International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 01 | Jan 2021 www.irjet.net p-ISSN: 2395-0072

Application of Big Data Tools for Seed Classification

Kavitha Reddy Gurrala

Supply Chain Professional

Abstract - *Rice is the staple food among most Asian and African Countries, and it is one of the highly consumed cereal grains. It is also one of the highly produced agricultural commodities across the world. With limited agricultural lands for the cultivation of rice, the yield and productivity of the crop plays a main role in meeting the dietary needs of the growing population across the world.*

However, the crop yield and productivity are highly dependent on the type and quality of the seed used. Thus, in order to properly estimate the crop output levels and to ensure higher crop productivity and yield, seed selection plays a key role in ensuring the seed quality for a healthy and strong crop. Therefore, usage of certified seeds for sowing is highly recommended, as the certification agencies control the quality of the seeds by testing for "varietal purity" and through "classification of seeds- to avoid contaminations or weed mix-up", such that only the intended variety and breed is provided to the farmers. Currently, seed certification employs experts for manual paddy classification which is quite slow to meet the timely market requirements for seed sowing and further it is highly unreliable, as it employs manual procedures and only a few samples for classification.

Therefore, this paper aims at identifying the appropriate classification algorithms that automate the classification of seeds in a timely and reliable fashion in order to meet the market requirements for certified seeds, using the seed characteristics/attributes extracted from Images obtained through Computer Vision Technologies.

Key Words: Seed Classification, Hyperspectral Data, Seed Characteristics, Quality Assessment, Yield and Productivity.

1.INTRODUCTION

As rice is grown using a sequence of processes starting with Seed Selection followed with Land Preparation, Crop Establishment, Water Management, Nutrition Management, Crop Health, Harvesting, and Post-Harvesting. Any delay within the certified seed availability delays the whole cycle of Rice Growth. Therefore, usage of automated rice classification methods is highly recommended to ensure the timely and reliable classification and certification of seeds.

Currently available, computer vision technologies are capable enough to extract seed characteristics/attributes from Images, thus facilitating the adaptation of Big data tools for the development of prediction and classification models based on the seed characteristics captured from the image pixel data. Thus, the introduction of Machine vision technologies integrated with Big data tools is highly recommended for auto-classification of the rice seeds for seed certification. Further, identification and development of the right prediction and classification algorithms from the available pool of Big Data tools (Open source programming languages and packages i.e. R) is the need of the hour to ensure timely availability of certified seeds for reliable crop yields to meet the growing demand requirements for paddy across the world.

2. LITERATURE REVIEW

Nearly half of the world population is dependent on rice for food calories/protein and by the year 2025 the world would need about 760 million tons of paddy. In order to face the increasing demand for paddy one needs to narrow the yield gaps and particularly focus on reducing the usage of different varieties for sowing- as this alone contributes for around 10% of the yield gaps (Duwayuri et.al, 1998, Rice Knowledge Bank, IRRI). In addition, there are about more than 40,000 different varieties of rice across the world, that differ by length, colour, aroma, flavour etc (Rice Association, 2020).

Further, availability of several rice varieties within the markets with different quality levels leads to seed

adulteration and mix up at a commercial level for economic gains, therefore governments across the world initiated several certification bodies to supply pure seeds to the farmers in order to protect the farmers from yield loses, as sowing of mixed rice varieties results in yield losses resulting from differences in the time required from sowing until harvest (Vergara et.al, 1966, Vemireddy et.al, 2015).

Certified bodies, play a crucial role in providing pure and quality seeds to resource-poor tribal and rural communities, as selecting the right seeds stands as the basis for healthier seedlings fostering higher yields (Mishra et.al, 2012). Nonetheless, good quality seeds are essential for ensuring the productivity levels. However, timely availability of certified seeds often affects farmer's ability to sow the crop in a timely fashion to match the weather and seasonal conditions for seed germination and harvest (Cromwell, 1996). Hence, fostering timely availability of good quality certified seeds at reasonable prices is the need of the hour to ensure the yield rates and profits to the farmers (Parimala et.al, 2013).

Further, certification agencies often employ skilled experts to identify morphological structures, shapes and colours from tiny grains using traditional methods employing large magnifying glasses, an illumination, and forceps. Besides, such inspection's call out for many seed inspectors involving a lot of their time for inspection and classification- thus increasing costs of inspection/classification and reducing the efficiency and reliability of inspection/classification (Kiratiratanapruk et.al, 2020). Similarly, additional methods such as High-Performance Liquid Chromatography (HPLC) and Gas Chromatography-Mass Spectrometer (GC-MS) used for seed classification and certification are also very slow, expensive and draw conclusions based on very few samples impacting the reliability of the certification process. Therefore, there exists an immediate need to replace the traditional classification methods by automated methods employing image processing and data mining techniques to enhance the speed and reliability of classification. In addition, deep learning models can also be chained to learn intricate features of increasing levels of abstraction for reliable classification (Qiu et.al, 2018, Maheshwari and Renuga Devi, 2019).

3. DATA SET

The "Rice Data Set" used for building the classification models was obtained from Kaggle (Seyma, 2020). The data set provided information extracted from two rice varieties- "Gonen" and "Jasmine".

The "Jasmine" rice variety used within the data set, has its origins in Thailand. It is characterized by its superior physical appearance, cooking quality, grain aroma etc. and officially named as "Khao Dawk Mali 105" in short "KDM 105". It is one of the finest qualities of rice in Thailand but suffers from average yield levels that range around 66% of the world average rates (Rahman et.al, 2009). In contrast, the rice variety "Gonen" used within the data set, has its origins in Turkey. It is characterized as the third highest average rice yielding varieties within the world (Manners, 2013). In addition, the Turkish seed varieties are also characterized with higher Germination Energy and Total Germination rates than the standard rice varieties (Dimitrovski et.al, 2017).

In addition, the data set projected ten attributes of the rice seeds such as "Area", "MajorAxisLength", "MinorAxisLength", "Eccentricity", "ConvexArea", "EquivDiameter", "Extent", "Perimeter", "Roundness", "AspectRation" that described the rice characteristics and it also included 18,185 entries for each of these rice attributes.

Finally, it had a column titled "Class" identifying the variety of rice-based on the values of the ten attributes. Appendix I, displays the data frame summary listing all

the attributes measured. Further, the "Rice Data set" would be employed to build the classification models using "R" (widely used open source programming language for data analysis) and to identify the algorithm that yields the highest accuracy of classification.

4. METHODOLOGY

The rice data set portrays the features extracted from the pre-processed images obtained from Hyperspectral Systems, further such features would be used to build traditional and deep learning classification models in order to select an optimal classification algorithm that yields the highest accuracy of classification. The figure 4.1 below displays the induction and deduction process for development of a learned models using the training data (which includes 70% of the total data set) and application of the learned models to the testing data (which includes 30% of the total data set) in order to judge the classification accuracy.

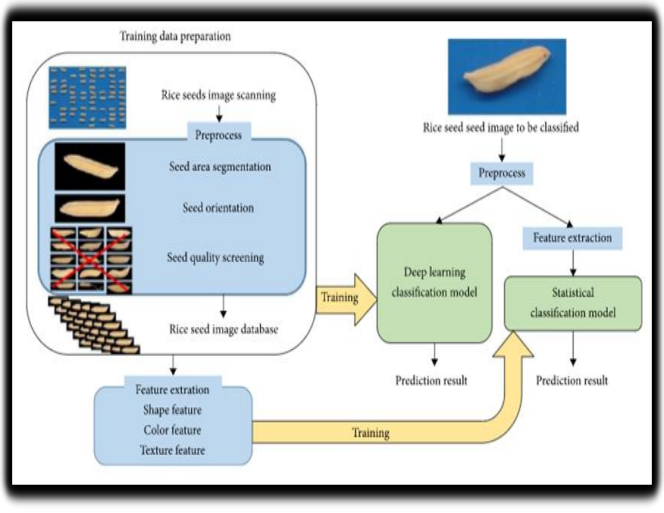

Figure 4.1, Rice Seed Classification Process, Adapted from (Kiratiratanapruk et.al, 2020).

Further, as all the rice characteristics provided within the data set (input variables) were "continuous" in nature, except the "Class" attribute (output/class/response variable of interest) that described the rice categories was "categorical" in nature, the data was explored using the "ggplot (data visualization package in R)" and "ggpubr (package used for customization of plots created using the ggplot package)" packages within "R" to create "Box Plots" in order to visualize and understand, how the distribution of the continuous attributes varied within the two rice categories.

In addition, the "CARET" (Classification and Regression Training) package within the "R" which is generally employed for data-splitting, pre-processing, feature selection, model finetuning using resampling and variable importance estimation (Max, 2019), was used for identifying the most important attributes within the data set using the "Learning Vector Quantization Model" and further used in building Classification models as listed in figure 4.2 below.

Besides, "R" was used to connect to "Spark" environment in order to enhance the computational speeds through parallel processing through installing spark 2.4 version and using the package "sparklyr" for the establishing the connection.

Further, a single computer was configured as a master node of the Spark machine in order run the algorithms. In addition, the spark environment was used to build several ml_models for classification as listed in figure 4.2 below (Spark, 2020).

Likewise, "R" was also used to connect to the "H2O" environment- the scalable open source machine learning platform, using the "h2o" package that generally aids with parallelized implementations of many supervised and unsupervised machine learning algorithms in order to build "Deep Leaning Neural Network Models" for Classification (Cran.r, 2020).

Both, the "ml_multilayer_perceptron_classifier models" within the "Spark" environment and the "Deep Leaning Neural Network Models" within the "H2O"

environment were developed with multiple scenarios changing the number of layers and nodes within the hidden layer architecture in order to identify the optimal hidden layer architecture that maximizes the classification accuracy within each of the environments.

Caret-based algorithms	Sparklyr-based algorithms	H2O-based algorithms
Generalized Linear Model for Classification (glm)	ml_logistic_regression_model(ml_log)	Deep Leaning NN Models (h20.DL)
Rule-Based Classifier (part)	ml_naive_bayes model (ml_nb)	
k-nearest neighbours-instance based classifier (knn)	ml_decision_tree model (ml_dt)	
Support Vector Machine with Linear Kernel (svmLinear)	ml_gradient_boosted_trees model (ml_gbt)	
Support Vector Machine with Polynomial Kernel (svmPoly)	ml_multilayer_perceptron_classifier (ml-nn)	
CART Decision Tree Model (rpart)		
C4.5 Decision Tree Model (J48)		
Ensembled Tree -Bagged Cart (treebag)		
Ensembled Tree-Random Forest (rf)		

Figure 4.2, Classification Models Employed

5. DATA ANALYSIS

Based on the Box Plots displayed within the figure 5.1 below, the mean values for the "Area", "MajorAxisLength", "MinorAxisLength", "ConvexArea", "EquivDiameter", "Extent", "Perimeter", "Roundness", of the "Gonen" variety is observed to be higher than the "Jasmine" variety. In contrast, the mean values for the "Eccentricity" and "AspectRation" of the "Jasmine" variety is observed to be higher than the "Gonen" variety.

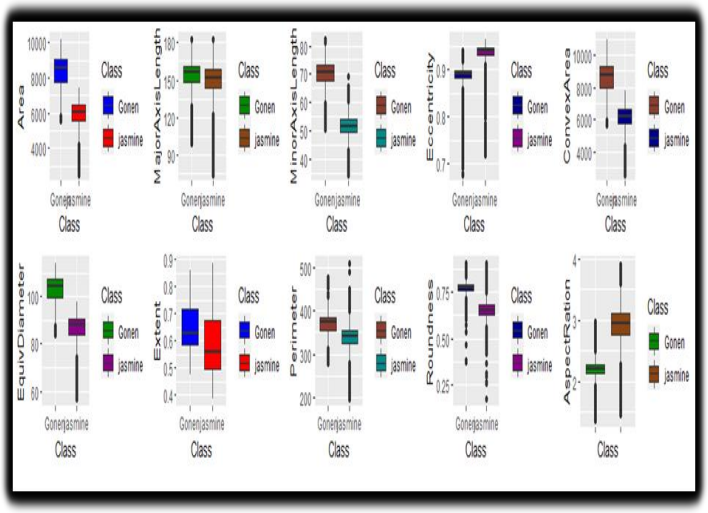

Figure 5.1, Box Plots
(Rice Characteristics & Class)

In addition, figure 5.2 below displays the output from the "Learning Vector Quantization Model" constructed to estimate the variable importance within the training data set. Based on the plot, it is evident that the "MinorAxisLength" stands out as the most important variable within the data set followed by "Eccentricity", "AspectRation" , "Roundness", "EquivDiameter", "AREA", "ConvexArea", and "Perimeter", "Extent", "MajorAxisLength" stands out as the least important attributes within the data set.

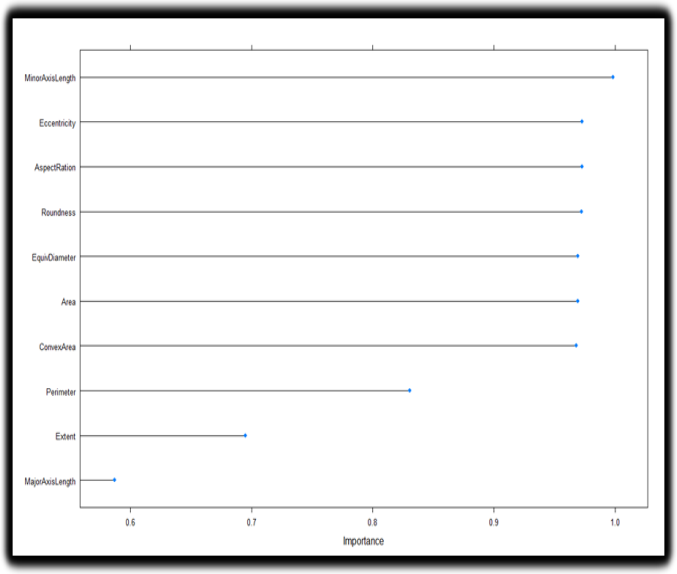

Figure 5.2, Feature Ranking -Caret R Package

As the data set provided information on only 18,185 entries for each of the ten rice attributes, it characterizes a small data set in size, therefore performance evaluation based on "Time required for training model convergence" and "Model prediction times using the test data" cannot be considered as key performance measures for the gauging the functionality of the Classification Algorithms, hence the selection of an optimal Classification Algorithms from the pool of classification algorithms employed was only based on highest accuracy of classification.

Further, the figures 5.3, 5.4 and 5.5 below, respectively summarizes the performance of the Caret-based Algorithms, Sparklyr-based algorithms and performance comparison of Caret vs Sparklyr based Algorithms.

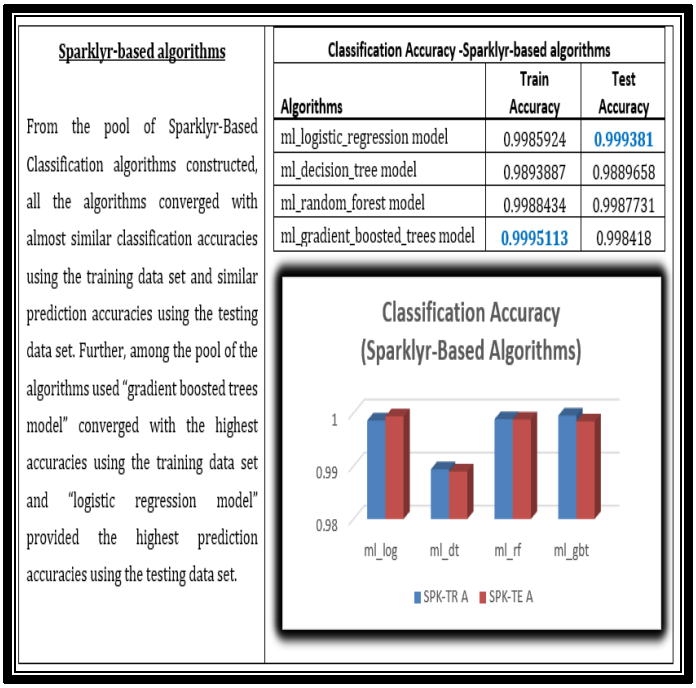

Figure 5.4, Classification Accuracy (Sparklyr-based-Models)

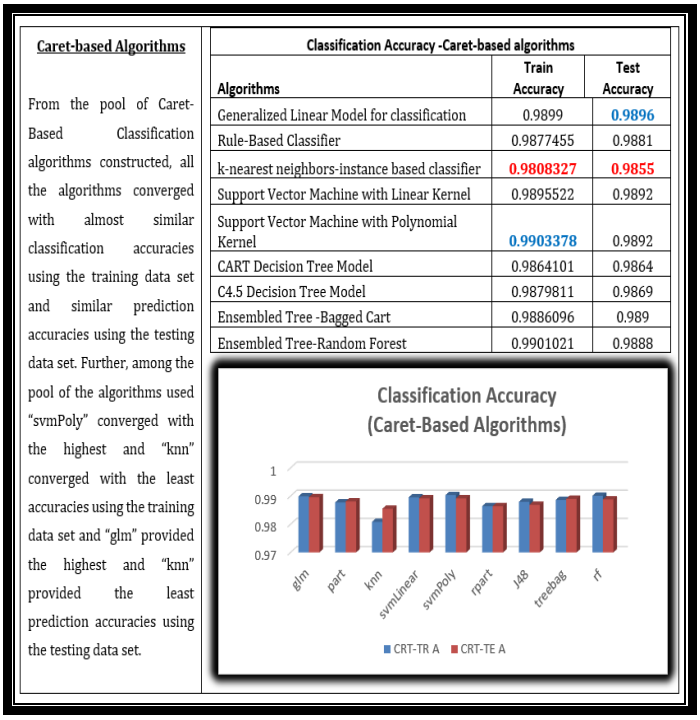

Figure 5.3, Classification Accuracy (Caret-based-Models)

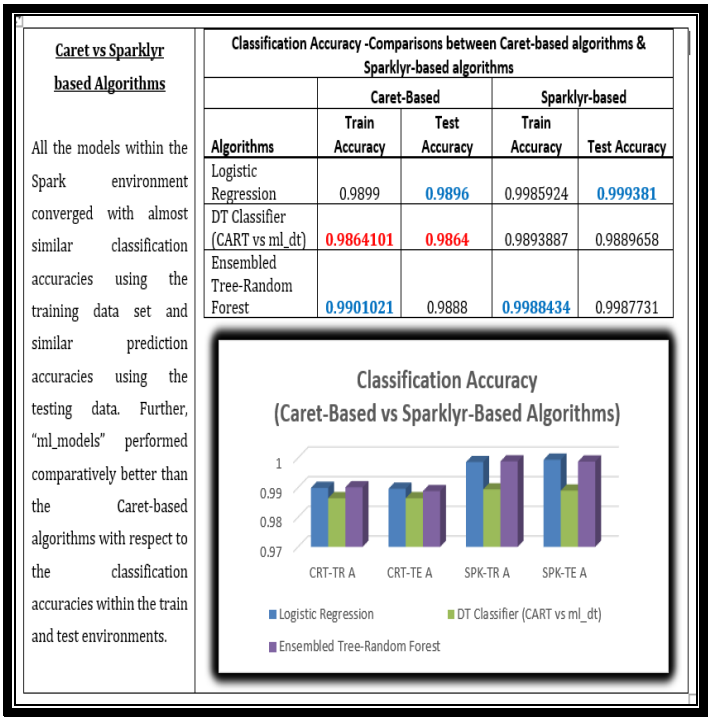

Figure 5.5, Classification Accuracy (Caret vs Sparklyr based Models)

In addition, the figure 5.4 below, summarizes the performance of the Neural Network Models with different Hidden layer structures within the Spark and H2O Environments. Based on the figure it is observed that NN Models within the H2O environment performed consistently with regards to classification accuracies within all hidden layer configurations adopted. Further, the NN models provided the optimal classification accuracies with a simple hidden layer structure employed. Even though model convergence times are not being considered as one of the measures for performance, the convergence times of the NN models were observed to be slightly higher within the Spark Environment comparative to the convergence times of the NN models within the H2O environment.

Classification Accuracy -Neural Network Models

		Sparklyr-based (ml_multilayer_perceptron_classifier)		H20 -based-Deep Leaning Neural Network Models	
Scenarios	Hidden Layer Structure	Train Accuracy	Test Acuracy	Train Accuracy	Test Acuracy
1	(100,75,50,25)	0.5000716	0.4996668	0.9811	0.9809
2	(75,50,25,10)	0.9765748	0.9765748	0.9846	0.9842
3	(50,25,25,10)	0.9937453	0.9916787	0.9826	0.9786
4	(25,25,25,25)	0.9801276	0.9827823	0.9859	0.9844
5	(10,10,10,10)	0.5	0.5	0.988	0.9879
6	(100,50,25)	0.9529659	0.954538	0.9832	0.9833
7	(75,50,25)	0.5002394	0.5008331	0.9823	0.9808
8	(50,25,10)	0.9906455	0.9913542	0.9847	0.9837
9	(25,25,25)	0.9900936	0.9905949	0.9879	0.9881
10	(10,10,10)	0.5	0.5	0.987	0.9866
11	(100,75)	0.5002237	0.5008331	0.9831	0.9828
12	(75,50)	0.9958242	0.9958537	0.9849	0.9848
13	(50,25)	0.9937722	0.9938585	0.9844	0.9835
14	(25,25)	0.4999665	0.4991669	0.9859	0.9857
15	(10,10)	0.5	0.5	0.9887	0.9894

Figure 5.6, Classification Accuracies Neural Network Models

6. CONCLUSIONS

Based on the performance of the models generated within different environments, the "ml_models" can be adapted by the certification agencies for automatic classification of Rice Varieties, as these algorithms provided highest classification accuracies comparative to similar algorithms from other packages.

Further, they can be used to formulate models using large data sets as they employ parallel processing of data and provide high computational performance compared to the algorithms within the Caret Package. In addition, based on the performance of NN models- the "Deep Leaning Neural Network Models" within the "H2O" environment can be adapted by the certification agencies for automatic classification of Rice Varieties, as they provided highest classification accuracies with simple hidden layer structures consuming less times for convergence compared to the "ml_multilayer_perceptron_classifier models" within the "Spark" environment.

7. FUTURE EXTENSIONS

As the optimality decision for selecting the best classification algorithms from the pool of algorithms available within the Caret Package, Spark Domain and H2O Domain was based on the "Model Classification Accuracies" alone. Hence, the "R" programs developed can be executed for larger data set sizes, so that additional performance metrics such as "Time required for training model convergence" and "Model prediction times using the test data" can also be considered towards defining an optimal classifier to be adapted by the Certification Agencies for timely and reliable classification of the Rice Varieties.

REFERENCES

[1] Apache Spark, "Machine Learning Library", Apache Spark, Accessed on: Dec. 01, 2020. [Online]. https://spark.apache.org/docs/latest/ml-guide.html

[2] Cran.r, "Package 'h2o", Cran.r, Oct 2020, Accessed on Dec. 01, 2020. [Online]-https://cran.r-project.org/web/packages/h2o/h2o.pdf

[3] Cromwell E, "Governments, farmers and seed in a changing Africa", CABI Publishing, Wallingford, UK, 1996.

[4] Dimitrovski. T, Andreevska. D, Andov. D, Simeonovska. E and Ibraim. J, "Some Seed Quality Properties of Newly Introduced Turkish Rice Varieties (Oryza sativa L.) Grown Under the Environmental Conditions of Republic of Macedonia", Congress Book, 2017.

[5] Duwayuri. M, Tran. D. V and Nguyen. V. N, "Reflections in Yield Gaps in Rice Production: How to narrow the Gaps", International Rice Commission, 1998, Cairo, Egypt.

[6] Kiratiratanapruk. K, Temniranrat. P, Sinthupinyo. W, Prempree. P, Chaitavon. K, Porntheeraphat. S and Prasertsak. A "Development of Paddy rice seed classification process using machine learning techniques for automatic grading machine", Journal of Sensors, 2020.

[7] Maheshwari. S and Renuga Devi. M, "Paddy Seed Classification and Identifying Varieties using Random Assessment Classification", International Journal of Engineering and Advanced Technology (IJEAT), vol. 9 issue.2, pp. 2682-2685, 2019.

[8] Manners. G, "Rice Booms in Turkey", Rice Today, Jan 2013. Accessed on Dec. 01, 2020. [Online]. https://ricetoday.irri.org/rice-booms-in-turkey/

[9] Max. K, "The Caret Package", github, Mar 2019, Accessed on Dec. 01, 2020. [Online]. http://topepo.github.io/caret/index.html

[10] Mishra. S, Chaudhury. S. S, and Arivudai Nambi. V "Strengthening of traditional paddy selection practices of tribal farm families with improved knowledge and skills in Koraput district, Odisha", Indian Journal of Traditional Knowledge, vol.11, pp. 461-470, 2012.

[11] Parimala. K, Subramanian. K, Kannan. S. M and Vijaya Lakshmi. K, "A Manual on Seed Production and Certification", Centre for Indian Knowledge Systems, Chennai Revitalizing Rainfed Agriculture Network, 2013.

[12] Qiu. Z, Chen. J, Zhao. Y, Zhu. S, He. Y and Zang. C, "Variety Identification of Single Rice Seed Using Hyperspectral Imaging Combined with Convolutional Neural Network", Applied Sciences, vol.8, issue. 2, 2018.

[13] Rahman. S, Aree. W, Sreeboonchita, S and Chavanapoonphol. Y, "Production Efficiency of Jasmine Rice Producers in North and North-Eastern Thailand", Journal of Agricultural Economics, 2009.

[14] Rice Association, "Types of Rice", Rice Association. Accessed on Dec. 01, 2020. [Online]. Available: http://www.riceassociation.org.uk/content/1/18/types-of-rice.html

[15] Rice Knowledge Bank, International Rice Research Institute(IRRI): http://www.knowledgebank.irri.org/decision-tools/rice-doctor/rice-doctor-fact-sheets/item/salinity

[16] Seyma. S, "Rice Seed Dataset (Gonen & Jasmine", Kaggle, Mar 2020. Accessed on Nov. 12, 2020. [Online]. Available: https://www.kaggle.com/seymasa/rice-dataset-gonenjasmine

[17] Vemireddy. L. R, Satyavathi. V. V, Siddiq. E. A and Nagaraju. J, "Review of methods for the detection and quantification of adulteration of rice: Basmati as a case study", Journal of Food Science and Technology, vol. 52, issue. 6, pp: 3187-3202, 2015.

[18] Vergara. B. S, Tanaka. A, Lilis. R and Puranabhavung. S, "Relationship between growth duration and grain yield of rice plants", Soil Science and Plant Nutrition, vol. 12, no. 1, 1966.

Appendix -I

Data Frame Summary

dat

Dimensions: 18185 x 12

Duplicates: 0

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	id [integer]	Mean (sd) : 9093 (5249.7) min < med < max: 1 < 9093 < 18185 IQR (CV) : 9092 (0.6)	18185 distinct values (Integer sequence)		18185 (100%)	0 (0%)
2	Area [integer]	Mean (sd) : 7036.5 (1467.2) min < med < max: 2522 < 6660 < 10210 IQR (CV) : 2461 (0.2)	5343 distinct values		18185 (100%)	0 (0%)
3	MajorAxisLength [numeric]	Mean (sd) : 151.7 (12.4) min < med < max: 74.1 < 153.9 < 183.2 IQR (CV) : 14.4 (0.1)	18185 distinct values		18185 (100%)	0 (0%)
4	MinorAxisLength [numeric]	Mean (sd) : 59.8 (10.1) min < med < max: 34.4 < 55.7 < 82.6 IQR (CV) : 18.8 (0.2)	18185 distinct values		18185 (100%)	0 (0%)
5	Eccentricity [numeric]	Mean (sd) : 0.9 (0) min < med < max: 0.7 < 0.9 < 1 IQR (CV) : 0 (0)	18185 distinct values		18185 (100%)	0 (0%)
6	ConvexArea [integer]	Mean (sd) : 7225.8 (1502) min < med < max: 2579 < 6843 < 11008 IQR (CV) : 2520 (0.2)	5450 distinct values		18185 (100%)	0 (0%)
7	EquivDiameter [numeric]	Mean (sd) : 94.1 (9.9) min < med < max: 56.7 < 92.1 < 114 IQR (CV) : 16.4 (0.1)	5343 distinct values		18185 (100%)	0 (0%)
8	Extent [numeric]	Mean (sd) : 0.6 (0.1) min < med < max: 0.4 < 0.6 < 0.9 IQR (CV) : 0.2 (0.2)	18007 distinct values		18185 (100%)	0 (0%)
9	Perimeter [numeric]	Mean (sd) : 351.6 (29.5) min < med < max: 197 < 353.1 < 508.5 IQR (CV) : 39 (0.1)	16246 distinct values		18185 (100%)	0 (0%)
10	Roundness [numeric]	Mean (sd) : 0.7 (0.1) min < med < max: 0.2 < 0.7 < 0.9 IQR (CV) : 0.1 (0.1)	18184 distinct values		18185 (100%)	0 (0%)
11	AspectRation [numeric]	Mean (sd) : 2.6 (0.4) min < med < max: 1.4 < 2.6 < 3.9 IQR (CV) : 0.8 (0.2)	18185 distinct values		18185 (100%)	0 (0%)
12	Class [factor]	1. Gonen 2. jasmine	8200 (45.1%) 9985 (54.9%)		18185 (100%)	0 (0%)