




An Automatic Rice Grain Classification for Agricultural Products Marketing

Manjula Sri Rayudu¹ , Lakshmi Kala Pampana¹ 

Shalini Myneni¹, Sruthi Kalavari¹ and Raghupathy Reddy Madapa² 

¹ VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, TS 500090, India

² Wells Fargo International Solutions Pvt. Ltd, Hyderabad, TS 500032, India

manjulasree_r@vnrvjiet.in

Abstract. Use of Technology in agriculture and marketing of agricultural products is essential in sustainable and quality production. Some of the usage areas of these technologies are quality control and classification of grains. To optimize the rice production and processing industry, ensuring best product quality and meeting consumer demands effectively, different varieties of rice grain need to be classified accurately and consistently. Manual classification of rice is laborious, time consuming, inconsistent, and inefficient. Our main objective is developing an Artificial Intelligence (AI) based automated model that can analyze and classify rice grains with high accuracy, allowing for higher throughput and increased productivity. In such, we proposed a Machine Learning (ML) based approach to classify five classes of rice varieties. Investigated the results of five classifiers namely, Logistic Regression (LR), K-Nearest Neighbors (KNN), Naive Bayes (NB), Decision Tree (DT) and Random Forest (RF). The RF classifier has given 99.40% accuracy in classifying the five varieties of rice grains.

Keywords: Agriculture Marketing, Machine Learning, Rice varieties.

1 Introduction

Rice is most grown crop and staple food in continents like Asia, provides most of the carbohydrates in their food diet. Around 905 of Asian countries prefer rice as their major food, whose demand and economical aspects are increasing day by day which is to be considered. Rice variety classification facilitates market segmentation based on consumer preferences and demands. Consumers often have specific preferences for rice grains, such as aroma, texture, cooking quality, and taste. By classifying rice varieties based on these characteristics, producers and retailers can offer a diverse range of rice options to cater to different consumer segments. This helps in meeting consumer preferences, enhancing customer satisfaction. So, for increased demand of production, manual classification method is very time consuming, laborious, inconsistent, and inefficient resulting lower throughput and less productivity. Hence, in this paper, an automatic rice grain classification model based on machine learning meth-

© The Author(s) 2023

C. Kiran Mai et al. (eds.), *Proceedings of the Fourth International Conference on Advances in Computer Engineering and Communication Systems (ICACECS 2023)*, Atlantis Highlights in Computer Sciences 18, https://doi.org/10.2991/978-94-6463-314-6_21

odology is presented to classify the five varieties of rice grains accurately. The objective of this work is to device automation for quality control and analysis with reduced effort, price and time and to provide demand based production.

In this work five different varieties of grains namely Basmati, Karacadag, Arborio, Ipsala and Jasmine are considered and is shown in Figure1. In the present work machine learning techniques are employed to classify rice varieties. Five classification models are chosen to classify five varieties of rice based on morphological and shape features. Five classification algorithms KNN (k- nearest neighbor), DT (decision-tree), LR (logistic-regression), NB (naïve-bayes) and RF (random-forest) algorithms are employed on the data. The dataset has 75000 images and 15000 of each grain type rice. The images in the dataset are RGB images with a resolution of 250x250.

2 Literature Surve

2.1 Literature survey on rice grain quality grading:

The existing literature in the classification of quality rice grains and rice varieties is presented below. The authors[1] have image processing methods to identify the quality of rice in terms of chalkiness and Normal on two different varieties of rice grains and employed classification on two different rice dataset using SVM classifiers. The binary categorization of rice yielded 98.5 % classification accuracy. The SVM algorithm took too long duration for training the model especially on larger dataset and the SVM algorithm underperforms for less variance features between the classes. The authors [2] employed a mechanism to identify the quality of grains from change in grain shadows between unfilled and filled grains and employed SVM and observed better False negatives compared to the calculation with other traditional methods. A deep CNN algorithm is used by the author[3] to categorize the rice kernels as 3 different classes of 7309 rice grains to avoid manual labor associated in the classification. . The authors [4] to categorize the quality of rice grains employed CNN,RCMNN and FRCNN to identify diseases rice kernels, a validation accuracy of 80.77% is yielded. But is limited to a smaller data set. They used 200 species of data to classify the rice variety for 3 classes using convolutional neural network (CNN). Neural network-based rice varieties classification was done by authors in [5], they employed feature extraction based classification with CNN.. In which the rice is classified as five classes with the Artificial, Deep and Convolutional Neural Networks (ANN,DNN and CNN) on large data. As per the literature, the deep learning algorithms are performed to the task of categorization of rice data using larger dataset with 75000 images. Even though these ANN, DNN, CNN achieved highest classification accuracy, considering the hardware requirement, it needs high complexity resources to train on huge data. Hence in this proposed work , machine learning algorithms on same dataset with 75000 images are experimented to get equivalent accuracy.

2.2 Literature survey on machine learning classification methods:

Many ML algorithms were discussed in the literature for various applications. In this section literature review on various methods for different applications, which are suitable to classify based on size, shape color and texture are presented in this section. The authors of [6] described K-nearest neighbors and mention that this is a weak rule because classification depends on a single sample. Encyclopedia of Machine Learning [7] described the concepts and the terms associated with machine learning. [8] described the optimized rules set from a large set of rules applicable, for the training of large data set efficiently. Random Forest classifier, the decision tree based classifier and SVM are discussed in [9] and [10] as well known for their simplicity and efficiency. [10] described feature based classification with Random Forest, Naive-Bayes, Ada-boost and Quadrant Discriminative Analysis. Feature extraction methods were implemented from images before employing [classification algorithms. The classification methods mentioned in [9] and [10] are also utilized in medical image analysis for the feature-based classification of diseases. [11-15]. Optimized KNN, GWO-neural network optimization and Neural network methods were employed for medical signal classification. [16][17][18].

3 Methodology

The present work is employed on five Indian & Italian promising rice varieties namely, Basmati, Arborio, Jasmine, Ipsala and Karacadag. A rice data-set with 75000 samples used for training and validation. The data set is further divided as 80% is allocated for training and 20% for testing data and then normalized. Figure 1 shows the images of all the five varieties. From the rice images 14 features are identified and extracted using morphological methods and computation tools. Manually Rice categories are primarily identified from their physical appearance, in terms of size, shape and color. Further identification is done through texture, weight, stickiness, chalkiness, odor etc. The rice categories are mostly basmati (which means fragrant) but differ in length and shape. Hence in this proposed method the morphological features & size and shape are chosen to be the features to perform vision-based classification. The methodology followed for the classification is shown in Figure 2. The chosen features are extracted by the researchers and presented as a second data set in [5] from the 75000 rice images. These feature data set parameters are used, and classification is done with 5 different classification methods.

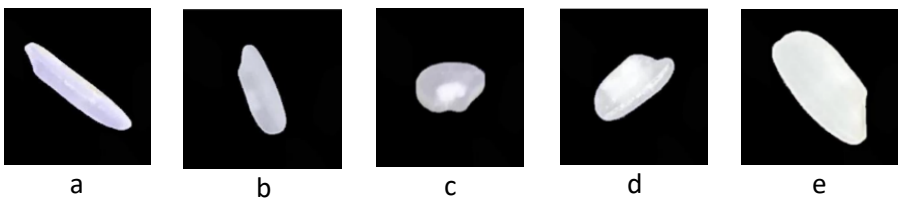


Fig.1. Rice grain images-(a)Basmati (b)Karacadag (c) Ipsala (d)Arborio (e)Jasmine

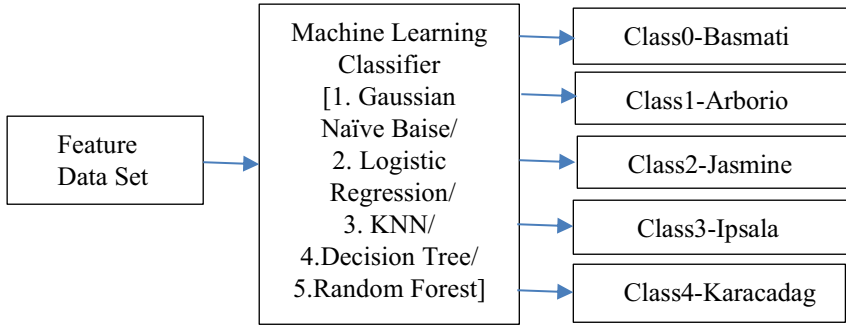


Fig. 2. Methodology

3.1 Significant Features:

There are about 106 features that are identified to distinguish rice varieties, out of which twelve morphological and two shape related features are identified as predominant features that can distinguish the chosen rice grain varieties. These features are extracted using morphological image processing techniques and measurement tools [5]. The following paragraph provides the morphological and shape features considered for this work. Morphological features include -Major and Minor axis length, Area, perimeter, eccentricity, extent, solidity, vertical and horizontal equivalent-diameter, convex area, roundness, and compactness Shape features are calculation of major and minor axis length divided by area. Statistical features are colour features and two features Mean and Standard Deviation are considered, along with the morphological and shape feature based classification. Feature data set from 15000 images of the five categories of rice [5] has been considered for the proposed work.

3.2 Classification Models:

Initially seven classification models are chosen KNN (K Nearest Neighbor), SVM (support Vector machine), DT (Decision Tree), LR (Logistic Regression), MLP (Multilayer perception), NB (Naive Bayes) and RF (Random Forest) algorithms are employed to classify the rice varieties. Though MLP, LR and NB are not suitable for the categorization of rice, all seven models were employed. Decision tree (DT) is chosen as it can classify categorical and nonlinear data, DT is chosen as it is easy to consolidate data into classes and reliability is high. Logistic regression is applicable to classify if the classes are linearly separable. Multilayer perception (MLP) is one of the most widely used Artificial neural network model, which has given its performance similar to NB. K Nearest Neighbor algorithm (KNN) is one of the simplest machine learning algorithms, based on the supervised learning method, is K-Nearest Neighbor. To get the most out of the model, it is essential to select a suitable value for k in the KNN. Since there are few votes, the model's error rate will be high if K is low, especially for new data points. As a result, the model is overfitted and extremely sensitive to input

noise. Conversely, if K has a high value, the model's boundaries become flimsy and there are more misclassifications. The model is underfitted in this instance. So, from the experimentation on K neighbors, it was observed that $K=5$ is the best fit for the model. SVM determines its neighbors based on sample data presented to the algorithm and assumes that estimates are made for new data. In the preliminary results SVM has shown similar results as that of KNN. Random Forest (RF) is a fast supervised machine learning algorithm based on majority voting. Instead of relying on one decision tree, the random forest takes the prediction from each tree and bases its prediction of the final output on the majority votes of predictions. Results and analysis of five models NB, LR, KNN, DT and RF are presented in this paper.

3.3 Evaluation Metrics:

To evaluate the ability of the classification models for predicting various types of rice grains diverse performance parameters are considered. They include Accuracy, Precision, Recall and F1 Score as provided in the set of equations (1-4), Which are evaluated from fundamental metrics true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) taken from the confusion matrix.

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (1)$$

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \quad (2)$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \quad (3)$$

$$\text{F1 score} = 2(\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (4)$$

4 Results and Discussion

Different machine learning algorithms are applied on the dataset. Initially, 14 Predominant features from morphological and shape features are identified manually from 106 total features. All the five classifiers' performance is observed and compared with earlier research. The Accuracy metrics are shown in Table 1. Out of the five models' random forest has shown highest accuracy of 99.24%; Calculated according to Eq1. Decision Tree has shown the prediction accuracy of 99.20%, nearer to that of random forest. KNN has provided an accuracy of 95.40%, Logistic regression and Naïve Base have provided relatively poor prediction accuracy of 78% and 76.27% respectively. These accuracies are average accuracies with morphological and shape features(A). Later statistical features are considered for classification. Classification accuracy on including statistical features are slightly higher and the average accuracy with Morphological, shape and color feature is shown in column(B). The improvement is meagre. They are presented in table 1.

Table 1. prediction accuracy of proposed and earlier research

Model	Accuracy (Earlier research)	Accuracy (proposed work) (A)	Accuracy (proposed work) (B)
Gaussian Naïve Base	-	76.27	76.20
Logistic Regression	90.03	78.01	78.08
KNN	92.23	95.40	95.45
Decision Tree	97.81	99.20	99.22
Random Forest	98.04	99.24	99.28

Here results of the two classifier models namely KNN and RF are presented elaborately. The confusion matrix explains predictive assessment of classification. The accuracy of a classification is evaluated from true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). Confusion charts for both the classifiers are shown in Fig3 and Fig4. Along with this, the performance metrics and accuracy of the two models are presented in Table 2 and Table3. The tables show individual class metrics for both KNN and RF classifier. Accuracy is computed for each class for both the classifiers. The model’s overall accuracy is the weighted average of all classes. Each model is trained on the same dataset by configuring the parameters. For example, by changing the number of nearest neighbors in KNN, the best accuracy is achieved with K=5.

Table 2. Accuracy of KNN classifier

Class	FP	FN	TP	TN	ACC
0	762	294	2705	11239	0.92
1	340	293	2684	11683	0.95
2	184	219	2870	11727	0.97
3	0	4	3007	11989	0.99
4	200	676	2248	11876	0.94

Table 3. Accuracy of RF classifier

Class	FP	FN	TP	TN	ACC
0	42	18	2981	11959	0.99
1	102	19	2958	11921	0.99
2	25	60	3029	11886	0.98
3	0	12	2999	11989	0.99
4	17	77	2847	12059	0.97

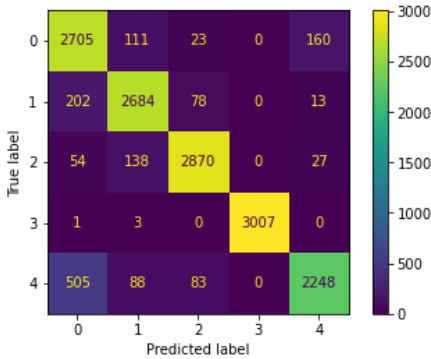


Fig. 3. Confusion chart of KNN classifier.

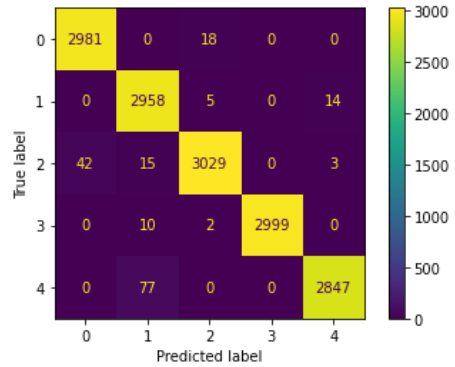


Fig. 4. Confusion chart of RF classifier

Precision, Recall and F1 score are presented in Table 4. From Table 4, it is inferred that there is imbalance of precision and recall between class 0 and Class4, even though accuracy is 92%. For an accurate RF model precision and recall should be balanced for all the classes. Finally chosen RF algorithms to classify all the five varieties of rice grains with 99.24% accuracy while maintaining the well-balanced precision and recall among all the classes and the metrics are listed in Table 2. F score is another performance metric for evaluating the machine learning model performance, for RF, the F1 score is nearer to accuracy. It shows the RF model on the specified dataset is a robust and scalable model.

Table 4. Prediction Results of KNN and RF Classifiers

Class	KNN			RF		
	Precision	Recall	F1 Score	Precision	Recall	F1 Score
0	0.78	0.90	0.84	0.99	0.99	0.99
1	0.89	0.90	0.89	0.97	0.99	0.98
2	0.94	0.93	0.93	0.99	0.98	0.99
3	1.00	1.00	1.00	1.00	1.00	1.00
4	0.92	0.77	0.84	0.99	0.97	0.98

* Class 0-Basmati, Class 1-Arborio, Class 2- Jasmine, Class 3-Ipsala, Class 4-Karacadag

The other metrics TPR, TNR, FPR and FNR of KNN and RF classifiers are presented in Table 5 The tables show individual class metrics for both KNN and RF classifier, which are obtained from the confusion chart parameters. The model’s overall accuracy is the weighted average of all classes. The ROC and PR(Precision-Recall) curves visualize the machine learning model performance in terms of True Positive Rate (TPR) and False Positive Rate (FPR). A good model always has the highest true positive rate at lowest false positive rate. Upon observation of RoC and PT curves, the RF classifier has the highest TPR 0.98 at lowest FPR 0.02 as compared to KNN . Compared to KNN classifier the RF classifier has high precision at high Recall.

Table 5. Performance metrics of KNN classifier

Class	FP	FN	TP	TN	TPR	TNR	FPR	FNR
0	762	294	2705	11239	0.90	0.93	0.06	0.09
1	340	293	2684	11683	0.90	0.97	0.02	0.09
2	184	219	2870	11727	0.92	0.98	0.01	0.07
3	0	4	3007	11989	0.99	1	0	0
4	200	676	2248	11876	0.76	0.98	0.01	0.23

Table 6. Performance metrics of RF classifier

Class	FP	FN	TP	TN	TPR	TNR	FPR	FNR
0	42	18	2981	11959	0.99	0.99	0.003	0.006
1	102	19	2958	11921	0.99	0.99	0.008	0.006
2	25	60	3029	11886	0.98	0.99	0.002	0.019
3	0	12	2999	11989	0.99	1	0	0.003
4	17	77	2847	12059	0.97	0.99	0.001	0.026

5 Conclusion

In conclusion, the developed automatic rice grain classification model in agricultural products marketing is a significant advancement in the industry. By automating the classification process, It minimizes human error and ensures that consumers receive products that meet their desired standards. Five classifiers have been configured to classify rice grains into five categories. Among those Gaussian Naïve Base and Logistic Regression models have shown poor accuracy. KNN, RF and DT models have shown significant improvement in prediction accuracy. The automatic classification model developed with Random Forest classifier has shown better performance in terms of Accuracy as well as F1 Score, which is a measure of balanced sample set. This machine learning model performed 99.24 % accuracy in classifying the rice grains. The model is very robust as it maintained a good balance between precision and recall among all the five classes. The model has a high True positive rate (TPR) 0.98 while having a very less false positive rate (FPR) 0.02. The developed algorithm classifies the rice varieties quickly. The algorithm is scalable. extendable to classify more varieties of rice grains with high accuracy and low complexity. And reproducible with same configuration of the model on the same data set.

References

1. Sun, Chengming, et al. "Evaluation and analysis the chalkiness of connected rice kernels based on image processing technology and support vector machine." *Journal of Cereal Science* 60.2 (2014): 426-432.

2. Liu, Tao, et al. "A shadow-based method to calculate the percentage of filled rice grains." *Biosystems Engineering* 150 (2016): 79-88
3. Lin, P., et al. A deep convolutional neural network architecture for boosting image discrimination accuracy of rice species. *Food and Bioprocess Technology* 11(2018):765-773
4. Ahmed, Tashin, Chowdhury Rafeed Rahman, and Md Faysal Mahmud Abid. Rice grain disease identification using dual phase convolutional neural network based system aimed at small dataset. *AgriRxiv* (2021): 20210263534.
5. Koklu, Murat, Ilkay Cinar, and Yavuz Selim Taspinar. Classification of rice varieties with deep learning methods. *Computers and electronics in agriculture* 187 (2021): 106285.
6. Mucherino, Antonio, et al. K-nearest neighbour- classification. *Data mining in agriculture* (2009): 83-106.
7. Sammut, C., Webb, G.I. (eds) *Encyclopedia of Machine Learning*. Springer Reference Springer, New York, 2 edition, (2017)
8. Huynh, V.Q.P., Fürnkranz, J. & Beck, F. Efficient learning of large sets of locally optimal classification rules. *Mach Learn* 112, (2023) 571–610
9. Abdulkareem, Nasiba Mahdi, and Adnan Mohsin Abdulazeez. "Machine learning classification based on Radom Forest Algorithm: A review." *International Journal of Science and Business* 5.2 (2021): 128-142.
10. Cristianini, N., Ricci, E. (2008). Support Vector Machines. In: Kao, MY. (eds) *Encyclopedia of Algorithms*. Springer, Boston. https://doi.org/10.1007/978-0-387-30162-4_415.
11. Lakshmi Kala Pampana and Manjula Sri Rayudu- Multi-Class Classification of Retinal Abnormality using Machine Learning Algorithms-*International Journal of Performability Engineering*,(Nov) 2022, 18(11), pp. 826–832.
12. Rayudu Manjula Sri; Pendum, Srujana; Dasari, Srilaxmi, "Prediction of Severity of Non Proliferated Diabetic Retinopathy Using Machine Learning Techniques", *Journal of Computational and Theoretical Nanoscience*, Volume17, issue 9&10, October 2020.
13. SriLaxmi Dasari,Boo Poonnguzhali, Manjula Sri Rayudu -an efficient machine learning approach for classification of diabetic retinopathy stages, *Indonesian Journal of Electrical engineering, and Computer science*, Volume30, issue 1 , 2023.
14. Lakshmi Kala Pampana and Manjula Sri Rayudu, "Classification of Retinal Vessel Tortuosity as Severity Grades using Random Forest Classifier", *National Conference on Advances in Science, Agriculture, Environmental & Biotechnology*, ISBN:978-93-90150-32-8Edn: 482, 2022.
15. Lakshmi Kala Pampana and Manjula Sri Rayudu, *Retinal Artery Vein Crossover Abnormality Classification using Optimized Machine Learning methods*", *National Conference on Advances in Electrical, Electronics & Computer Engineering*, ISBN : 978-93-90150-31-1 Edn: 487, 2022.
16. LV Rajani Kumari, Y Padma Sai- Classification of arrhythmia beats using optimized K-nearest neighbor classifier, *Intelligent Systems: Proceedings of ICMIB* ,p 349-359,2021.
17. Mulam, Harikrishna and Mudigonda, Malini. "EOG-based eye movement recognition using GWO-NN optimization" *Biomedical Engineering / Biomedizinische Technik*, vol. 65(1), 2020, pp. 11-22.
18. Mulam H, Mudigonda M. Empirical mean curve decomposition with multiwavelet transformation for eye movements recognition using electrooculogram signals. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*. 2020;234(8):794-811

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

