# A NOVEL APPROACH FOR RICE GRAIN CLASSIFICATION USING MACHINELEARNING TECHNIQUES

**Mrutyunjaya M S[1*], Harishkumar K S[2]**

**Article History:** **Received:** 01.02.2023 **Revised:**07.03.2023 **Accepted:** 10.04.2023

**Abstract:**
Rice provides essential nutrition and energy to the global population, contributing more than twenty percent of total caloric intake. Choosing the appropriate rice amongst the available verities in the market has become a biggest challenge due to their identical physical appearance. Biological and chemical methods such as DNA analysis and alkaline testing are costly, time consuming, and not accessible to the average person, making them unsuitable for the identification of rice grain varieties. Further, few image processing techniques have been incorporated for identification of different rice varieties which seems to be less accurate for verities with same morphological features. This paper describes a hybrid approach for enhancing the accuracy of rice variety identification using machine learning and digital image processing techniques. A total of 1,00,000 rice grain images were selected, with 20,000 for each variety. With the use of image processing techniques, the images were preprocessed in preparation for feature extraction. Initially 12 morphological, 4 shape and 25 texture features were extracted using bounding box and region property methods of image processing on 5 different verities of rice, later these extracted features were fed into different supervised classification algorithms like Support Vector Machines (SVM), K - Nearest Neighbors (K-NN), Decision Tree (DT) and Naıve Bayes (NB). The algorithms with the highest average classification accuracy of 99.24 is achieved with K - nearest neighbors. According to the performance measurement results, the study was successful in identifying the different types of rice.

**Keywords:** Classification, Rice, DNA, SVM, Naive Bayes.

[1*]Assistant Professor School of Computer Science Engineering, Presidency University, Bangalore, Karnataka, India, Email: mrutyunjaya@presidencyuniversity.in
[2]Assistant Professor School of Computer Science Engineering, Presidency University, Bangalore, Karnataka, India, Email: harishkumar@presidencyuniversity.in

***Corresponding Author:** Mrutyunjaya M S
*Assistant Professor School of Computer Science Engineering, Presidency University, Bangalore, Karnataka, India, Email: mrutyunjaya@presidencyuniversity.in

**DOI:** XYZ

## 1 Introduction

Most of the rural population of a developing country relies on agriculture for their livelihood, making it the sector's central importance. Rice is a dietary mainstay for over half of the global population, offering vital sources of energy, protein, vitamins, and minerals. Rice is also a key in the global food system for its high production and low cost. In addition, rice is an important part of culture and tradition in many countries, and it is often used in religious ceremonies and festivals. It is an important crop for food security, poverty alleviation, and rural development.

People's expectations of the quality and safety of the food they consume rise in tandem with their growing level of awareness. This is why Accurate, prompt, and impartial methods are necessary to evaluate the quality of cereal grains. Hobson et al. [1] have suggested methods using image processing to classifying rice types according to size, shape, and colour. Eight distinct types of Japanese rice were successfully catalogued. A camera was set up so that it would focus on individual grains of rice set against a black matte background. The grains' distinctive forms and textures were the primary focus of the image analysis. Their work made use of the following variables. The shortest and most basic metric is the La metric, which measures the mean length. Calibrating the image gave us the length and width in terms of pixels. The chain code of each shape is used to create a set of shape features that are based on diameters. To ensure standardization, the diameters were measured using pixels located on opposite halves of the chain code. Simply put, the aspect ratio (Ra) is the ratio between the minimum and maximum diameters. The compactness ratio (Rj) is a number between 0 and 1 that describes how compact a shape is (spherical). This information was used to correctly identify eight distinct types of Japanese rice. Digital imaging methodology developed to study various traits for rice variety identification has been presented by Kiruthika et al. [2].

## 2. LITERATURE SURVEY

In this paper, we attempted to create a superior system that could outperform all other systems. We came across many results of existing systems and went through research findings throughout this procedure. We examined many scholars' discoveries by researching various publications and research theses, and we used numerous strategies based on their research and conclusions.

Both Pasmati single mattai and IR16 rice were used for the research. They analysed the rice in terms of its length, area, and aspect ratio, all of which are already established standards. The work entails the steps of converting the image of the rice grains into a binary image. This image was segmented in order to separate it into distinguishable parts. For the purpose of this work, edge-based segmentation is employed. After obtaining Region Props through blob analysis, which measures the area and bounding box of image regions, we were able to use these metrics to determine the best candidates for feature matching[3].

To ensure the rice was pure white and to count the number of broken seeds among the head grains, Yadav Jindal., [4] used digital image analysis. It was possible to extract and quantify morphological and shape features of the different rice grain varieties. 45 morphological features were used for classification by Dubey et al. [5]'s artificial neural network. They were able to achieve an 88% success rate in classifying all grains, and this rate was found to increase with the number of features used in the analysis. Using 74 extracted morphological features, Zapotoczny et al. [6] discussed the use of morphological characteristics to classify five distinct types of barley. PCA, LDA, and NDA were employed to effectively classify barley varieties. Using neural networks and image processing techniques, patterns can be recognized. Abirami et al. [7] were able to classify different varieties of basmati rice with an accuracy of 98.7% by extracting various morphological features. [8]The multi-class support vector machine classification (M-SVM) algorithm was used by Sethy & Chatterjee [9] to classify 6 different types of rice, with 92% accuracy, using extracted geometric and texture features. The machine learning techniques of Multi-Layer Perceptron classifier (MLP), Support Vector Machine classifier (SVM), K-Nearest Neighbours classifier (K-NN), and Decision Tree classifier (DT) were used by Koklu Ozkan [10]. to categorise seven distinct types of dry beans based on their morphological and shape features, with SVM yielding the highest classification accuracy (93.13%).

## 3. METHODOLOGY

The methodology of rice variety classification is described in detail in the

Figure 1 shown below.

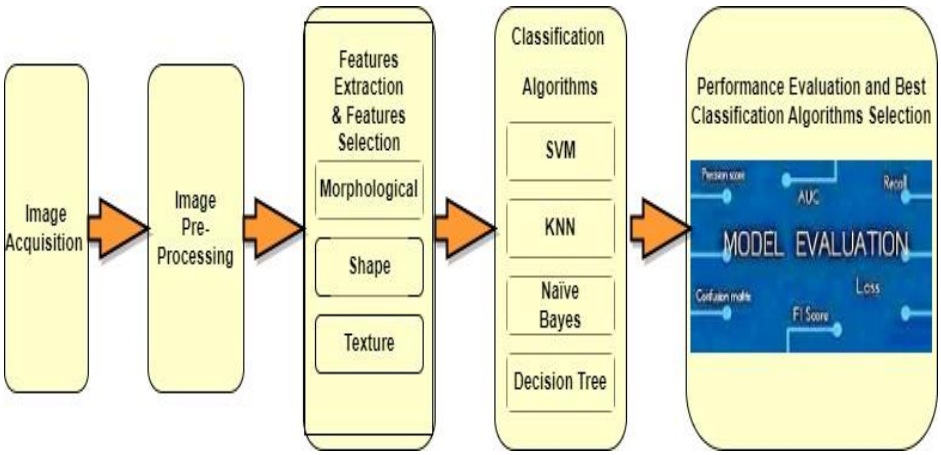


**Fig. 1.** Aproposed Architecture for classification.

**3.1Image Acquisition**:
A closed box was used to house a camera with a CCD imaging sensor for the purpose of capturing images of five different types of rice grains. The box was equipped with an internal lighting system and a barrier to shield external light, and the height of the camera was set to 15 cm [11]. A total of one hundred thousand images were taken of the different types of rice, with twenty thousand images of each variety. These images were then stored in a computer for further preprocessing. The background of the box was kept black to facilitate this process. The varieties of rice used were Arborio, Basmati, Ipsala, Jasmine, and Karacadag [12].

**3.2Image processing:**
Global Thresholding, also known as the Otsu method, was used to convert the collected images first into grey scale images and subsequently into binary images. The following equation is used by OTSU's method to choose the optimal Thresholding value K in order to maximize the value of G.

G = Pj(Ij − IT )2 + Pb(Ib − IT )2 (1)

Where, Pj = proportion milled rice pixels, Pb = proportion of background pixels, Ij = milled rice mean gray value, Ib = background's mean gray value, IT = whole image mean gray value. By applying morphological opening and closing operations the unwanted background pixels were eliminated. Below figure shows the different stages of image preprocessing operations[13].

Features Extraction: In this study 42 features were extracted for each rice variety. 12 morphological, 5 shape and 25 texture features. Morphological and shape features were extracted by using bounding box and region props techniques available in image processing [14]. By building a Grey level co-occurrence matrix for each image type and varying the pixel distances and angles, texture features were retrieved.

**3.3Morphological Features**

**3.3.1** Area: The area of any object in an image is determined by the total number of pixels enclosed within its boundary.

**3.3.2** Perimeter: The total number of pixels along the outer edge of a shape determines its perimeter.

**3.3.3** Eccentricity (E): The eccentricity of the circle also symbolizes the unique characters of each individual in the region, and how the moments in the region are distinct and special.

**3.3.4** Equivalent Diameter (ED): It is the circumference of a circle whose area is equal to that of a rice grain. The formula for computing equivalent diameter is following Equation**.**

$$\boldsymbol{ED = \sqrt{\frac{4A}{\pi}}} \qquad \textbf{(2)}$$

**3.3.**5 Solidity (S):Is The ratio of the number of pixels in the rice grain region to the number of pixels in the convex body. The formula for Computing solidity is specified In Equation 3,

$$\boldsymbol{S = \frac{A}{CA}} \qquad \textbf{(3)}$$

**3.3.6** Convex Area (CA): The Convex Area of a rice grain is the number of pixels of the smallest convex polygon that can fit around it.

**3.3.7** Extent (Ex): The extent of the bounding box is the ratio of the number of pixels in the bounding box to the number of pixels in the region of the rice grain.

**3.3.8** Compactness (Co): It is computed by finding ratio of equivalent diameter vs major axis length. The calculation formula is given in Equation.

$$Co\frac{ED}{dmax} \quad (4)$$

**3.3.9** Thinness ratio: It measured the roundness of the seed.

$$TR = \frac{(P^2)}{4\pi A} \quad (5)$$

**3.3.10** Major axis length (Dmax): Dmax of grain in the image is the length of the longest ellipse that encapsulates a single grain, measured in pixels.

**3.3.11** Minor axis length(Dmin): Dmin of a grain in the image is the length of the shortest ellipse that completely encloses a single grain, measured in pixels.

**3.3.12** Aspect Ratio (Ra): The ratio of a single grain's shortest to longest diameters is known as the Aspect Ratio.

$$R_a = \frac{D_{min}}{D_{max}} \quad (6)$$

### 3.4. Texture

Texture analysis involves describing the texture content of different regions in an image. This involves quantifying characteristics like roughness, smoothness, bumpiness, or silkiness based on how the pixel intensities vary across the image. Texture analysis can be used to segment an image by identifying texture boundaries. This technique is particularly useful when objects in an image are defined primarily by their texture, rather than their intensity, and standard threshold methods are ineffective. [15]. The study employed various texture features as listed below to analyze the texture content of the images:

- **Co-occurrence Matrices:** The spatial gray level co-occurrence (GLCM) is a statistical method used to describe the texture of an image. It is based on the spatial relationship between pixels in an image. GLCM is used to extract features from an image by counting how often different combinations of gray levels occur in an image. The method utilizes a 'G * G' GLCM matrix Pd, which is defined by a displacement vector d = (dx, dy). The entry (i, j) of Pd represents the number of occurrences of the gray level combination (i, j) at a distance of 'd' apart. This can be expressed as Pd (i, j) = ((r, s), (t, v)): I(r, s) = i, I(t, v) = j.
- **Contrast:** is a measure of the amount of variation in the local gray-level values in a given image. It provides information on how different the gray levels are in a local region of the image. A high contrast value indicates that the gray-level differences between neighboring pixels are large, while a low contrast value indicates that the differences are small.
- **Correlation:** Correlation is a measure of the likelihood that two specific pixels in a gray-level co-occurrence matrix will occur together. The correlation between the gray-level values of adjacent pixels provides information about how closely related they are. A high correlation value suggests that the gray-level values of the neighboring pixels are similar, while a low correlation value indicates that they are dissimilar [16].
- **Energy:** The total amount of energy present in a GLCM is calculated by summing the squared elements of the matrix. This can be expressed as:

$$Energy = \sum i \sum j \left(P_{ij}\right)^2 \quad (13)$$

Where Pij is the value of the matrix at position (i,j). It is also known as angular uniformity or the second moment of angularity. Energy provides information on how uniform or homogeneous the texture is in different regions of the image. A low energy value suggests that the gray-level values of neighboring pixels differ, while a high energy value indicates that the values are similar[17].

- **Homogeneity:** It is a measure of the similarity of gray levels in the image. The homogeneity of a gray-level co-occurrence matrix (GLCM) is calculated by summing the product of the probabilities of the two gray levels occurring together in each pixel:

$$H = \sum i \sum j \left(P_{ij}\right)^2 \quad (14)$$

Where Pij is the probability of gray level i being adjacent to gray level j. A high homogeneity value indicates that the values of neighboring pixels are close to each other, while a low homogeneity value indicates that the values of neighboring pixels differ significantly.

### 3.5 Features Selection

Once with features were extracted to study the distribution and relationships among features across different rice varieties a statistical data visualization plots like sea-born and pair plots were generated as shown in figure no. These plots can be used to visualize distributions of data and to identify relationships between variables. Sea-born also offers functions to make it easier to visualize regression models and to plot statistical time series [18].

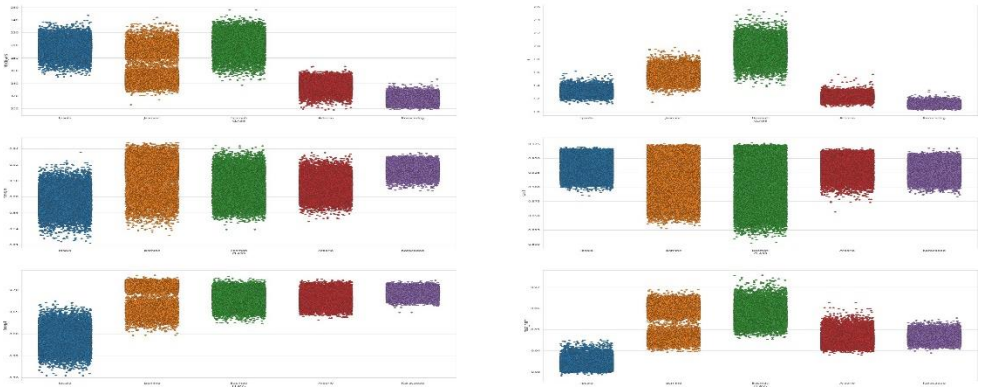

**Fig 2.** Sea-born graph for selected attributes

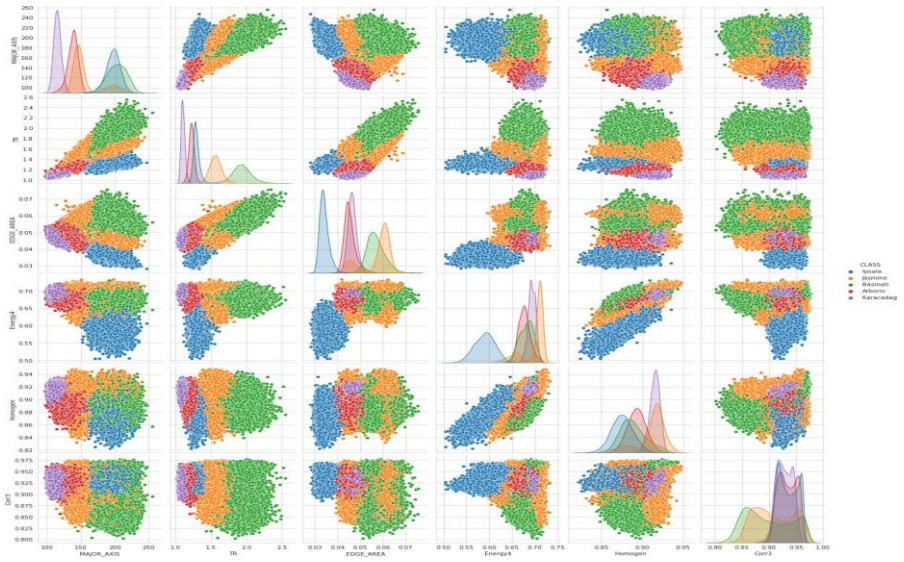

**Fig 3.** Pairplot graph for selected attributes

## 4. Result And Discussion

For this study, a dataset of 100,000 images was used, consisting of 20,000 images for each of the five rice varieties. To train the classification models, 75% of the images (75,000) were randomly selected for the training set, while the remaining 25% (25,000) were used for testing purposes. The extracted features were classified using traditional classification algorithms, including decision tree classifier (DT), k-nearest neighbor classifier (KNN), Naive Bayes classifier (NB), and support vector machine classifier (SVM).

### 4.1 Decision Tree Classifier (DT):

Decision Tree classification algorithms are supervised learning algorithms used to classify data points into specific categories. They are based on decision trees, which are a graphical representation of the possible outcomes of a decision process. Decision tree algorithms are used in a variety of applications, including supervised learning, recommendation systems, and medical data analysis. They are often used in combination with other classification algorithms to improve accuracy and provide more accurate results.

**Table 1:** Decision Tree Classification Report

| Rice Type | Precision | Recall | f1-score | Support |
|---|---|---|---|---|
| Arborio | 0.98 | 0.98 | 0.98 | 4943 |
| Basmati | 0.99 | 0.99 | 0.99 | 5002 |
| Ipsala | 1.00 | 1.00 | 1.00 | 4923 |
| Jasmine | 0.99 | 0.99 | 0.99 | 5126 |
| Karacadag | 0.99 | 0.98 | 0.98 | 5006 |
| accuracy | | | 0.99 | 25000 |
| macro avg | 0.99 | 0.99 | 0.99 | 25000 |
| weighted avg | 0.99 | 0.99 | 0.99 | 25000 |

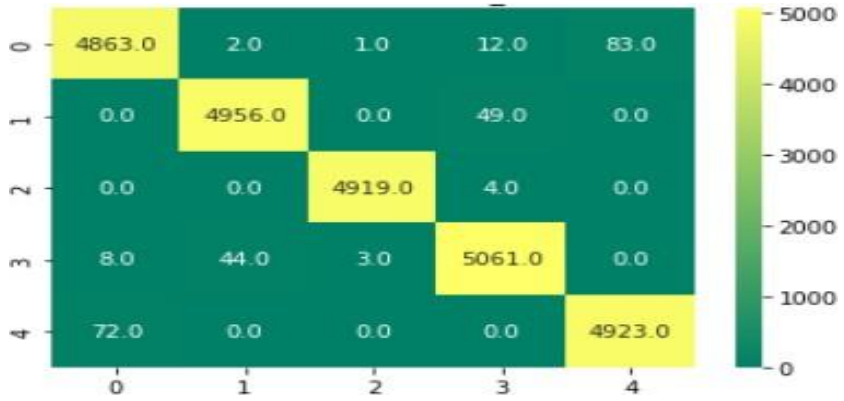


**Fig 4.** Decision Tree Classifier Confusion Matrix

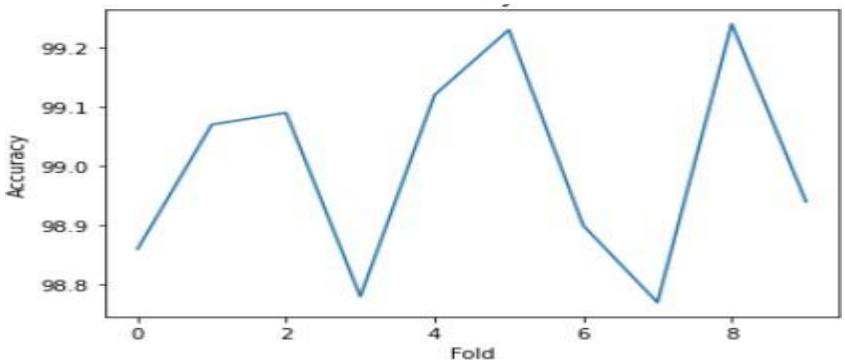


**Fig 5.** Decision Tree Classifier Fold Accuracy Visualizer

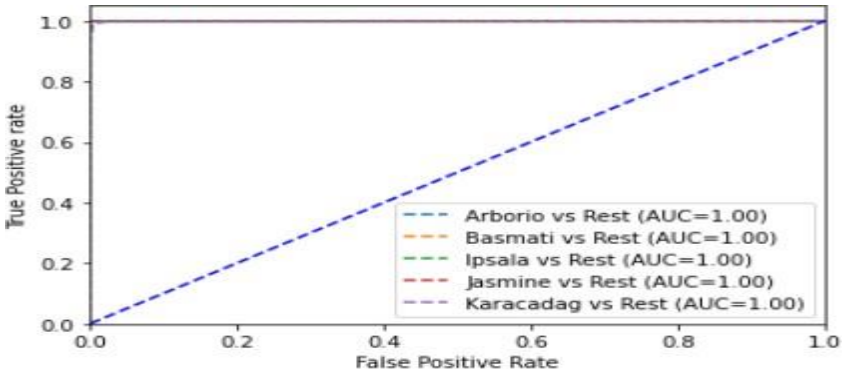


**Fig 6.** Multi-class decision tree classifier ROC curve

### 4.2 K-Nearest Neighbor Classifier (K-NN)

KNN Is a supervised machine learning algorithm that is often used for solving classification problems. Unlike some other algorithms, KNN does not build a model before making predictions; rather, it stores all the data and waits for new data points to be introduced. When a new data point is presented, the algorithm classifies it based on its similarity to other data points. The KNN algorithm operates on the principle that a data point is classified based on a majority vote of its k-nearest neighbors, where user specifies the value of k as an integer value. Thus, KNN is a type of lazy learning.

**Table 2: KNN Classification Report**

| Rice Type | Precision | Recall | f1-score | Support |
|---|---|---|---|---|
| Arborio | 0.99 | 0.98 | 0.99 | 4943 |
| Basmati | 0.99 | 0.99 | 0.99 | 5002 |
| Ipsala | 1.00 | 1.00 | 1.00 | 4923 |
| Jasmine | 0.99 | 0.99 | 0.99 | 5126 |
| Karacadag | 0.99 | 0.99 | 0.99 | 5006 |
| accuracy | | | 0.99 | 25000 |
| macro avg | 0.99 | 0.99 | 0.99 | 25000 |
| weighted avg | 0.99 | 0.99 | 0.99 | 25000 |

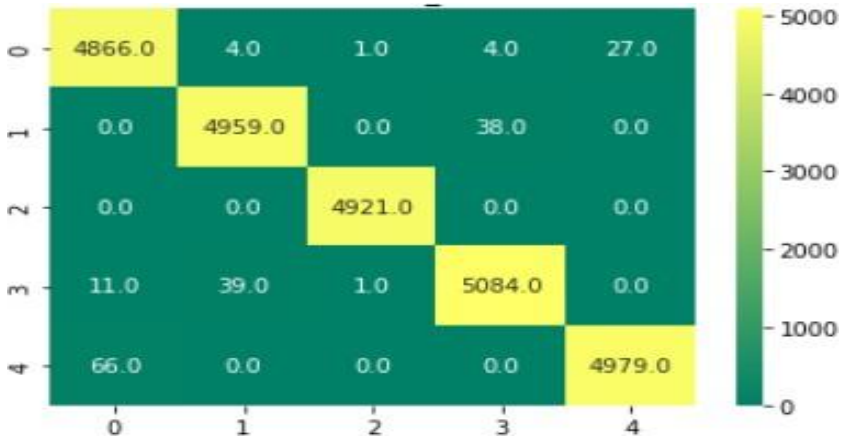


**Fig 7.** K-NN Confusion Matrix

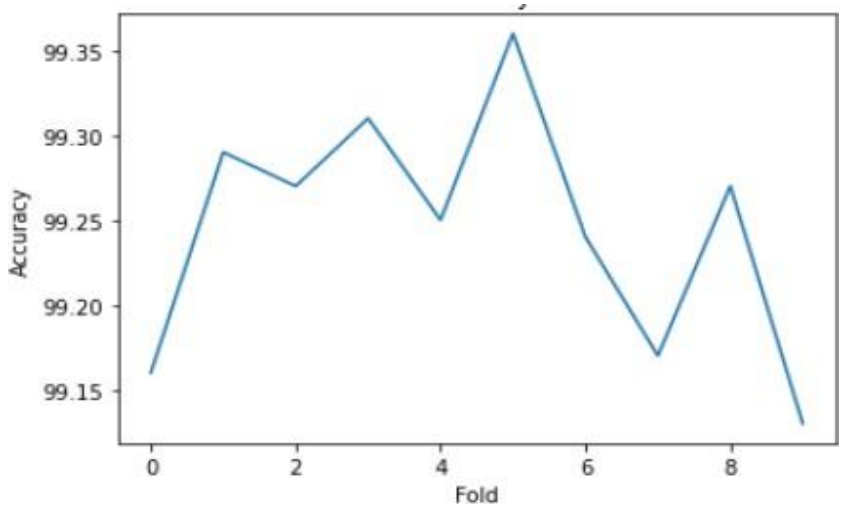


**Fig 8.** K-NN Fold Accuracy Curve Visualizer

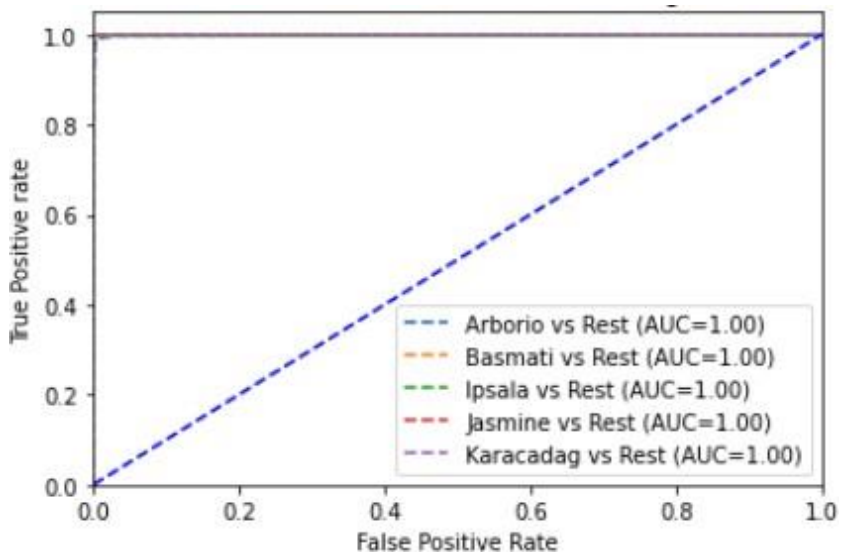


**Fig 9.** Multi-class K-NN ROC curve

### 4.3 Naive Bayes Classifier

Naive Bayes classification algorithms are a family of supervised learning methods that use Bayes' theorem to make predictions, assuming that each feature is independent of all other features. Naive Bayes is a popular and simple technique for supervised learning that has many applications, including spam filtering, text classification, and medical diagnosis. Because it is fast to train and easy to implement, it is a common choice among machine learning practitioners.

**Table 3:** Naive Bayes Classification Report

| Rice Type | Precision | Recall | f1-score | Support |
|---|---|---|---|---|
| Arborio | 0.98 | 0.98 | 0.98 | 4943 |
| Basmati | 0.98 | 0.96 | 0.97 | 5002 |
| Ipsala | 1.00 | 1.00 | 1.00 | 4923 |
| Jasmine | 0.95 | 0.98 | 0.97 | 5126 |
| Karacadag | 0.98 | 0.98 | 0.98 | 5006 |
| accuracy | | | 0.98 | 25000 |
| macro avg | 0.98 | 0.98 | 0.98 | 25000 |
| weighted avg | 0.98 | 0.98 | 0.98 | 25000 |

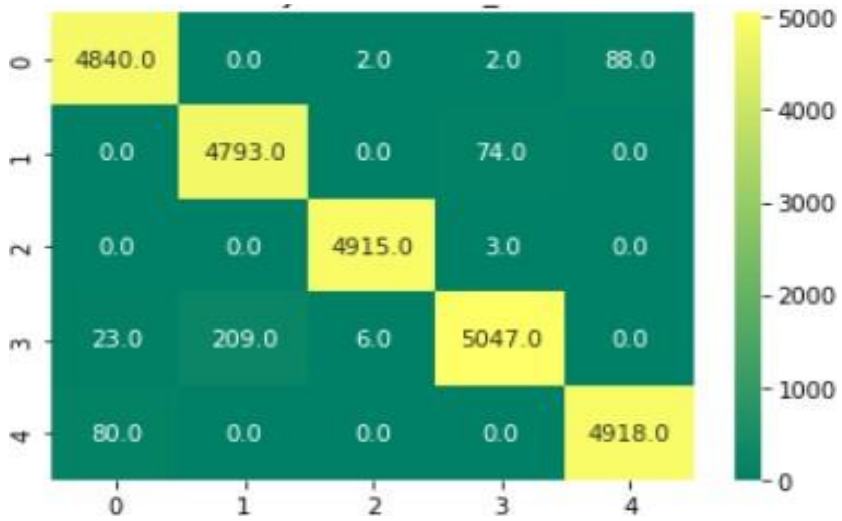


**Fig 10.** Naive Bayes classification Confusion Matrix

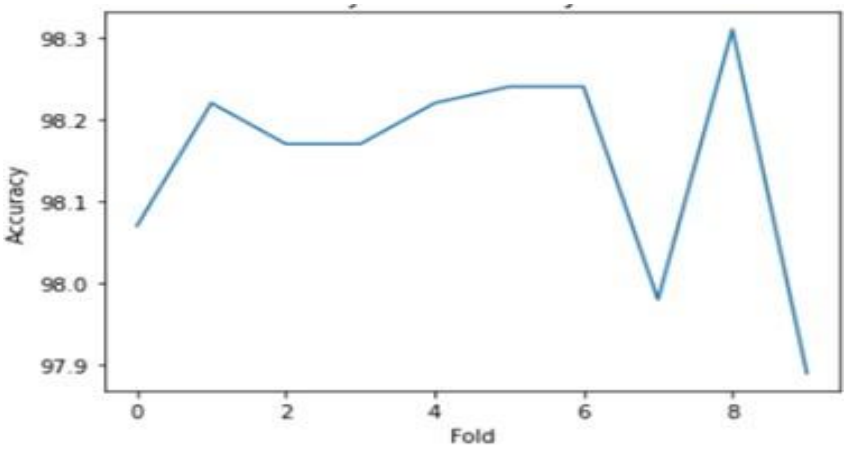


**Fig 11.** KNN Fold Accuracy Curve Visualizer

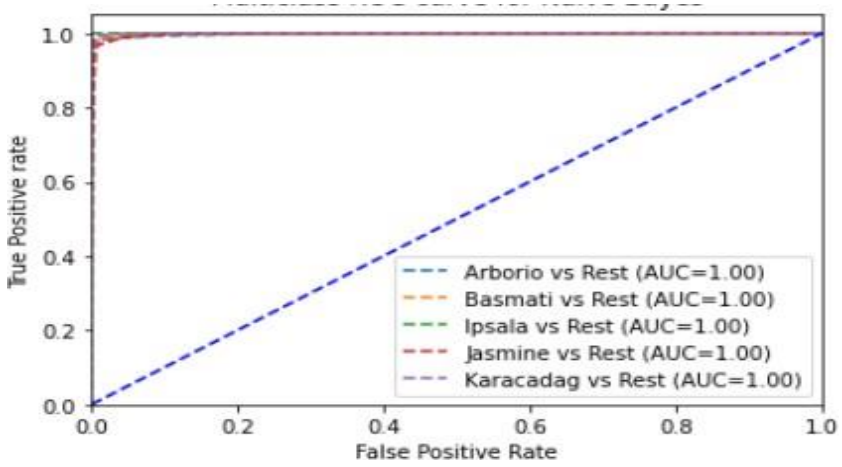


**Fig12.** Multiclass ROC Curve NB

**4.4 Support Vector Machine Classifier (SVM)**

SVM is a supervised learning algorithm that is widely used for classification and regression tasks. It is a powerful and influential classification model. Its main goal is to find a hyperplane in an N-dimensional space (where N-1 is the number of features) that can best separate data points into different classes. The algorithm strives to identify the maximum margin hyperplane, which is the plane that has the farthest distance from the nearest data points of each class. The support vectors are the data points that are closest to the decision boundary of the hyperplane. SVM is a versatile algorithm that can handle both linear and non-linear datasets. It is also useful in high-dimensional spaces and can handle large feature sets. Because it is resistant to outliers and can produce precise results even with unbalanced data, it is a common choice for classification tasks.

**Table 4:** Support Vector Machine Classification Report

| Rice Type | Precision | Recall | f1-score | Support |
|---|---|---|---|---|
| Arborio | 0.99 | 0.98 | 0.99 | 4943 |
| Basmati | 0.98 | 0.97 | 0.98 | 5002 |
| Ipsala | 1.00 | 1.00 | 1.00 | 4923 |
| Jasmine | 0.97 | 0.98 | 0.98 | 5126 |
| Karacadag | 0.99 | 0.99 | 0.99 | 5006 |
| accuracy | | | 0.99 | 25000 |
| macro avg | 0.99 | 0.99 | 0.99 | 25000 |
| weighted avg | 0.99 | 0.99 | 0.99 | 25000 |

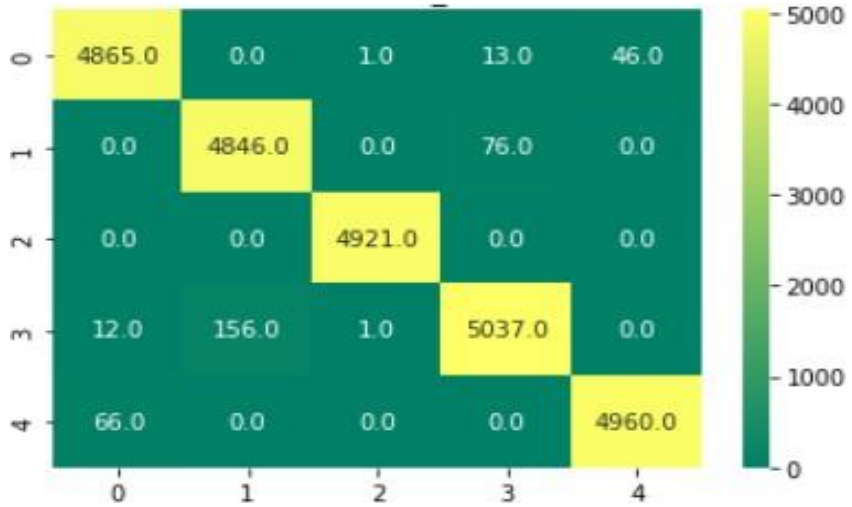


**Fig13.** SVM Confusion Matrix

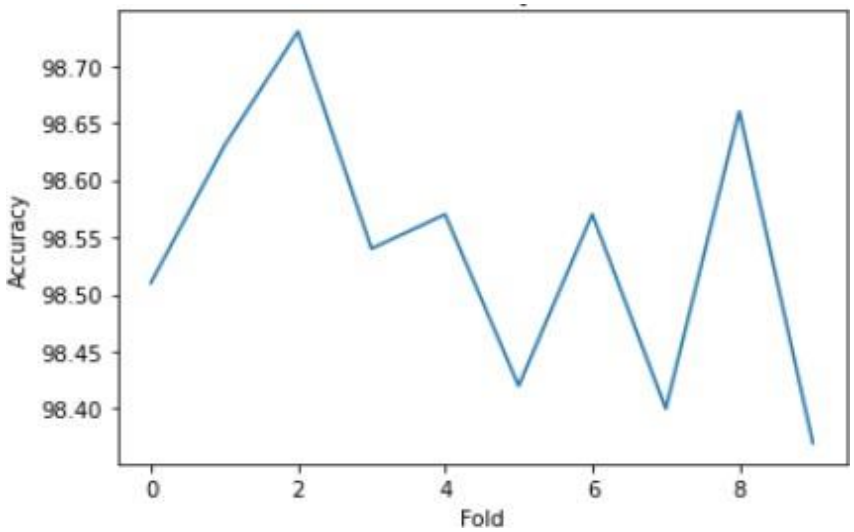


**Fig 14.** SVM Fold Accuracy

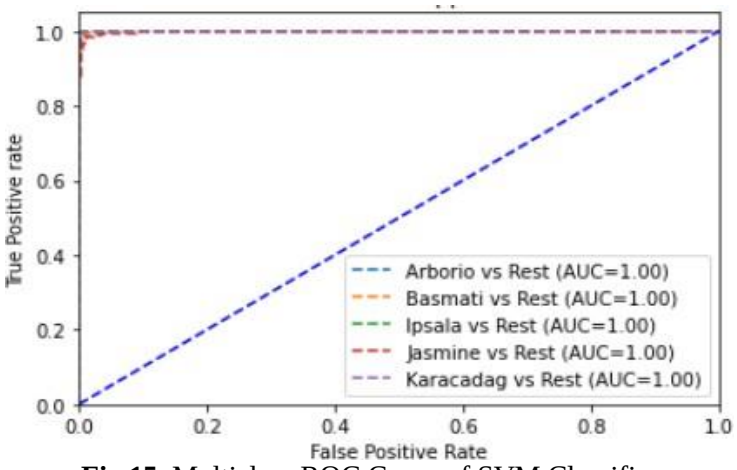


**Fig 15.** Multiclass ROC Curve of SVM Classifier

## 5. Conclusion

This study highlights the potential of using digital image analysis and machine learning to quantitatively characterize rice grains. By combining various descriptors, such as morphological, shape, and texture features, the study demonstrates that it is possible to classify different rice varieties with high accuracy. More- over, the techniques used in the study can be applied to commonly available low-cost imaging hardware. The classification task involved 100,000 samples from five different rice varieties. The results showed that decision tree classification achieved an accuracy of 99.04%, k-nearest neighbor achieved 99.24%, naive Bayes achieved 98.15%, and support vector machine achieved 98.54%. Interestingly, the study found that increasing the number of features positively contributed to higher classification accuracy.

## 6. Future Enhancement

This work could be extended to focus on identifying rice according to more precise guidelines. One possible approach to improve classification accuracy is to adopt ensemble methods such as boosting, bagging, and stacking. These methods involve combining the predictions of multiple models, which can lead to better accuracy compared to individual models. It may be possible to further improve the accuracy of classification of rice based on morphological, shape, and texture features using more sophisticated techniques.

## References

1. D.M. Hobson, R.M. Carter, Y .Yan. “Characterizations and Identification of Rice Grains through Digital Image Analysis”, IEEE- 2007.
2. R.Kiruthika, S.Muruganand, Azha Periasamy “Matching of Different Rice grains Using Digital Image processing”. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering Vol. 2, Issue 7, 2013.
3. Ilkay CINAR, Murat KOKLU. “Classification of Rice Varieties Using Artificial Intelligence Methods”. International Journal of Intelligent Systems and Applications in Engineering IJISAE, 2019, 7(3), 188–194.
4. Yadav, B. and V. Jindal, Monitoring milling quality of rice by image analysis. Computers and Electronics in Agriculture, 2001. 33(1): p. 19-33.
5. Dubey, B., et al., Potential of artificial neural networks in varietal identification using morphometry of wheat grains. Biosystems Engineering, 2006. 95(1): p. 61-67.
6. Zapotoczny, P., M. Zielinska, and Z. Nita, Application of image analysis for the varietal classification of barley:: Morphological features. Journal of Cereal Science, 2008. 48(1): p. 104-110.
7. Abirami S, Neelamegam P & Thanjavur India K H (2014). Analysis of Rice Granules using Image Processing and Neural Network Pattern Recognition Tool. DOI: 10.1.1.673.5557.
8. Harishkumar, K. S., Yogesh, K. M., & Gad, I. (2020). “Forecasting air pollution particulate matter (PM2. 5) using machine learning regression models”. Procedia Computer Science, 171, 2057-2066.
9. Sethy P K Chatterjee A (2018). Rice Variety Identification of Western Odisha Based on Geometrical and Texture Feature. Interna- tional Journal of Applied Engineering Research 13(4) : 35-39
10. Koklu M & Ozkan I A (2020). Multiclass classification of dry beans using computer vision and machine learning techniques. Computers Electronics in Agriculture 174: 105507. DOI: 10.1016/j.compag.2020.105507.
11. Ikegami 2020 Accessed: 14 May 2020]; Available from: https://www.ikegami.com.
12. Ilkay CINAR, Murat KOKLU. “Identification of Rice Varieties Using Machine Learning Algorithms” Journal of Agricultural Sci- ences

2022, 28 (2): 307 – 325.
13. Harish Kumar, K. S., & Gad, I. (2020). "Time series analysis for prediction of PM2. 5 using seasonal autoregressive integrated moving average (SARIMA) model on Taiwan air quality monitoring network data". Journal of Computational and Theoretical Nanoscience, 17(9-10), 3964-3969.
14. Pazoki A, Farokhi F Pazoki Z (2014). Classification of rice grain varieties using two Artificial Neural Networks (MLP and Neuro- Fuzzy). The Journal of Animal Plant Sciences 24(1): 336-343.
15. R M Carter, Y. Yan., "Digital imaging based classification and authentication of granular food products", Institute of Physics publishing, Measurement Science and Technology V.17, pp.235-240, 2005.
16. Harishkumar, K. S., Gad, I., & Yogesh, K. M. (2021, February). "Spatio-temporal clustering analysis for air pollution particulate matter (pm 2.5) using a deep learning model". In 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS) (pp. 529-535). IEEE.
17. Douik A and M. Abdellaoui (2010). "Cereal grain classification by optimal features and intelligent classifiers". International journal of communications and control, Vol-5(4), pp. 506-516.
18. Harishkumar, K. S. "Multidimensional Data Model for Air Pollution Data Analysis." In 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1684-1689. IEEE, 2018.