National College of Ireland

# A Comparative study between Traditional Machine Learning and Deep Learning Models to classify Rice Types

MSc Research Project
MSc in Data Analytics

Zainab Maqsud Ghulam Hussain Mohamed Ali
Student ID: x21135614

School of Computing
National College of Ireland

Supervisor: Mohammed Hasanuzzaman

**National College of Ireland**
**Project Submission Sheet**
**School of Computing**

| | |
|---|---|
| **Student Name:** | Zainab Maqsud Ghulam Hussain Mohamed Ali |
| **Student ID:** | x21135614 |
| **Programme:** | MSc in Data Analytics |
| **Year:** | 2022 |
| **Module:** | MSc Research Project |
| **Supervisor:** | Mohammed Hasanuzzaman |
| **Submission Due Date:** | 15/12/2022 |
| **Project Title:** | A Comparative study between Traditional Machine Learning and Deep Learning Models to classify Rice Types |
| **Word Count:** | 6456 |
| **Page Count:** | 19 |

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

**ALL** internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

| | |
|---|---|
| **Signature:** | Zainab Maqsud Ghulam Hussain Mohamad Ali |
| **Date:** | 1st February 2023 |

## PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies). | ☐ |
| **Attach a Moodle submission receipt of the online project submission**, to each project (including multiple copies). | ☐ |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | ☐ |

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# A Comparative study between Traditional Machine Learning and Deep Learning Models to classify Rice Types

Zainab Maqsud Ghulam Hussain Mohamed Ali
x21135614

**Abstract**

Rice is one of the most commonly cultivated grains on the planet. When purchasing rice packets at the market, the physical appearance is the first thing that comes to mind for buyers. There are also rice varieties with numerous distinguishing traits. Typically, these attributes consist of shape, texture, and colour. On the basis of these distinguishing properties of the various types of rice, classification and identification of quality are possible. Machine learning models help facilitate the recognition of trends and patterns. Traditional machine learning models like Support Vector Machines (SVM), Random Forest and Decision Tree are compared with deep learning models like Convolutional Neural Network when it comes to classifying different types of rice (CNN). Common types of rice used in this experiment include Arborio, Basmati, Ipsala, Jasmine, and Karacadag. Over 75,000 images of rice grains are included in the dataset, with 15,000 depicting each kind. Images were preprocessed to prepare them for feature extraction. Values for quantifying feature set performance were produced using these models. Based on the results, we can see that the Random Forest model had the highest classification accuracy at 92.66%, followed by the SVM model at 89.33%, the Decision Tree model at 77.33%, and the CNN model at 50.66%. The classification increases with the inclusion of each new feature. On the basis of the obtained performance measurement results, it is feasible to conclude that the study was successful in identifying rice types.

Keywords —Rice Classification, Machine Leaning, Random Forest, Support Vector Machines, Decision Tree, Convolutional Neural Network, Image Processing, Deep Learning

## 1 Introduction

Rice is the most valuable grain product in the world in terms of its production value. It is abundant in carbs and starch. Moreover, it is of considerable importance in terms of nutrition in the rest of the globe, due to its nutritive and economical qualities, and it is frequently employed in the industrial sector. There are many quality standards for rice cultivars grown in Turkey. These include physical appearance, cooking qualities, flavour and fragrance characteristics, and efficiency Tipi et al. (2009). Manual rice grain classification methods can be costly and unreliable when attempting to determine physical characteristics from quality criteria. Because human judgements are subjective, inconsistent, and slow. Alternatives to automated methods that are non-destructive, economical, quick, and accurate include machine vision systems.

Applications of image processing and computer vision in farming are essential because they provide non-destructive analyses and are more cost-effective than human methods Mahajan et al. (2015). The advantages of image processing-based computer vision systems can be compared to the shortcomings of manual methods Barbedo (2016). Because the human element plays such a crucial role, manual grain evaluation or classification may take a significant amount of time and come at a high cost. When manual methods are used for the evaluation, the process may be different because the reliability of the results is dependent on the experience of the evaluation specialists. In addition, when analysis was carried over on a large scale, it may be difficult to make prompt decisions using manual approaches Patrício and Rieder (2018). Rice is a product that is manufactured in a number of places and is utilized all over the world. Rice's price on the market takes into account a variety of different aspects. These algorithms can manage voluminous amounts of data. Implementing these strategies will allow rice farmers to produce crops of a higher quality, while still adhering to food safety regulations, while doing so in a cost-effective, automated and non-destructive manner. In recent years, a wide variety of computerized feature representations have been utilized to evaluate the categorization and purity of rice. Rice grain splits, break rate, whiteness, and measures (such as perimeter, length, etc.) are some examples of measurable characteristics. Methods based on image processing enable the extraction of several features from grain products. In addition, it has been demonstrated that these characteristics can be categorized utilizing machine learning and deep learning techniques such as Convolutional Neural Network (CNN) and Support Vector Machines (SVM). This was proved by a number of studies. The SVM technique was applied in a study that is described in the literature, and it was successful in classifying 1700 pieces of rice data into two categories with an accuracy rate of 98.5% Sun et al. (2014). In a separate investigation, the SVM approach was utilized to analyze 200 data points covering sixteen categories, with an accuracy of 87.16% Liu et al. (2016). In the research that consisted of three different categories and 7399 data points, the deep CNN approach had a success rate of 95.5% Lin et al. (2018). In a second investigation with 200 data points and three various types of rice and using Convolutional Neural Network for classification methods and feature extraction, the researcher attained an accuracy of 88.07% Ahmed et al. (2021).

The objective of this research is to create a non-destructive system that enhances the accuracy of classification by making use of photographs of a number of different types of rice. Rice classification using machine learning models offers several benefits in the agriculture industry. Firstly, it enables automated and efficient grain sorting, which can save time and resources. Secondly, it improves quality control and inspection by reducing human error and increasing accuracy. Thirdly, the technology provides predictive analysis for optimized crop management and yield, which is crucial for farmers to make informed decisions and improve their crop yields. These benefits make rice classification an essential tool for farmers and the agriculture industry. On the other hand, manual methods for rice classification have several limitations that make them less efficient and effective compared to other methods. One of the biggest limitations is that they are time-consuming and labor-intensive, requiring a significant amount of effort and manpower to classify even a small batch of rice. Additionally, human error can often come into play, leading to inaccurate results and inconsistencies in the classification process. The lack of standardization and subjectivity in manual methods also contributes to their limited accuracy and consistency. Furthermore, manual methods are not well-equipped to handle large volumes of rice efficiently, and it can become difficult to identify subtle differences

in rice quality, which can affect the final classification. These limitations highlight the need for more advanced and automated methods for rice classification. The objective of comparing rice classification using machine learning and deep learning models is to evaluate the effectiveness and accuracy of the models in identifying and classifying different types of rice. This comparison aims to determine which model is better suited for rice classification and which model provides more accurate results.

### 1.1 Research Question

RQ – How can traditional machine learning(Random forest, decision tree, support vector machine) and deep learning(Convolutional neural network) models help in classifying different types of rice(Basmati, Ipsala, Arborio, Jasmine, and Karacadag) and which is better at doing so?

## 2 Related Work

This section examines the existing literature in the selected field. We define the factors and ideas that classify any type of grains. Using Deep learning and Machine learning approaches as the subparts, many researchers have investigated to achieve higher accuracy and precision.

### 2.1 Deep Learning and Image Processing Approach

Kaya and Saritas (2019) developed an Artificial Neural Network and image processing techniques in a dynamic environment to differentiate vitreous durum wheat kernels, which suggest high results. To enable dynamic categorization, video frames from each category or a collection of classes were placed on a spinning conveyor belt. Due to the lack of a classification of wheat kernels in real-time, the study had severe limitations. This problem can be solved by selecting a reference zone and enumerating the number of items within it.

Using both image data and automatic identification, a number of research on seed detection were carried out by Granitto et al. (2005). Their database contains the unique species of 236 different types of weeds. ANN (Artificial Neural Network) classifier was able to successfully classify 92.54% of their test image by making use of 12 factors, which included morphological, color, and texture data. Included among the 12 morphological criteria were measurements of seed uniformity, size, and ratios for plane mass distribution. As a criterion for selecting features, the accuracy of Naive Bayes was used, and it was discovered that they were nearly true for the data set. In addition, they stated that morphological characteristics have the greatest capacity for discrimination, that color is not a legitimate categorization set of criteria because the majority of the species were light to dark brown or black, and that texture characteristics are even less reliable than morphological characteristics. All of these claims were made in light of the fact that the majority of species were brown or black.

Chatnuntawech et al. (2018) developed the relationship between spatiospectral deep learning and hyperspectral imaging where they propose a non-destructive method for classifying rice varieties. According to the author, hyperspectral imaging also revealed spectral data on the rice seeds in addition to spatial information. CNN is used to extract

features that is spatio-spectral properties from every rice seed without the use of data processing procedures. They showed that the proposed approach (ResNetB) recorded highest classification accuracies compared to the most widely used classification methods that use SVM using two different types of rice datasets. For the paddy rice dataset, which included the rice types that are frequently mistaken with the others, the suggested technique outperformed SVM with the both spectral and spacial information, achieving a mean classification accuracy of 91.09%as opposed to SVM's 79.23%.

The results of the study demonstrated by Kezhu et al. (2014) shows a neural network classification to detect soybean illnesses. The accuracy rate was 100% for soy seeds with no defects, and 93.75% for those with defects. Both the accuracy of the classifier and the performance of the classifier for damaged and worm-eaten soybean were 92%. A satisfactory detection outcome was achieved for heterogeneous seeds, with an accuracy performance of 90%. This study demonstrates that the BP neural network may provide effective technical assistance for finding a single flaw in soybean. However, it is conceivable for multiple defects to coexist on a particular soybean seed, making a comprehensive examination of soybean quality very practical and helpful.

In the study conducted by Yi et al. (2014) shows that morphologically same seeds can be distinguished with an accuracy of 97% using seed image analysis, and that automated approaches can distinguish seeds with an accuracy of 90%, which is higher than that of qualified experts. It was proven that qualities once thought to be "ideal," in which the importance of texture was ignored, are not always the case; in fact, texture turned out to be a key differentiator in their tests for the species that they used.

Sethy et al. (2020) analyzed the effectiveness of 13 different number of CNN models using a deep feature and transfer learning plus the SVM approach. The deep features of pooling layers of vgg16, AlexNet and vgg19, were retrieved at the first. These layers were fc8, fc7, and fc6. The SVM was utilized to make classification decisions based on these features. When compared to the characteristics of fc8 and fc7, the feature of fc6 made a significantly larger impact to the classification. Therefore, in order to choose the most appropriate classification model, just AlexNet's fc6 layer, as well as vgg16 and vgg19, was taken into consideration. Among the deep feature approaches, the resnet50 plus the SVM classification model was found to be the best option, with an F1 score of 98.38% and a training time of 69 seconds. The research could perform better with more fine tuned CNN model in the hopes of achieving greater results.

## 2.2 Machine Learning Techniques

Azmi et al. (2021) developed machine learning techniques to determine the RF-based moisture content of rice. The processed data is input to various ANNs models, including K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Random Forest, and Multi-layer Perceptron (MLP). In comparison to the other four models, the Random Forest technique with one input feature (RSSI WSN) has the highest accuracy of 87%. Random Forest is a robust model in this instance since it can accurately predict the level of moisture in rice with only one input feature.

Ibrahim et al. (2019) employed a multi-class support vector machine (SVM) for the

purpose of rice grain classification. The performance of this study is examined using 90 test images that produced a classification accuracy of 92.22%. This study can help to assist the agricultural technology industry in the future with the automatic classification of rice grain.

Using SRC techniques and image data, Kuo et al. (2016) were able to identify 30 distinct varieties of rice grains without harming them. The different kinds of paddy were biologically and visually distinct. This resulted in the creation of image-based approaches for recognizing rice grains. Under a microscope, scientists produced images of the rice grains with a resolution of approximately 95 pixels per millimeter using the suggested method. The high resolution made it possible to see even the smallest details of the rice grains. After the textural, physical and visual traits of the grains have been measured, an SRC classifier is made to use the traits as inputs and foresee the types of grains. The classification algorithm was 89.1% accurate as a whole.

Han et al. (2015) implements a new computer vision based approach to classify crop diseases, including 1) a new technique for identifying features using a combination of super pixels as well as marker controlled watershed segmentation techniques for ignoring over segmentation, minimizing the difficulty to follow image processing tasks, and enhancing the quality of the results; and 2) a new method for calculating and investigating features based on textural, gradational, and spectral characteristics. In addition, they compared classification techniques based on SVM and ANN. The SVM-based technique is quite effective at disease identification and severity estimation. The early results of this effort are promising.

Mohapatra et al. (2021) developed a Naive Bayes classifier as an approach for classifying the quality of rice. They use the Naive Bayes classifier method with the electronic nose dataset to make rice quality predictions. Complement achieves the highest level of accuracy, 98%, compared to the other methods. The remaining methods, which include Gaussian, Bernoulli, and Multinomial, display comparably lower levels of precision. The ROC value that they obtained was high, and it was 1.00, which indicated that their method performed perfectly.

# 3 Methodology

Literature Review done shows classification of various types of grains done with Machine Learning and Deep Learning. We compare both the techniques to check which model performs best. Machine learning has an edge over traditional techniques due to the large amount of data that can be quickly searched through to analyze patterns. The main goal is to construct a model by using Deep Learning and Machine Learning techniques to an image dataset. This will allow the model to easily classify different types of rice, which will help the FMCG (Fast Moving Consumer Goods) companies to identify rice types in less time and accurately without the need of manpower.

The proposed technique should not produce incorrect results for consumers in the actual world. Therefore, the model's precision should be good. In order to attain greater

classification results during testing, the trained model must have a high degree of accuracy. False positive rates and false negative rates must be considered while picking a Machine Learning or deep learning algorithm for the classification of rice. Because if the percentages of false negatives and false positives are large, the rice categories will be erroneous. The developed approach must be capable of producing classifiers in real time, requiring fewer computer resources and a very quick execution time.

## 3.1 Dataset

We used the 'Rice Image Dataset' from an open source website 'Kaggle'. The Dataset contains five types of rice that is often grown in many parts of the world, Basmati, Ipsala, Karacadag, Arborio and Jasmine. The dataset consists of a total number of 75,000 rice grain images, with 15,000 from each of these different kinds of grains. For the image dataset, models are made using Random Forest, SVM, Decision Trees, and CNN (Convolutional Neural Network), and classification is done. Using the confusion matrix values of the models, statistical results of accuracy, sensitivity, specificity, false positive rate, and false negative rate were calculated, and the results of each model are given.

## 3.2 Model Training using Machine Learning Algorithms

The evaluation the performance of various supervised machine learning algorithms on a training dataset through training and testing. The algorithms selected were based on their effectiveness in handling classification problems. The dataset was randomly split into training and testing parts, which is considered the optimal scenario.

- Support Vector Machines (SVM)

Support Vector Machine (SVM) is an approach to machine learning based on supervised learning that can be used for regression and classification. The objective of SVM is to construct a classification prediction model based on the supplied attributes of the original value of the testing data. The SVM approach can utilize many kernels, with polynomial, linear, sigmoid and RBF being the most prevalent Sanjaa and Chuluun (2013). In SVM, a hyperplane separates the categories, which in this case are the five kinds of rice. After training with the dataset, the classifier successfully classify types of rice.

- Random Forest

Random Forest is an integrated learning method that is also referred as bootstrap aggregating. It may be utilized for regression, classification and other tasks Huang and Zhang (2019). It is a fast method in comparison to other algorithm for supervised learning since it needs very little training time than those other algorithms do. It possesses a high accuracy for use with large datasets. Random Forest has excellent scalability and parallel processing for high-dimensional data classification, and it is highly resistant to outliers and noise while also avoiding overfitting. Random Forest also has a high level of robustness to outliers and noise. Classifiers known as random forests combine the results of several different tree predictors, with the results of each tree being determined by the values of a vector that is selected at random. In addition to this, the trees in the forest are all laid out in the same pattern. While building a tree, we make the assumption that

'n' refers to the number of training observations and 'p' refers to the number of variables that serve as features in a training set. We choose k greater than p as the number of factors that need to be considered while making a choice at a tree's decision node. During the testing phase, we will select a bootstrap sample out from 'n' observed values in the training set. Then, we will utilize the rest data to calculate the error caused by the tree. Therefore, for each node in the tree, we make a decision from among arbitrary 'k' variables and then compute the best split based on the 'k' variables that are present in the training set. Unlike other tree algorithms, trees are never cut and grow continuously. Random Forests are able to manage a large number of variables inside a data set. During the building process, they also generate an internal neutral estimation of the generalization error. In addition, they are proficient at predicting missing data. The lack of repeatability is a major drawback of Random Forests, as the creation of the forest is arbitrary. In addition, it is difficult to comprehend the final model and its consequences because it is composed of multiple distinct decision trees.

• Decision Tree

A decision tree is a non-parametric supervised learning technique used for classification and regression tasks. It has a tree-like, hierarchical structure with branches, root nodes, internal nodes, and leaf nodes. The objective of using a Decision Tree is to construct a training model that can be employed to figure out the class of the attribute value by training basic decision rules generated from historical data. When utilizing Decision Trees to determine a class label for only a record, predictions are made beginning at the tree's root. We do a comparison between the values of the root attribute and the attribute of the record. Based on the comparison, we proceed along the branch that corresponds to that value and then proceed to the subsequent node.

• Convolutional Neural Network (CNN)

CNN is a technique for deep learning that is often used in applications such as voice recognition, natural language processing, image processing, and large data sets. The convolution layer creates vast quantities of data while simultaneously reducing complexity. The model proposed follows sequential pattern that means all the steps will be done one after another. The image is kept, shrunk, then transferred to the subsequent layer. In order to prevent altering the classification, it is also essential to make the most appropriate modifications to this layer. The activation layer, which allows data to be brought into specified ranges, is applied after the other layers. Following these operations, features are reduced to the level of the neural network, and learning operations are performed to derive conclusions in the connected layer serving as the classification layer.

• Model Evaluation

The performance of the model will be evaluated depending on the specified parameters. The confusion matrix will be used to evaluate the efficiency of the model. The matrix consists of true, false, positive, and negative values which are called as true positive, true negative, false positive, and false negative.

1. True Negative (TN)

It is said to have True Negative values when the value that was predicted to be negative in this case actually turns out to be true.

2. True Positive (TP)
It is referred to be True Positive values when the value that was predicted to be positive in this context actually turns out to be true.

3. False Negative (FN)
Negative values that turn out to be false are those in which the value that was predicted to be negative actually ends up being true.

4. False Positive (FP)
False positive values are those that occur when the expected value, which in this case is positive, actually turns out to be false.

Based on the values that we presently have, we are now able to calculate a few evaluation metrics to use , such as accuracy, specificity, sensitivity.

5. Accuracy
To calculate accuracy, we divide the total number of right predictions by the total number of observations in the data set. Accuracy is expressed as a percentage. The accuracy can range from 1.0 to 0.0, with 1.0 being the highest.

Accuracy = (TP+TN)/(TP+TN+FP+FN)

6. Specificity
The level of specificity (SP) can be calculated by the number of negative predictions that were right divide by the total number of negatives. In some cases, it is also referred to as the True Negative Rate (TNR). The specificity of 1.0 is the highest possible, while 0.0 is the least specific.

Specificity = TN/(TN+FP)

7. Sensitivity
You can figure out a test's sensitivity (SN) by dividing the total number of positive results by the number of accurate positive predictions. It can also be called the Recall rate (REC) or the True Positive rate, depending on the situation (TPR). The scale for sensitivity goes from 1 to 0.

Sensitivity = TP/(TP+FN)

# 4 Design Specification and Implementation

The approach uses a Python-based implementation. The dataset includes 15000 pictures of rice falling in all of the five categories as Arborio, Basmati, Ipsala, Jasmine and Karacadag. The data is collected and a list of images are generated based on each cat-

egories. The data is split 80-20 ratio. As we have image data in separate folders for each rice class, we have implemented a function. The function runs through each folder and list of image files in each folder. Then it moves the training images in train folder and test images in test folder.

The pre-processing of image for feature extraction is done by image segmentation. Image segmentation is a method in which a digital image is broken down into various subgroups called Image segments which helps in reducing the complexity of the image to make further processing or analysis of the image simpler. Segmentation in easy words is assigning labels to pixels. For this we have implemented a function that takes the image then takes the grayscale of the image, find the adaptive threshold, contours and region of interest (ROI). Adaptive thresholding takes the grayscale image as input and outputs the binary image representing the segmentation. For each pixel in the image, a threshold is calculated. Contours are the line joining all the points along the boundary of an image that are having the same intensity. These are identified using the thresholds. ROI Segmentation is selecting a specific region in the frame and providing it's dimensions in the rectangle method so that it will draw the rectangle-shaped ROI on the frame.

Below Figure 1 shows a graph of total images in the list for each category.

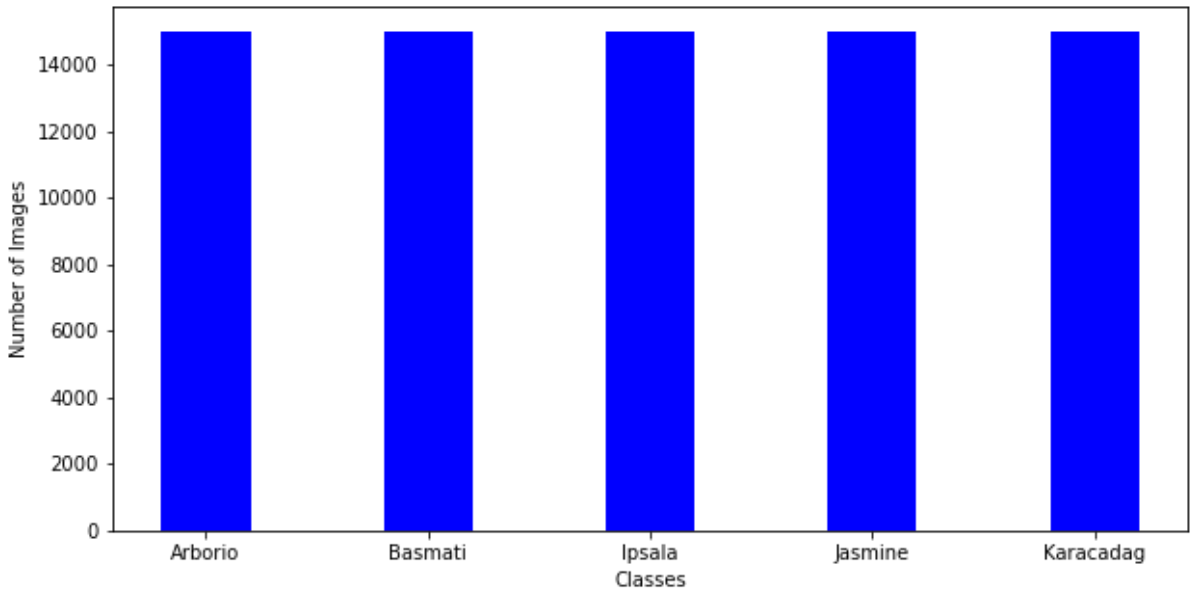


Figure 1: Total images in the list of each category

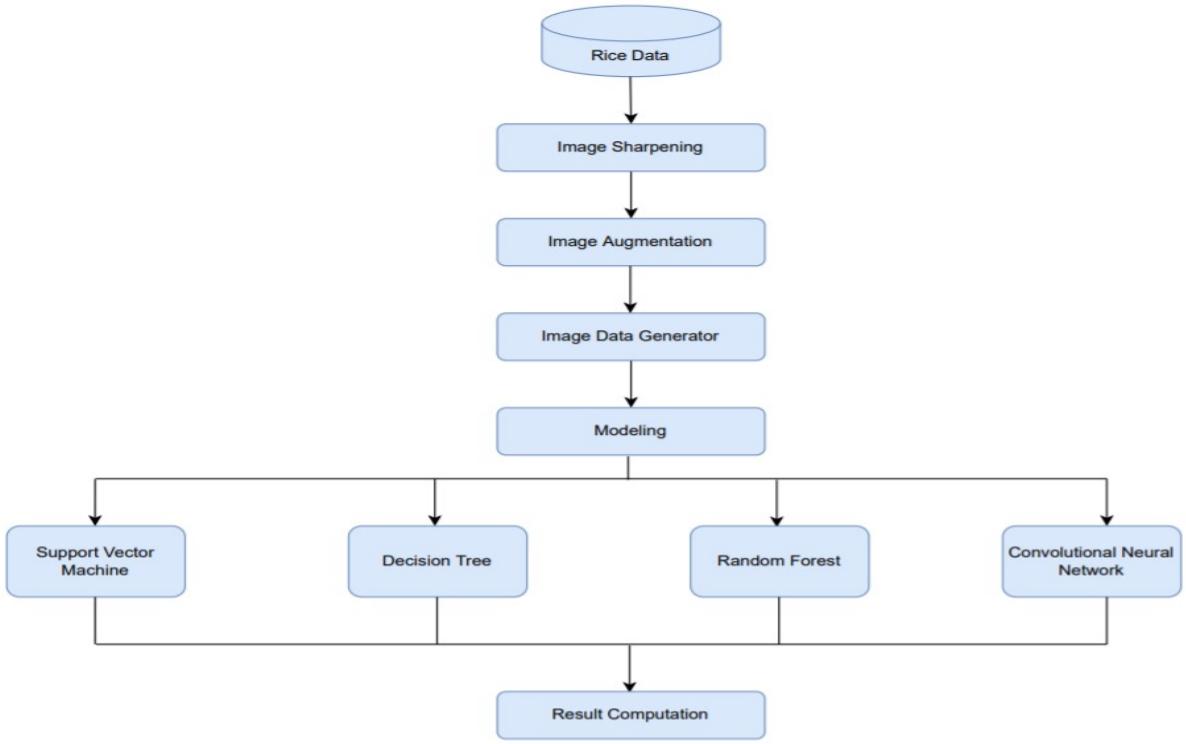


Figure 2: Implementation using Machine Learning Models and Neural Network

All the python libraries that were required to implement in the project were imported. The data is collected using glob function and a list of images are generated based on each categories. The data is then passed for Image Sharpening and Data Exploration where we load the images and visualize the function. We check the image in grey scale since the images is read in RGB format so in grey scale it will read in (0,1) , then we check the adaptive threshold of images and contours in images to avoid overfitting. The process of image sharpening helps to enhance the contrast and features of an image. This is done by focusing the image's edges and increasing the image's overall sharpness. After the image is augmented, we select the number of images for training and testing the data. The train_test_split function splits 180 images to 149 training and 30 for testing data. The dataset was evaluated using SVM, Random Forest, Decision tree and CNN methods to compare the best performing model. The hyper parameters used for SVM are gamma, c and kernel where gamma and c are the penalties that model gets to improve the performance and a linear kernel was used because it was giving best performance for SVM. Decision tree and Random forest worked best with the default hyper parameters. Since CNN uses two-dimensional data, no reshaping was necessary. In CNN, we employ the Sequential model, which is a form of model in which the layers are added sequentially. This is the most basic sort of model and it is utilized when the data flow is linear. The first layer is a conv2D layer with a kernel size of 64, generating 64 images. The Activation parameter has been set to 'relu' due to its better performance. After all layers the last dense layer will produce output as [0,1,0,0,0] that signifies that the 2nd class is true. Finally the result is computed in a table to check the model accuracy of each model.

Table 1 below lists the hyperparameters that have been chosen for the implemented models.

Table 1: List of hyperparameters used to implement models.

| **Model** | **Hyperparameter** | **Values** |
|---|---|---|
| SVM | Gamma | Gamma=$1/(\mathrm{n}_features * variance(X))$ |
| | C | C=20 |
| | Kernel | Kernel=Linear |
| Decision Tree | Criterion for tree splitting | Criterion = gini |
| | Splitting Strategy | Strategy = best |
| | Max depth | Max depth = None |
| | $\mathrm{Max}_features$ | Max features = sqrt($\mathrm{n}_features$) |
| Random Forest | No. of estimators | No. of estimators = 100 |
| | Criterion | Criterion = gini |
| | $\mathrm{Max}_depth$ | $\mathrm{Max}_depth = None$ |
| | $\mathrm{Max}_features$ | $\mathrm{Max}_features = sqrt(n_features)$ |
| CNN (Conv2D layer) | Kernel Size | Kernel Size = 64 |
| | Strides | Strides = 5 |
| | Filters | Filters = 5 |

# 5 Evaluation

This research compares the performance of SVM, Random Forest, Decision Tree, and CNN on the rice image dataset. These algorithms were put through a series of test samples using a variety of performance metrics. The results of those tests are given and graphed in this section. Here the classes are named as 'Arborio': 0, 'Basmati': 1, 'Ipsala': 2, 'Jasmine': 3 and 'Karacadag': 4.

## 5.1 Support Vector Machines (SVM) Confusion Matrix

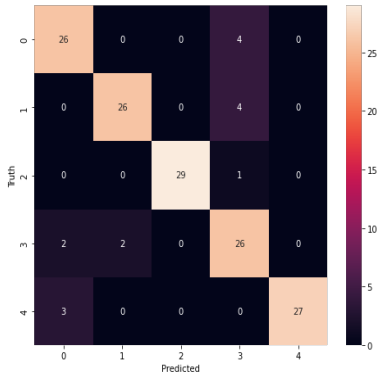


Figure 3: The Confusion Matrix for SVM

From the confusion matrix we can see that class 2 that is Ipsala rice images were classified better than other classes with maximum of 29 true positives and only 1 false negative.

```
Sensitivity :  1.0
Specificity :  1.0
```

```
print(classification_report(y_test, y_predicted))
```

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.84 | 0.87 | 0.85 | 30 |
| 1 | 0.93 | 0.87 | 0.90 | 30 |
| 2 | 1.00 | 0.97 | 0.98 | 30 |
| 3 | 0.74 | 0.87 | 0.80 | 30 |
| 4 | 1.00 | 0.90 | 0.95 | 30 |
| accuracy | | | 0.89 | 150 |
| macro avg | 0.90 | 0.89 | 0.90 | 150 |
| weighted avg | 0.90 | 0.89 | 0.90 | 150 |

Figure 4: The Classification report for SVM

## 5.2 Random Forest Confusion Matrix

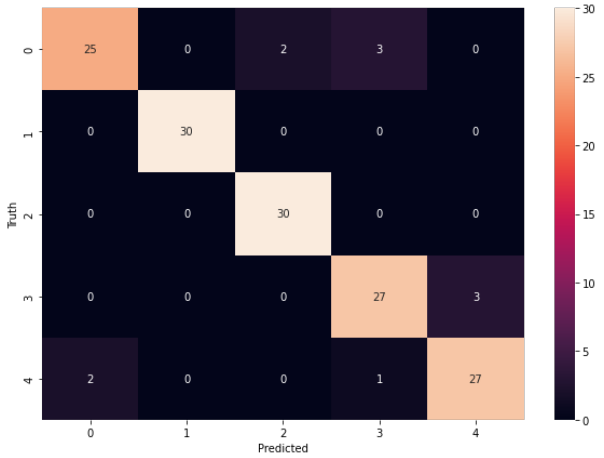


Figure 5: The Confusion Matrix for Random Forest

From Random Forest confusion matrix we can see that classes 1 and 2 were classified perfectly than other classes with maximum of 30 true positives and 0 false negatives.

```
Sensitivity :  1.0
Specificity :  1.0
```

```
print(classification_report(y_test, y_predicted))
```

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.93 | 0.83 | 0.88 | 30 |
| 1 | 1.00 | 1.00 | 1.00 | 30 |
| 2 | 0.94 | 1.00 | 0.97 | 30 |
| 3 | 0.87 | 0.90 | 0.89 | 30 |
| 4 | 0.90 | 0.90 | 0.90 | 30 |
| accuracy | | | 0.93 | 150 |
| macro avg | 0.93 | 0.93 | 0.93 | 150 |
| weighted avg | 0.93 | 0.93 | 0.93 | 150 |

Figure 6: Classification Report for Random Forest

## 5.3 Decision Tree Confusion Matrix

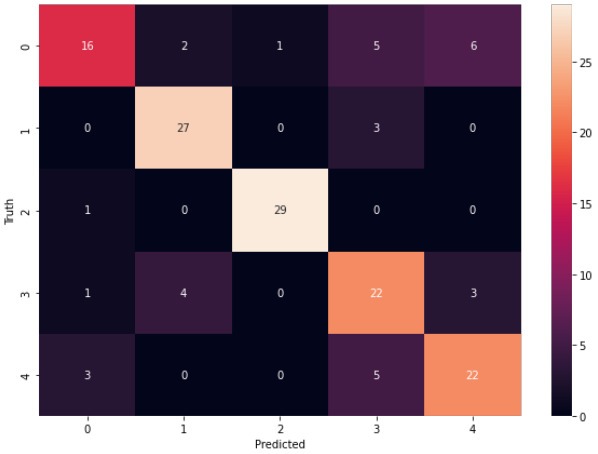


Figure 7: The Confusion Matrix for Decision Tree

Based on the Decision Tree confusion matrix, class 2 is identified more accurately than other classes, with a maximum of 29 true positives and one false negative.

```
Sensitivity :  0.8888888888888888
Specificity :  1.0
```

```
print(classification_report(y_test, y_predicted))
```

```
              precision    recall  f1-score   support

           0       0.76      0.53      0.63        30
           1       0.82      0.90      0.86        30
           2       0.97      0.97      0.97        30
           3       0.63      0.73      0.68        30
           4       0.71      0.73      0.72        30

    accuracy                           0.77       150
   macro avg       0.78      0.77      0.77       150
weighted avg       0.78      0.77      0.77       150
```

Figure 8: Classification Report for Decision Tree

## 5.4 Convolutional Neural Network (CNN) Accuracy score

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d (Conv2D)             (None, 20, 20, 5)         61445

 max_pooling2d (MaxPooling2D  (None, 10, 10, 5)        0
 )

 dropout (Dropout)           (None, 10, 10, 5)         0

 flatten (Flatten)           (None, 500)               0

 dense (Dense)               (None, 128)               64128

 dense_1 (Dense)             (None, 5)                 645

=================================================================
Total params: 126,218
Trainable params: 126,218
Non-trainable params: 0
_________________________________________________________________
```

Figure 9: Implementation of CNN

```
loss,accuracy= model.evaluate(X_test, y_test, verbose=0)
accuracy*100
```

```
50.66666603088379
```

```
print(classification_report(y_test, y_predicted))
```

```
              precision    recall  f1-score   support

           0       0.93      0.83      0.88        30
           1       1.00      1.00      1.00        30
           2       0.94      1.00      0.97        30
           3       0.87      0.90      0.89        30
           4       0.90      0.90      0.90        30

    accuracy                           0.93       150
   macro avg       0.93      0.93      0.93       150
weighted avg       0.93      0.93      0.93       150
```

Figure 10: Accuracy and Classification Report of CNN

## 5.5 Model Accuracy Scores

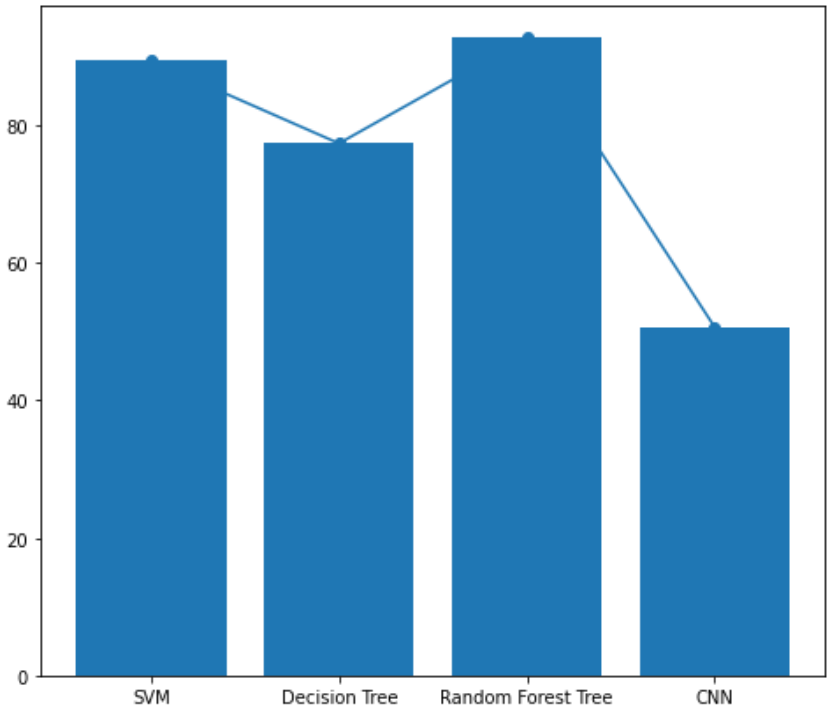


Figure 11: Plot for Model Accuracy Scores

It can be seen rather clearly from the graph that random forest has the best level of model accuracy.

## 5.6 Discussion

Based on the plot of model accuracy scores, it is evidently clear that the Random Forest machine learning algorithm is more accurate than any of the other algorithms that were considered for the classification of rice types. Below Figure 12 shows the performance of each model. When evaluating the performance of a classifier, which is sometimes referred to as its prediction accuracy, it is common practice to place a greater priority on false positives and false negatives. In the actual world, false positives are more detrimental than false negatives. Because we don't want consumers to be able to identify wrong type of rice grains, false positives are taken into account while finding the optimal classifier. According to the data presented in the table, the Random Forest algorithm achieved the highest scores across all of the evaluation metrics that were carried out. The Random Forest method has the highest rates of accuracy, sensitivity, and specificity, as well as the lowest false positive rates (FPR). As a result of this, Random Forest is the method that is most accurate for classifying types of rice. The features and characteristics of the rice classification dataset may have contributed to the better performance of Random Forest in comparison to other machine learning and deep learning models. Rice classification involves several high-dimensional features such as grain shape, size, color, and texture, and Random Forest's ability to handle high-dimensional datasets effectively have played a role in its better performance. Additionally, the presence of non-linear relationships in rice classification have been captured effectively by Random Forest. It also has low variance and high bias, making it less likely to overfit the data. Additionally, Random Forest is robust to noise and outliers in the data and can calculate the importance of each feature in the dataset, which can be used to select the most relevant features for classification. These factors combined likely contributed to the improved accuracy of Random Forest compared to other machine learning or deep learning models in rice classification.

| | Model | Accuracy |
|---|---|---|
| 0 | SVM | 89.333333 |
| 1 | Decision Tree | 77.333333 |
| 2 | Random Forest Tree | 92.666667 |
| 3 | CNN | 50.666666 |

Figure 12: Model Performance Table

# 6 Conclusion and Future Work

To conclude, we have seen how different models of machine learning and deep learning have performed in terms of accuracy to classify categories of rice. The presented system can help the agricultural companies to identify rice types in less time and accurately without the need of manual help. There have been many studies and comparisons conducted with rice in the past, and these may be found in the literature.

The Random Forest model demonstrated exceptional performance in comparison to other models such as SVM and Decision Tree. The RF model's ability to handle multiple features and high variance in datasets has made it a preferred choice for various machine learning problems. The results indicate that Random Forest models provide better accuracy of 92.66% and lower false positive rate in comparison to other models, which makes them more robust to overfitting. On the other hand, the SVM model showed limitations in dealing with large datasets and high-dimensional features, which may have led to its lower performance. Overall, the Random Forest model emerged as the superior choice due to its ability to handle complex data structures, its ability to mitigate overfitting, and its ability to provide better accuracy and stability. When looking at the studies with rice, the classification success that was attained with the algorithm CNN was mostly high. But it was not in this case since the accuracy rate was the lowest with 50.66%. It is important to note, however, that each data set is distinct from the others, and that the number of features of rice that are included in each data set is also unique. In Future we can add more layers to CNN model and train the model so that can help in achieving better performance. Some of the models that have more layers that can be implanted in future are VGG16 having 16 layers for CNN, VGG19 having 19 layers of CNN, ResNet50 having 50 layers of CNN. These can be implanted in future and their performance can be computed and compared with that of CNN. The addition of more layers will not affect our CNN implementation as these will be two different implementations of models. Although, the performance can be measured and compared with our implementation. On the basis of the obtained performance measurement results, it is feasible to conclude that the study was successful in identifying rice types and when compared with deep learning model, the traditional model performed better.

# References


Ahmed, T., Rahman, C. R. and Abid, M. F. M. (2021). Rice grain disease identification using dual phase convolutional neural network based system aimed at small dataset., *AgriRxiv* (2021): 20210263534.

Azmi, N., Kamarudin, L. M., Zakaria, A., Ndzi, D. L., Rahiman, M. H. F., Zakaria, S. M. M. S. and Mohamed, L. (2021). Rf-based moisture content determination in rice using machine learning techniques, *Sensors* **21**(5): 1875.

Barbedo, J. G. A. (2016). A review on the main challenges in automatic plant disease identification based on visible range images, *Biosystems engineering* **144**: 52–60.

Chatnuntawech, I., Tantisantisom, K., Khanchaitit, P., Boonkoom, T., Bilgic, B. and Chuangsuwanich, E. (2018). Rice classification using spatio-spectral deep convolutional neural network, *arXiv preprint arXiv:1805.11491* .

Granitto, P. M., Verdes, P. F. and Ceccatto, H. A. (2005). Large-scale investigation of weed seed identification by machine vision, *Computers and Electronics in Agriculture* **47**(1): 15–24.

Han, L., Haleem, M. S. and Taylor, M. (2015). A novel computer vision-based approach to automatic detection and severity assessment of crop diseases, *2015 Science and Information Conference (SAI)*, IEEE, pp. 638–644.

Huang, S.-y. and Zhang, B. (2019). Research on improved k-means clustering algorithm based on hadoop platform, *2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*, IEEE, pp. 301–303.

Ibrahim, S., Zulkifli, N. A., Sabri, N., Shari, A. A. and Noordin, M. R. M. (2019). Rice grain classification using multi-class support vector machine (svm), *IAES International Journal of Artificial Intelligence* **8**(3): 215.

Kaya, E. and Saritas, İ. (2019). Towards a real-time sorting system: Identification of vitreous durum wheat kernels using ann based on their morphological, colour, wavelet and gaborlet features, *Computers and Electronics in Agriculture* **166**: 105016.

Kezhu, T., Yuhua, C., Weixian, S. and Xiaoda, C. (2014). Identification of diseases for soybean seeds by computer vision applying bp neural network, *International Journal of Agricultural and Biological Engineering* **7**(3): 43–50.

Kuo, T.-Y., Chung, C.-L., Chen, S.-Y., Lin, H.-A. and Kuo, Y.-F. (2016). Identifying rice grains using image analysis and sparse-representation-based classification, *Computers and Electronics in Agriculture* **127**: 716–725.

Lin, P., Li, X., Chen, Y. and He, Y. (2018). A deep convolutional neural network architecture for boosting image discrimination accuracy of rice species, *Food and Bioprocess Technology* **11**(4): 765–773.

Liu, T., Chen, W., Wu, W., Sun, C., Guo, W. and Zhu, X. (2016). Detection of aphids in wheat fields using a computer vision technique, *Biosystems Engineering* **141**: 82–93.

Mahajan, S., Das, A. and Sardana, H. K. (2015). Image acquisition techniques for assessment of legume quality, *Trends in Food Science & Technology* **42**(2): 116–133.

Mohapatra, D., Tripathy, J. and Patra, T. K. (2021). Rice disease detection and monitoring using cnn and naive bayes classification, *Soft Computing Techniques and Applications*, Springer, pp. 11–29.

Patrício, D. I. and Rieder, R. (2018). Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review, *Computers and electronics in agriculture* **153**: 69–81.

Sanjaa, B. and Chuluun, E. (2013). Malware detection using linear svm, *Ifost*, Vol. 2, IEEE, pp. 136–138.

Sethy, P. K., Barpanda, N. K., Rath, A. K. and Behera, S. K. (2020). Deep feature based rice leaf disease identification using support vector machine, *Computers and Electronics in Agriculture* **175**: 105527.

Sun, C., Liu, T., Ji, C., Jiang, M., Tian, T., Guo, D., Wang, L., Chen, Y. and Liang, X. (2014). Evaluation and analysis the chalkiness of connected rice kernels based on image processing technology and support vector machine, *Journal of Cereal Science* **60**(2): 426–432.

Tipi, T., Yildiz, N., Nargeleçekenler, M. and Çetin, B. (2009). Measuring the technical efficiency and determinants of efficiency of rice (oryza sativa) farms in marmara region, turkey, *New Zealand Journal of Crop and Horticultural Science* **37**(2): 121–129.

Yi, X., Eramian, M., Wang, R. and Neufeld, E. (2014). Identification of morphologically similar seeds using multi-kernel learning, *2014 Canadian Conference on Computer and Robot Vision*, IEEE, pp. 143–150.