |

| | | | | |
|---|---|---|---|---|
| [3] | 7399 | 3 | CNN | 95.5% |
| [4] | 200 | 3 | CNN | 88.07% |
| [6] | 750 | 5 | BP-ANN | 86.85% |
| [8] | 300 | 6 | Multiclass-SVM | 92% |
| [21] | 3810 | 3 | DCNN | 95.50% |
| Proposed Algorithm | 75,000 | 5 | RF | 99.85% |
| | | | DT | 99.68% |

The literature is full of studies involving rice. Table 5 provides a comparison of earlier studies that used rice and similar investigations. The algorithms utilized in this investigation had the highest classification success when compared to the studies using rice in Table 5. However, it should be emphasized that each data set in Table 5 is distinct from the others and that each data set contains a different number of rice-specific characteristics. This table is being supplied solely for informational purposes.

## References

[1] M. Koklu, I. Cinar, and Y.S. Taspinar, "Classification of rice varieties with deep learning methods", Computers and electronics in agriculture, vol. 187, 2021.

[2] C. Sun, T. Liu, C. Ji, M. Jiang, T. Tian, D. Guo, L. Wang, Y. Chen, and X. Liang, "Evaluation and analysis the chalkiness of connected rice kernels based on image processing technology and support vector machine", J. Cereal Sci. vol. 60, no.2, pp. 426–432, 2014.

[3] P. Lin, X. L. Li, Y. M. Chen, and Y. He, "A deep convolutional neural network architecture for boosting image discrimination accuracy of rice species". Food Bioprocess Technol. Vol. 11 no.4, pp/ 765-773, 2018.

[4] Ahmed, T., Rahman, C. R., Abid, M., and Mahmud, F., "Rice grain disease identification using dual phase convolutional neural network-based system aimed at small dataset", com. Vis. And patt. Recong. 2021.

[5] Dubey, B. P., Bhagwat, S. G., Shouche, S. P., and Sainis, J. K. "Potential of artificial neural networks in varietal identification using morphometry of wheat grains", Biosystems engineering, vol. 95, no. 1, pp.61-67, 2006.

[6] OuYang, A. G., Gao, R. J., Sun, X. D., Pan, Y. Y., and Dong, X. L., "An automatic method for identifying different variety of rice seeds using machine vision technology", In 2010 Sixth International Conference on Natural Computation, pp. 84-88, 2010.

[7] Neelamegam, P., Abirami, S., Priya, K. V., and Valantina, S. R., "Analysis of Rice Granules using Image Processing and Neural Network Pattern Recognition Tool", In 2013 IEEE conference on information & communication technologies, pp. 879-884, 2013.

[8] Sethy, P. K., and Chatterjee, A., "Rice variety identification of western Odisha based on geometrical and texture feature", Int. Jour. of App. Eng. Res., vol.13, no. 4, pp. 35-39, 2018.

[9] Chen, Yung-Sheng, and Wen-Hsing Hsu, "Colored rice quality inspection system using machine vision", Journal of Cereal Science vol. 88, pp. 87-95, 2019.

[10] Koklu, M., and Ozkan, I. A., "Multiclass classification of dry beans using computer vision and machine learning techniques", Computers Electronics in Agriculture vol. 174, 2020.

[11] Baykan, O., A. Babalik, and F. Botsalı, "Recognition of wheat species using artificial neural network". 4. "International Symposium On Advanced Technologies, Konya", vol. 7, no. 3, pp. 28-30, 2019.

[12] Babalik, A., et al., "Effects of Feature Selection Using Binary Particle Swarm Optimization on Wheat Variety Classification", in International Conference on Advances in Information Technology, 2010, pp. 11-17, 2010.

[13] Zapotoczny, P., M. Zielinska, and Z. Nita, "Application of image analysis for the varietal classification of barley: Morphological features", Journal of Cereal Science, vol. 48, no.1, pp. 104-110, 2008.

[14] Kaur, H. and B. Singh, "Classification and grading rice using multiclass SVM", International Journal of Scientific and Research Publications, vol.3, no.4, p. 1-5, 2013.

[15] Pazoki, A., F. Farokhi, and Z. Pazoki, "Classification of rice grain varieties using two Artificial Neural Networks (MLP and NeuroFuzzy", The Journal of Animal & Plant Sciences, vol.24, no.1, pp. 336-343, 2014.

[16] S. K. Singh, A. Sinha and S. Yadav, "Performance Analysis of Machine Learning Algorithms for Erythemato-Squamous Diseases Classification," 2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics, pp. 1-6, 2022.

[17] P. Saxena, S. K. Singh, G. Tiwary, Y. Mittal and I. Jain, "An Artificial Intelligence Technique for Covid-19 Detection with eXplainability using Lungs X-Ray Images," 2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics, pp. 1-6, 2022.

[18] Jamshed, A., Mallick, B., and Kumar, P. “Deep learning-based sequential pattern mining for progressive database. Soft Computing”, vol.24, no. 22, pp. 17233-17246 2020.

[19] Jamshed, A., Mallick, B.,and Bharti, R. K. “An Analysis of Sequential Pattern Mining Approach for Progressive Database by Deep Learning Technique” In 2022 6th International Conference on Intelligent Computing and Control Systems, pp. 1409-1415, 2022.

[20] Shanthi, D., Kuncha, P., Dhar, M. M., Jamshed, A., Pallathadka, H., and JE, A. L. K. The Blue Brain Technology using Machine Learning. In 2021 6th International Conference on Communication and Electronics Systems, pp. 1370-1375, 2021

[21] Cinar, I., and Koklu, M., “Classification of rice varieties using artificial intelligence methods” Int. J. Intell. Syst. Appl. Eng. Vol. 7, no. 3, pp. 188–194, 2019.

[22] Goel, S., Agrawal, R., and Jain, P., “Analysis of Techniques and Methods for Automated EEG signal for Epilepsy Diagnosis: A Review” International Journal of Computer Sciences and Engineering vol. 6 no. 4, pp.429-439,2018.

[23] Goel, Sachin,et. al. "Advancement in Healthcare Systems by Automated Disease Diagnostic Process Using Machine Learning." IJEA vol.14, no.3 2022: pp.1-15.