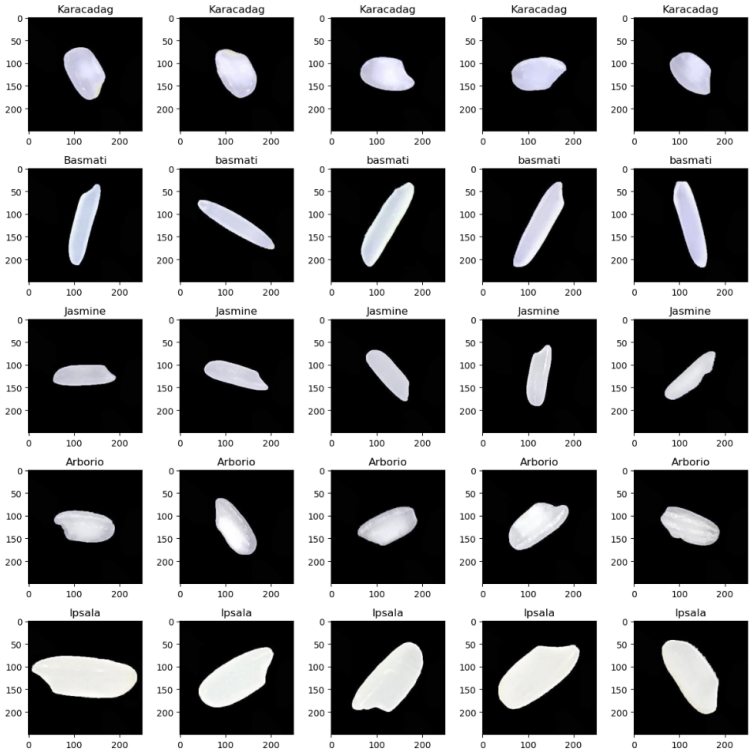
Machine Learning - Final Project
Rice Image Classification

אילן סיריסקי 207810458
אלדד צמח 209161447

Data Set Description [Link]

Datasets belonging to five rice varieties as Arborio, Basmati, Ipsala, Jasmine and Karacadag, which are often cultivated in Turkey, were used in the study.
The image dataset consists of 75,000 rice grain images, 15,000 from each variety. In RGB images contained in this dataset, the size of the image in which each grain of rice is located is 250 × 250 pixels.
We split the data into 80% training and 20% testing. For each class, we will have 12,000 images in the train set, and 3,000 images in the testing set.
For the CNN model, we had to reduce the size of the dataset set to a total of 25,000 images. For each class here, we will have 4,000 images in the train set, and 1,000 images in the testing set.



Our Question

1. Can we reach a high accuracy rating in classifying the different types of rice?
2. What are the two hardest rice variants to distinguish between them and Why?
3. Are there tools that can increase the success rate of classification?
4. Is there a certain structure in the rice variants that causes them to be similar or is there a hierarchy?
5. Which model has the best accuracy for the classification?
6. What were the features with the most impact?

Related Work

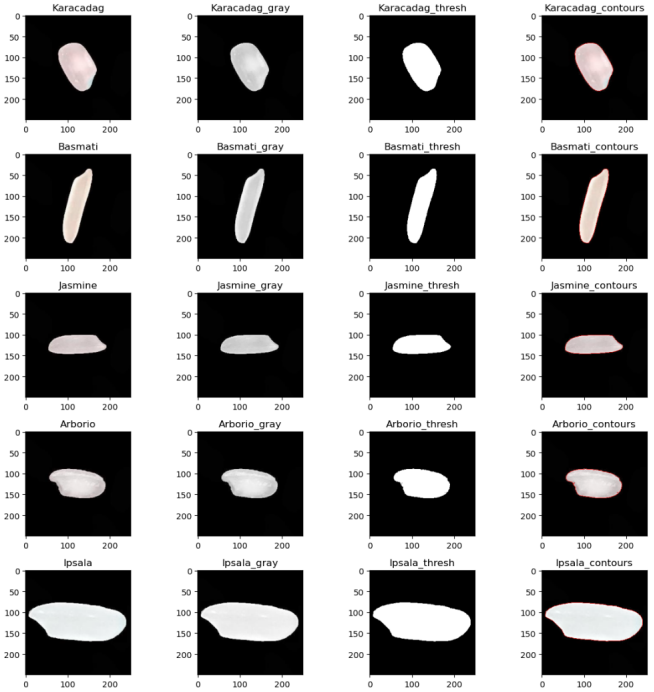
In recent years, many digital image features have been used to evaluate rice classification and quality. These include geometric parameters (length, perimeter, etc.), fracture rate, whiteness and determination of rice grain cracks can be given examples. Various features of grain products can be extracted by using systems based on image processing. Furthermore, these features are seen to be classified using algorithms such as ANN, SVM, LR, DNN and CNN from machine learning algorithms.

In a study in the literature, a two-class dataset containing 1700 rice data was carried out and 98.5% classification success was achieved using the SVM algorithm. In another study, 843 pieces of data were examined from sixteen classes and 87.16% accuracy was obtained using the SVM algorithm. In the study, which used three classes and 7399 pieces of data, a 95.5% success rate was achieved with the deep CNN algorithm. In another study conducted with three different types of rice and 200 pieces of data, the researchers used CNN for classification procedures after feature extraction and achieved 88.07% success.

Data Analysis - Image Processing

We did the image processing in 3 parts:

1. Convert to grayscale.
2. Threshold the image to get a binary map and trace contours of the rice grain.
3. Use contours to create features like length, size, shape etc.



We ended up having 29 different features, most of them are different kinds of Order Moments that are based on the contour in part 3.

For example, ‘m00’ is the zeroth order moment that represents the total mass or area of the region enclosed by the contour. It is calculated by summing the pixel intensities over the contour area. More on the feature explanation in the code notebook.

Our Approach

We will be using a total of 5 different models and they are: Logistic Regression, Linear SVM, Random Forest, MLP, CNN.

Project Difficulties

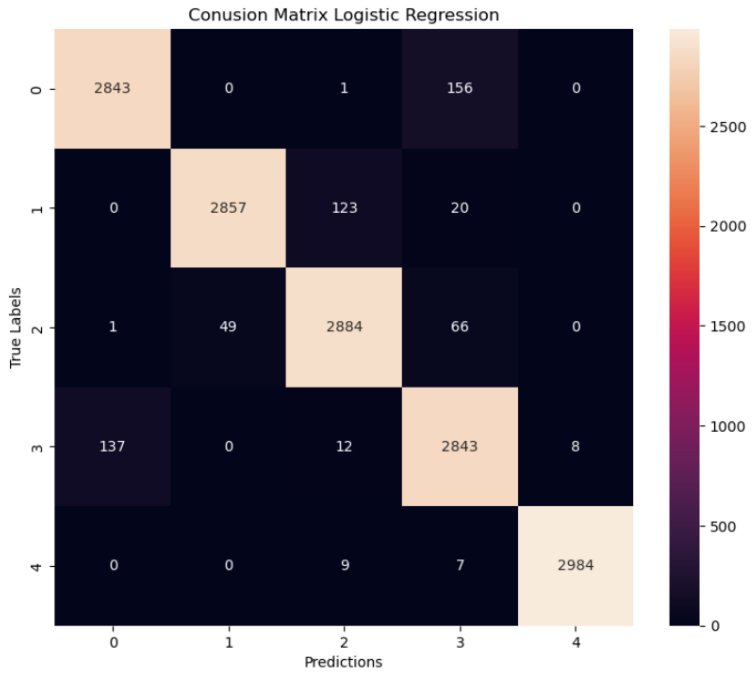
We expected problems with 2 kinds of rice, because both of them look almost identical, and we had trouble seeing the difference between them.
In the beginning we couldn't classify who is who between these 2 and we thought that the models couldn't classify them too. The 2 kinds are Arborio and Karacadag.
Karacadag on the left, Arborio on the right.

Results

Logistic Regression

----------Classification Report for LogisticRegression----------

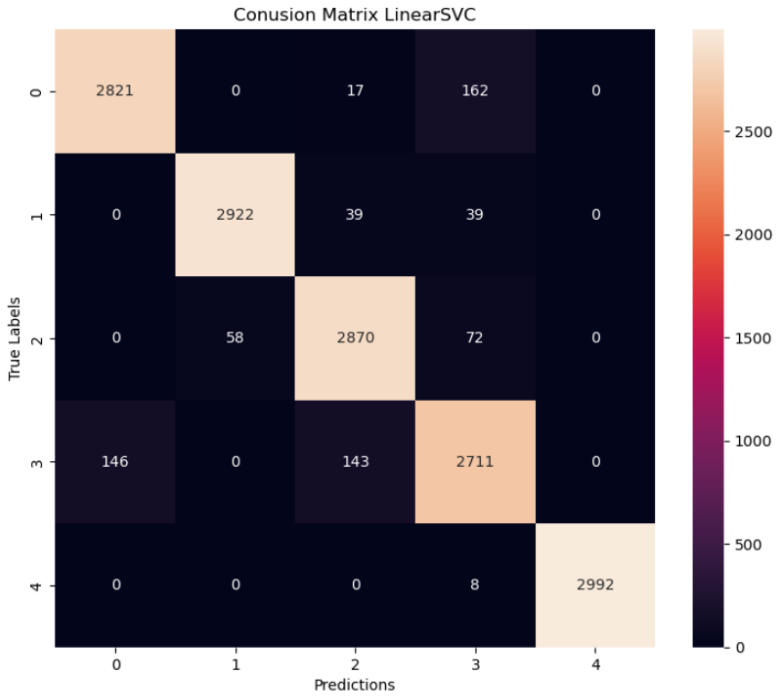
	precision	recall	f1-score	support
Karacadag	0.95	0.95	0.95	3000
Basmati	0.98	0.95	0.97	3000
Jasmine	0.95	0.96	0.96	3000
Arborio	0.92	0.95	0.93	3000
Ipsala	1.00	0.99	1.00	3000
accuracy			0.96	15000
macro avg	0.96	0.96	0.96	15000
weighted avg	0.96	0.96	0.96	15000



Linear SVC (sklearn SVM)

----------Classification Report for LinearSVC----------

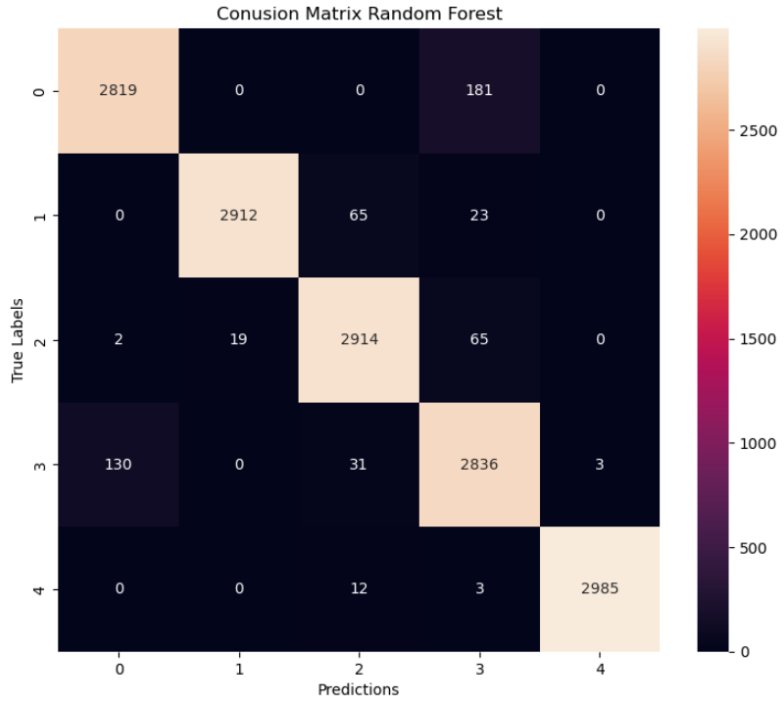
	precision	recall	f1-score	support
Karacadag	0.95	0.94	0.95	3000
Basmati	0.98	0.97	0.98	3000
Jasmine	0.94	0.96	0.95	3000
Arborio	0.91	0.90	0.90	3000
Ipsala	1.00	1.00	1.00	3000
accuracy			0.95	15000
macro avg	0.95	0.95	0.95	15000
weighted avg	0.95	0.95	0.95	15000



Random Forest

----------Classification Report for RandomForestClassifier----------

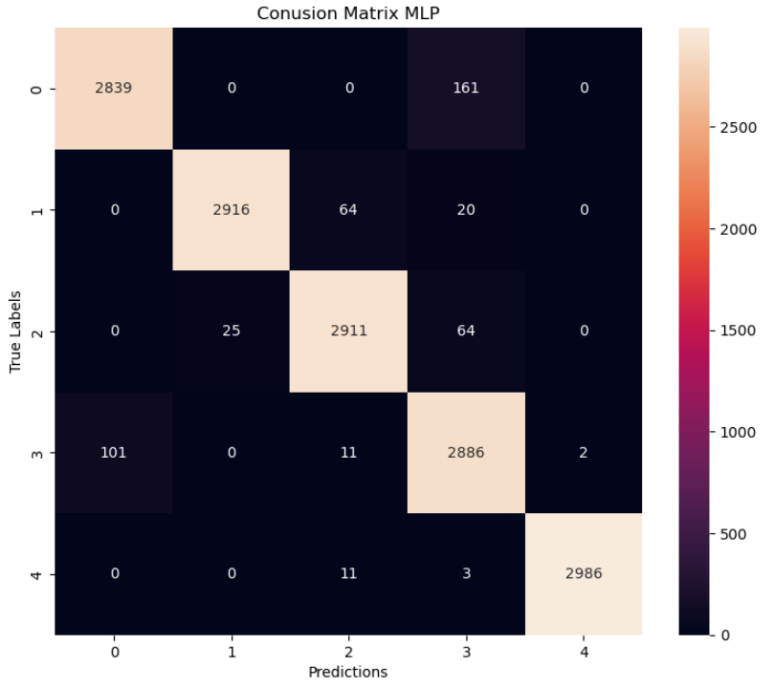
	precision	recall	f1-score	support
Karacadag	0.96	0.94	0.95	3000
Basmati	0.99	0.97	0.98	3000
Jasmine	0.96	0.97	0.97	3000
Arborio	0.91	0.95	0.93	3000
Ipsala	1.00	0.99	1.00	3000
accuracy			0.96	15000
macro avg	0.96	0.96	0.96	15000
weighted avg	0.96	0.96	0.96	15000



MLP

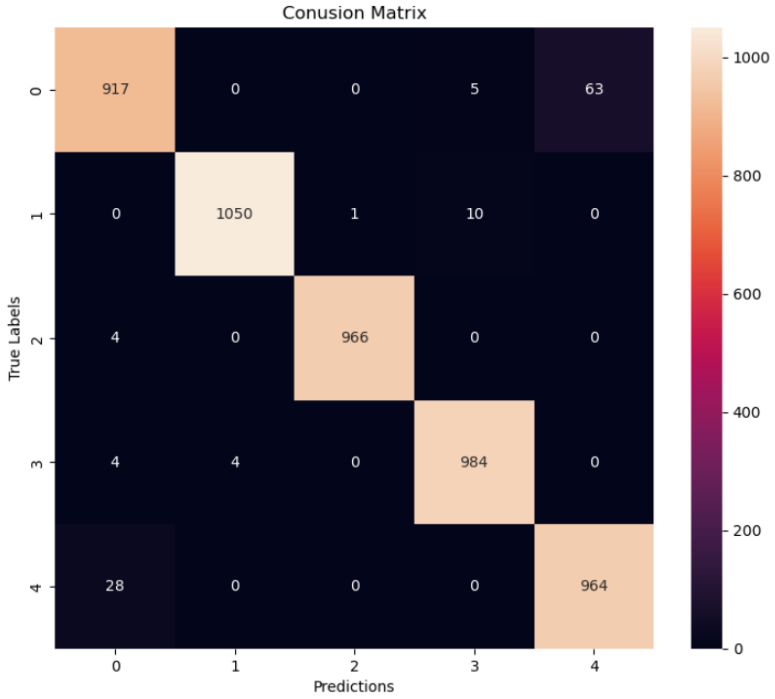
----------Classification Report for MLPClassifier----------

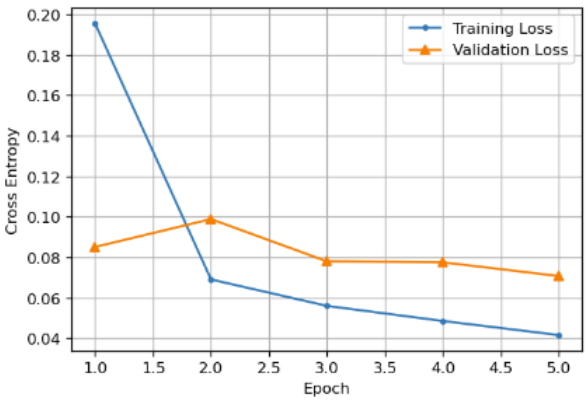
	precision	recall	f1-score	support
Karacadag	0.97	0.95	0.96	3000
Basmati	0.99	0.97	0.98	3000
Jasmine	0.97	0.97	0.97	3000
Arborio	0.92	0.96	0.94	3000
Ipsala	1.00	1.00	1.00	3000
accuracy			0.97	15000
macro avg	0.97	0.97	0.97	15000
weighted avg	0.97	0.97	0.97	15000

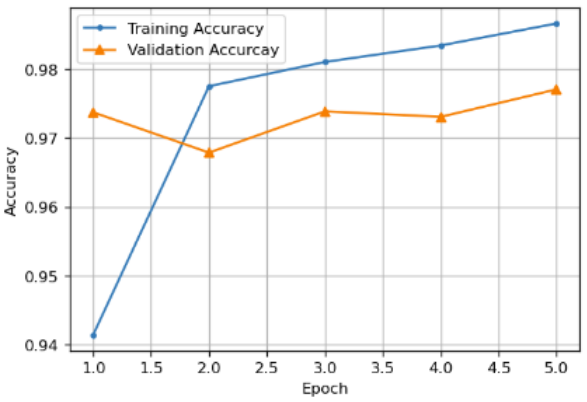


CNN

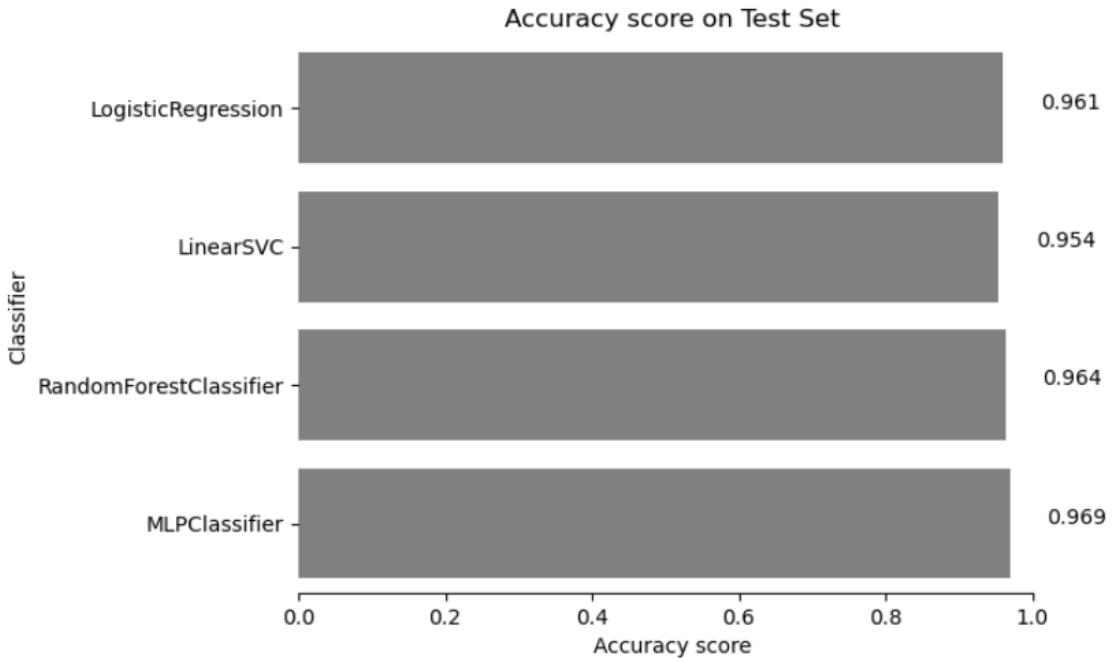
	precision	recall	f1-score	support
0	0.96	0.93	0.95	985
1	1.00	0.99	0.99	1061
2	1.00	1.00	1.00	970
3	0.98	0.99	0.99	992
4	0.94	0.97	0.95	992
accuracy			0.98	5000
macro avg	0.98	0.98	0.98	5000
weighted avg	0.98	0.98	0.98	5000

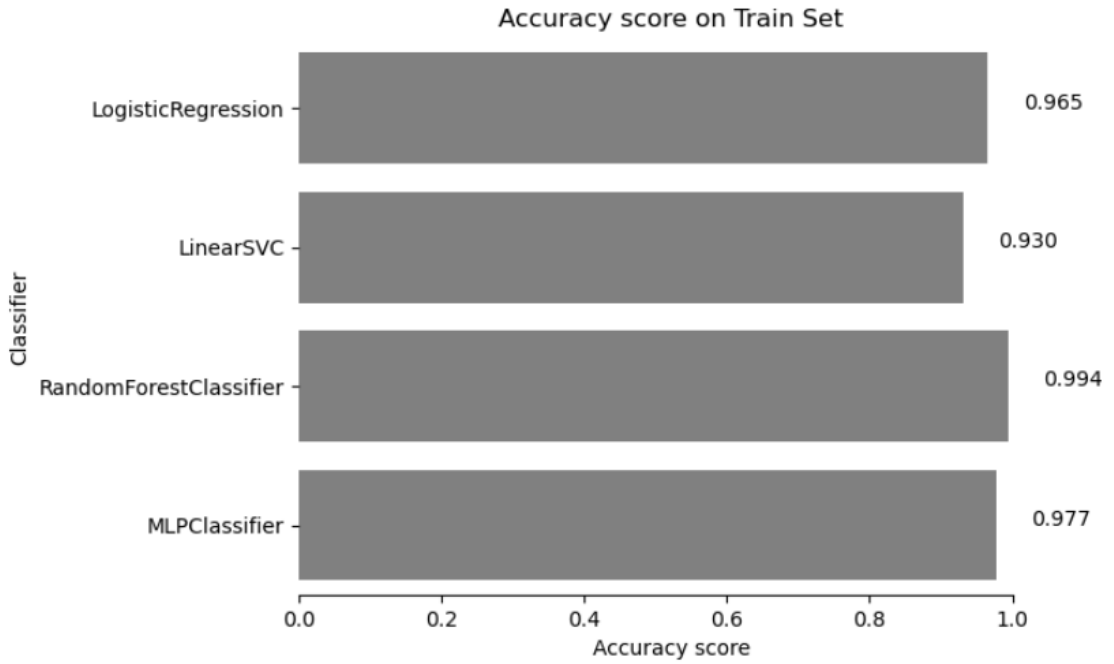






Result Analysis



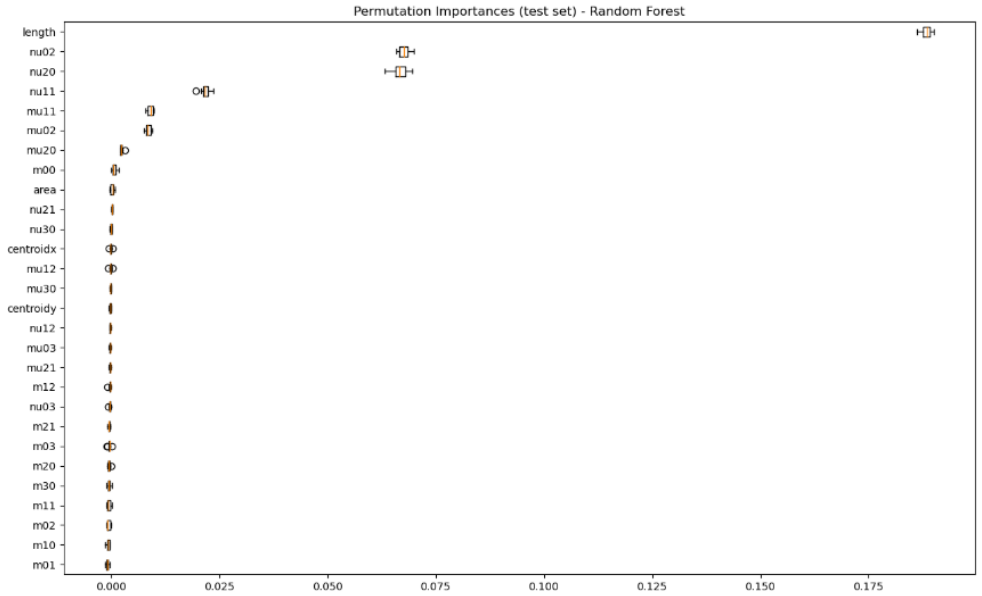


No CNN in those results because of smaller dataset.

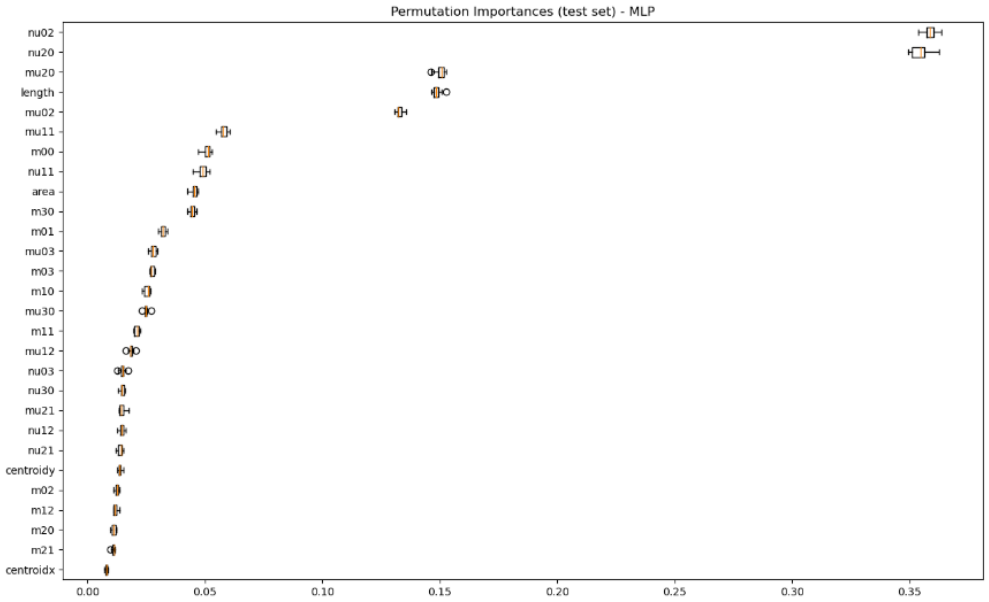
Feature Analysis

We check out the important features for our two best classifiers - Random Forest and MLP.

Random Forest



MLP



Answering Our Questions

1+5. For all of our ML models we reached more than 93% prediction accuracy. Where the SVM model was the lowest, while the Random Forest Classifier was the highest with 99.4% accuracy.
The MLP and the CNN models did quite well as well both around 98% accuracy.
Random Forest and MLP classifier perform best on both test and train set.
Logistic Regression Classifier is only slightly worse but perhaps can be improved through hyperparameter tuning.

2+4. The 2 types of rice that were the hardest to distinguish between them are "Arborio" and "Karacadag". The reason for that might be because they both have the same shape and size. The only difference might be the size of the lattice in their shape.
In some models there were also mistakes distinguishing between "Jasmine" and "Basmati" rice but a lot less than the 2 previously mentioned rice grains.

3+6. From the above plots, we can conclude that the second order invariant moments - "nu02" and "nu20" along with the length of the rice grain are significant for correct classification. Further, we can use the central and invariant moments to create more features like rotational invariants and try to improve the performance of our classifiers.

“nu02”, “nu20”: These features are scale-invariant central moments, calculated by normalizing the central moments by a combination of “m00” and other moments. They provide rotation and scale-invariant information about the shape.
“length”: This feature represents the perimeter or arc length of the contour.