# Agro Informatic Data Classification using Machine Learning Techniques

G. Shobana[1]
Department of Computer Applications
Madras Christian College
Chennai, India
gmshobana@gmail.com

K. Uma maheswari[2]
Department of Computer Science
Bharathi Women's College
Chennai, India
uma.tvr1981@gmail.com



***Abstract*—Agro-informatics is one of the newly emerging fields where machine learning and data analytics are applied to the agricultural data to assess grain quality, predict crop disease at an earlier stage, classify intra-species of crops and help in the analysis of several other agriculture related issues. Manual investigation of crops for its quality, size and disease is a tedious time-consuming process. The conventional methods also involve high production costs. Advanced technology involves automatic image capturing, data acquisition, pre-processing and application of machine learning for the analysis. In this paper, dry-beans data are classified using a proposed feature reduction framework using Principal Component Analysis and Linear Discriminant Analysis with machine learning algorithms. Multi-layer perceptron, Random Forest and Support Vector Machine performed well with 93% for multinomial classification while 98% for binary classification. With balanced and comprehensive groups of two bean samples, binary classification achieved higher accuracy than multinomial classification with seven groups of beans.**

***Keywords*—Agro-informatic; dry-bean; multi-layer perceptron; random forest;support vector machine.**

## I. INTRODUCTION

Agriculture has become more challenging with increasing population and unpredictable climatic changes. Eighty-two different countries across the world are currently facing food insecurity according to the statistics provided by World Food Programme (WFP). Around 828 million people struggle to get their regular meals and stay hungry. Forty-five countries are on the verge of experiencing famine [1]. The crop yield can be increased by adopting advanced technology to identify the most suitable soil for cultivation of a specific crop. Sustainable methodologies must be incorporated with the use of organic soil enrichment procedures. Soil type, temperature, water requirement and landscape vary for different crops. Cotton plant requires black soil for best yield. The quality and size of the seed determines the future yield [2]. Hence selection of high-quality seeds become imperative. Investigating the seeds manually is time consuming and less accurate. To overcome this challenge, machine learning can be employed for the analysis. Machine learning has become an integrated part of the analytic procedure for many domains like Cheminformatics, Bioinformatics, Geoinformatics and Agroinformatics as shown in Fig. 1.

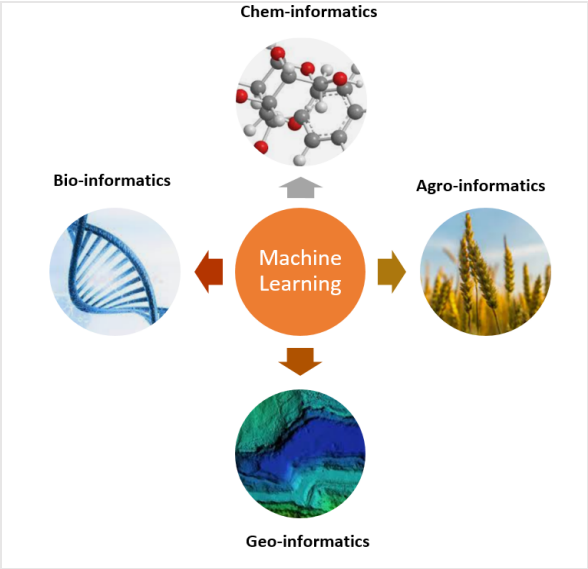


**Fig. 1.** Machine learning and its applications in different fields

Image processing techniques overcome the disadvantages that occur due to conventional seed measurement techniques. These advanced methods provide better results and accuracy. Agroinformatics involves application of machine learning methods and other computer-based computational techniques for the analysis of proliferating agricultural data. In this paper, dry beans are classified using machine learning algorithms. Both binary and multinomial classifications are performed to compare their prediction accuracy. In multinomial classification, the accuracy achieved was 93% while with binary classification the accuracy was 98%. The proposed methodology is a framework with three steps of normalization, Principal Component Analysis (PCA) and application of machine learning models. Linear Discriminant Analysis (LDA) was also applied instead of PCA and the results were observed. Framework with LDA had higher prediction accuracy compared to PCA. The procedure was implemented using Scikit-learn [3]. Cravero et al have found that SVM, RF and Neural Network perform well compared to other machine learning models with Agro-informatic data [4]. A nonlinear dimensionality reduction method is kernel PCA. Scikit-Learn can be used to implement LDA, PCA, and Kernel PCA. These are the transformers that assist in choosing crucial aspects. Three significant tasks are included in the pipeline-structured



eISSN 1303-5150

proposed process. The scaling procedure, the transformer, and the estimator are components of each pipeline.

## II. Related work

Extensive studies are carried out to apply machine learning to the agricultural domain, where data is enormous. Agroinformatics involves efficient application of computer-based techniques and algorithms to classify or predict agricultural data.

Mohamed et al have proposed additional attributes that map morphological data to the texture ratios. They have built an indigenous dataset with three Giza rice types namely 171,176 and 186. With 291 features, SVM produced an accuracy of 95%. They have adopted a comprehensive image acquisition procedure that greatly reduces pre-processing. Several other grains and seeds can be classified by applying their proposed algorithm [5].

Sudarshan et al have constructed self-organizing maps to perform automated efficient dry bean classification [6]. Deachrut et al have applied transfer learning for 1200 images of paddy seed. They obtained an accuracy of 83.33% for both MobileNetV2 and InceptionV3. They conclude by stating InceptionV3 as the best pre-trained weight that has the least test loss of 28.41%. Their proposed work may be used efficiently to identify various types of other rice varieties and helps in identifying high quality rice [7].

Yasir et al applied MobileNetV2 architecture for smart seed classification and the result was 95% on the test data. The experiment involved 14 types of seed varieties and the procedure of implementing deep learning convolutional neural networks yielded high accuracy [8].

Somsawut et al used 1164 images to classify Jasmine rice germination and achieved 89% accuracy by implementing convolutional neural networks [9].

Analyn et al classified fresh Excelsa beans using Mask R-CNN. Their proposed method automatically classified beans that are black, sour, cut or insect damaged from fresh beans, 87.5% accuracy was obtained through their procedure [10].

Meo Vincent et al classified different coffee beans using SVM and they yielded an accuracy of 70%. SVM has proved an efficient algorithm for agricultural data [11].

Vinay et al employed convolutional neural networks to classify wheat rust diseases. They used 2000 images of wheat plants with a stochastic gradient descent method they obtained 97.16% [12].

Panuwat et al applied Decision tree, Random forest, Gradient boosting and Naïve-Bayes algorithms to classify rice leaf diseases. The images were drawn from the UCI repository for the study. Random forest obtained an accuracy of 69.44%. They explored three types of rice leaf diseases namely Brown spot, Leaf smut and Bacterial leaf blight disease [13]. Jirasak et al applied three methods namely CNN model, VGG19 and MobileNet to classify potato and leaf diseases. Among the three models, VGG19 obtained 95.56% and 96.30% for potato and grape leaf disease classification [14].

## III. Materials and Methods

### Feature Description and Dataset

Machine learning algorithms have effectively forecasted agricultural productivity and helped with early detection of a number of crop illnesses. These prediction models can be used to examine the soil's quality and the suitable crops it supports. Pests that reduce the crops' typical production are another important worry during the growing phase.

The Dry bean dataset is available in the open repository and was added to the repository in 2020. 13,611 Samples with 17 attributes were obtained from the UCI repository [15]. 7 bean types were taken for the investigation. They are Barbunya, Bombay, Cali, Dermosan, Horoz, Sekar and Sira.

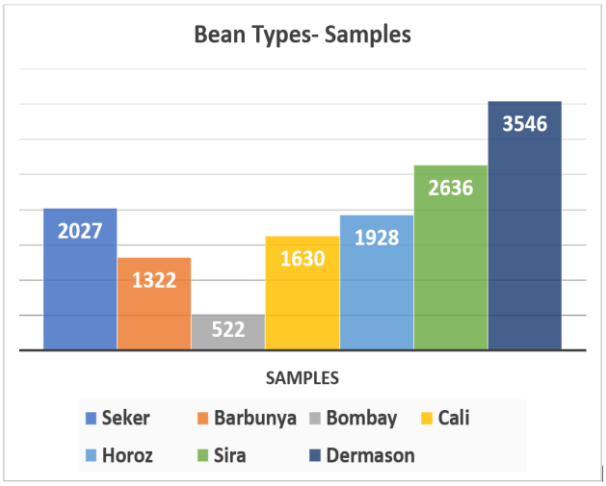


Fig. 2. Distribution of the dry bean samples [15]

Fig. 2. displays the distribution of dry beans by class, with 0 denoting Seker, 1 denoting Barbunya, 2 denoting Bombay, 3 denoting Cali, 4 denoting Horoz, 5 denoting Sira, and 6 denoting Dermason. There are 2027 Seker samples, 1322 Barbunya samples, 522 Bombay samples, 1630 Cali samples, 1928 Horoz samples, 2636 Sira samples, and 3546 Dermason samples. Dry beans are photographed in high resolution, and using image processing methods and specific technical measures, the features are extracted. The machine learning approach is then used with the generated features to make the prediction, Solidity defines the convex shell pixel ratio, also called convexity.

TABLE I. Samples of Bean Types

| Class | Bean Types | Samples |
|---|---|---|
| 0 | Seker | 2027 |
| 1 | Barbunya | 1322 |
| 2 | Bombay | 522 |
| 3 | Cali | 1630 |
| 4 | Horoz | 1928 |
| 5 | Sira | 2636 |
| 6 | Dermason | 3546 |

,





eISSN 1303-5150

```
RangeIndex: 13611 entries, 0 to 13610
Data columns (total 17 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Area             13611 non-null  int64
 1   Perimeter        13611 non-null  float64
 2   MajorAxisLength  13611 non-null  float64
 3   MinorAxisLength  13611 non-null  float64
 4   AspectRation     13611 non-null  float64
 5   Eccentricity     13611 non-null  float64
 6   ConvexArea       13611 non-null  int64
 7   EquivDiameter    13611 non-null  float64
 8   Extent           13611 non-null  float64
 9   Solidity         13611 non-null  float64
 10  roundness        13611 non-null  float64
 11  Compactness      13611 non-null  float64
 12  ShapeFactor1     13611 non-null  float64
 13  ShapeFactor2     13611 non-null  float64
 14  ShapeFactor3     13611 non-null  float64
 15  ShapeFactor4     13611 non-null  float64
 16  Class            13611 non-null  int64
dtypes: float64(14), int64(3)
memory usage: 1.8 MB
```

**Fig. 3.** Features and its datatype

Table.1 shows the distribution of the dry bean types. Multinomial and binary classification of the data was performed to understand the performance of machine learning algorithms for both type of classification. Sekar and Sira with 2027 and 2636 samples were taken for the binary classification study, while all the seven types were taken for the multiclass classification study. Fig. 3 displays the features and their data types. It also reveals that there are no missing values in the dataset taken for the investigation. The dataset has both integer and float data types. The column 'Class' represents the classification of the dry-bean types.

Proposed methodology

The proposed framework involves three steps of normalization, feature reduction and application of machine learning algorithms as shown in Fig.4. The output of the methodology with LDA and PCA was compared. High classification accuracy was obtained when LDA was used in the pipeline rather than PCA.

Procedure:

Step 1: Obtain the dry-bean dataset from the open UCI repository.

Step 2: Perform Pre-processing if required.

Step 3: Create a pipeline for each machine learning model to be tested.

Step 4: Normalize or perform scaling.

Step5: Linear Discriminant Analysis technique for feature extraction.

Step 6: Record the performance of the machine learning models with the extracted features.

Step 7: Repeat step 3 to 6, replacing LDA with PCA.

Step 8: Compare the results and select the best model.

70:30 is the ratio of training and validation. With LDA the machine learning algorithms perform well compared to PCA. Multinomial classification results in 93% accuracy while binary classification obtains 98% accuracy.

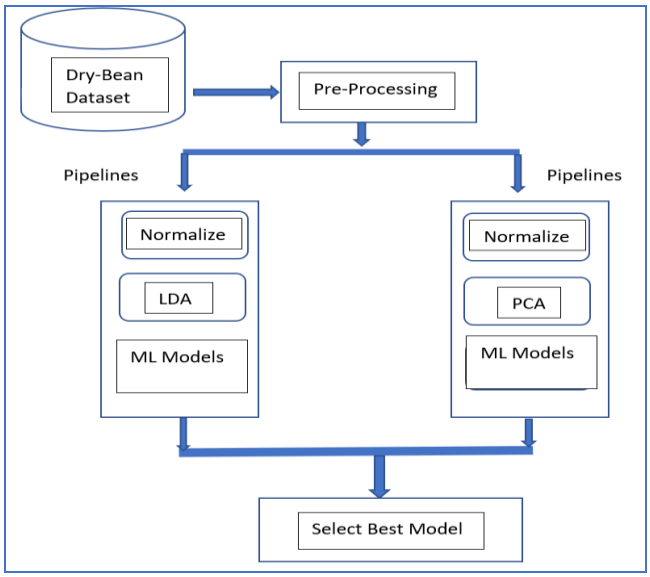


Fig. 4. Proposed Methodology

## IV. Performance Comparison

Dry bean dataset was donated to the UCI open repository by Murat koklu and Ilker Ali Ozkan during the year 2020. This is a less explored dataset and the researchers have explored the performance of conventional machine learning algorithms. Fig.5 shows the results of the proposed method and the conventional method.

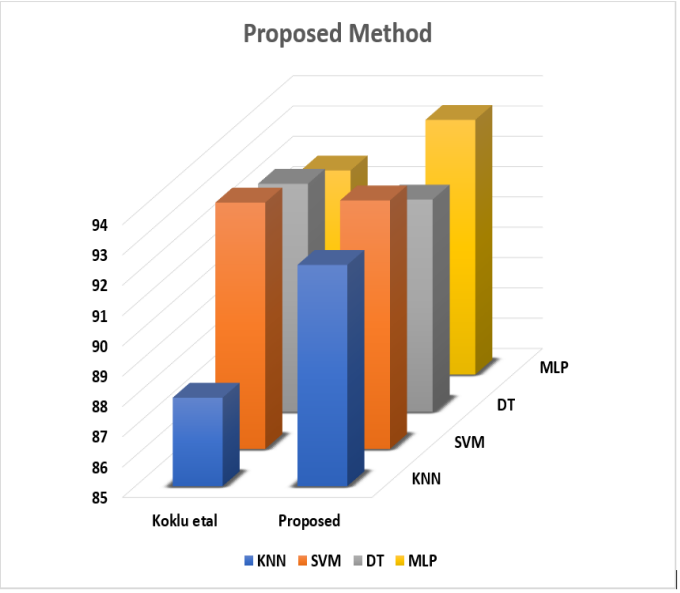


**Fig. 5**. Proposed method and conventional method



eISSN 1303-5150

Support Vector machine and Multilayer perceptron shows a very significant difference. Fig. 6 shows the performance of the framework with LDA and PCA. With the feature extraction of LDA, the framework has increased the overall classification accuracy of all the models with more than 92% while with PCA the performance was around 85%.

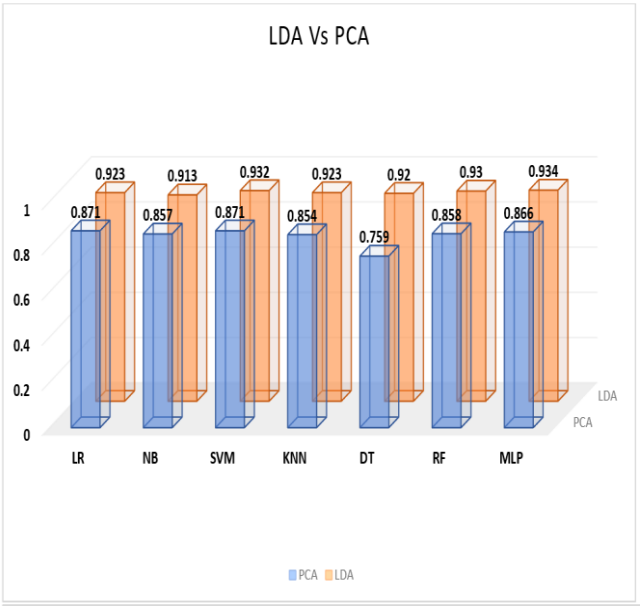


**Fig. 6**. Performance of LDA and PCA in framework

## V. Results and Discussions

The proposed framework with LDA has enhanced the performance of all the models. Fig.7 shows that Multilayer perceptron, random forest and SVM have high prediction accuracy.

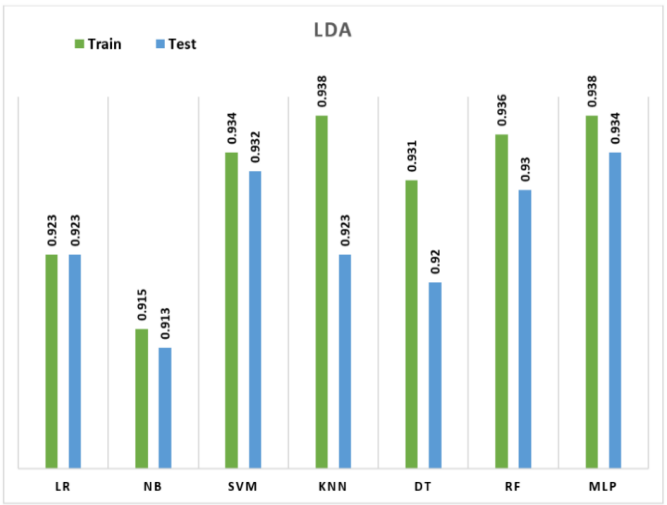


**Fig. 7**. Training and testing results with LDA

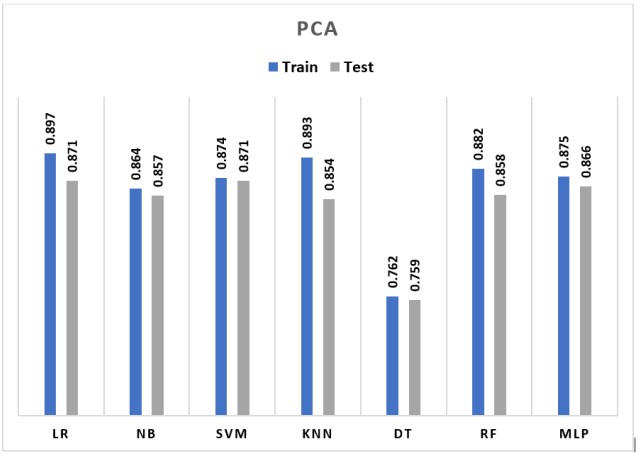


**Fig. 8**. Training and testing results with PCA

TABLE II. LDA and PCA training and testing results

| *Models* | *Training LDA* | *Testing* | *Training PCA* | *Testing* |
|---|---|---|---|---|
| **LR** | 0.923 | 0.923 | 0.897 | 0.871 |
| **NB** | 0.915 | 0.913 | 0.864 | 0.857 |
| **SVM** | 0.934 | 0.932 | 0.874 | 0.871 |
| **DT** | 0.931 | 0.920 | 0.893 | 0.854 |
| **KNN** | 0.938 | 0.923 | 0.762 | 0.759 |
| **RF** | 0.936 | 0.930 | 0.882 | 0.858 |
| **MLP** | 0.938 | 0.934 | 0.875 | 0.866 |

Fig.8 displays the training and testing results of PCA. Table 2. Shows the training and testing results of framework with LDA and PCA respectively for multinomial or multiclass classification [17]. When the same dataset was applied to classify only two types of bean types, the prediction accuracy was higher.

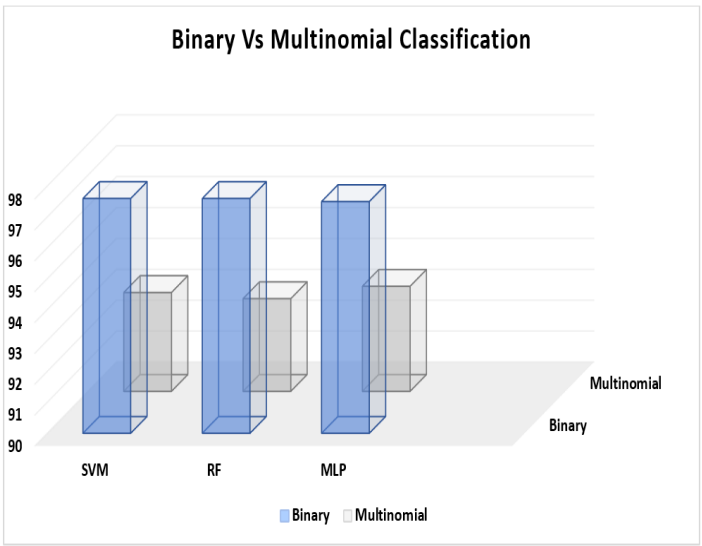


**Fig. 9**. Binary and multinomial performance comparison



eISSN 1303-5150

Fig. 9 shows the significant difference between the binary and multinomial classification. With Kernel PCA in the pipelined structure, the highest accuracy was 91%. SVM and MLP were the best performing models.

TABLE III. Binary and multinomial classification of the dataset

| Models | Training Multinomial | Testing | Training Binary | Testing |
|---|---|---|---|---|
| **SVM** | 0.934 | 0.932 | 0.978 | 0.976 |
| **RF** | 0.936 | 0.930 | 0.986 | 0.976 |
| **MLP** | 0.938 | 0.934 | 0.988 | 0.975 |

Table 3. shows the training and validation results obtained from binary and multiclass classification of the dataset. Sekar and Sira bean types were taken for the binary classification study. This methodology also performs efficiently for disease classification [18].

## VI. CONCLUSION

Agroinformatics involves the application of computers to analyze agricultural data. Machine learning techniques are employed in various fields like crop, soil and water management. They are also used in quality seed classification and for identifying crop diseases. In this paper, dry beans were classified with the pipelined framework that involved LDA and PCA. Framework with LDA yielded 93% accuracy. The binary and multinomial classification was also explored for the dry bean dataset. Binary classification with a balanced dataset achieved a high prediction accuracy of 98%. The future work would involve application of the framework with other grains or seed classification. Performance of the multinomial classification with balanced dataset and with augmented feature set can be examined.

## REFERENCES

[1] https://www.wfp.org.

[2] Rayda Ben Ayed, Mohsen Hanana, "Artificial Intelligence to Improve the Food and Agriculture Sector", Journal of Food Quality, vol. 2021, Article ID 5584754, 7 pages, 2021. https://doi.org/10.1155/2021/5584754.

[3] www.scikit-learn.org.

[4] Cravero, A.; Pardo, S.; Sepúlveda, S.; Muñoz, L. Challenges to Use Machine Learning in Agricultural Big Data: A Systematic Literature Review. Agronomy 2022, 12, 748.

[5] Habib, Mohamed. (2021). An enhanced seeds categorization and classification based on multiple features-set. Indian Journal of Computer Science and Engineering. 12. 779-789. 10.21817/indjcse/2021/v12i4/211204007.

[6] Sudarshan S, "Classification of Dry Beans Using Image Features," 2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2021, pp. 0950-0956, doi: 10.1109/UEMCON53757.2021.9666732.

[7] D. Jaithavil, S. Triamlumlerd and M. Pracha, "Paddy seed variety classification using transfer learning based on deep learning," 2022 International Electrical Engineering Congress (iEECON), 2022, pp. 1-4, doi: 10.1109/iEECON53204.2022.9741677.

[8] Y. Hamid, S. Wani, A. B. Soomro, A. A. Alwan and Y. Gulzar, "Smart Seed Classification System based on MobileNetV2 Architecture," 2022 2nd International Conference on Computing and Information Technology (ICCIT), 2022, pp. 217-222, doi: 10.1109/ICCIT52419.2022.9711662.

[9] S. Nindam, T. -O. Manmai and H. J. Lee, "Multi-Label Classification of Jasmine Rice Germination Using Deep Neural Network," 2022 7th International Conference on Business and Industrial Research (ICBIR), 2022, pp. 264-268, doi: 10.1109/ICBIR54589.2022.9786383.

[10] A. N. Yumang, M. Chloe M. Sta. Juana and R. L. C. Diloy, "Detection and Classification of Defective Fresh Excelsa Beans Using Mask R-CNN Algorithm," 2022 14th International Conference on Computer and Automation Engineering (ICCAE), 2022, pp. 97-102, doi: 10.1109/ICCAE55086.2022.9762416.

[11] M. V. C. Caya, R. G. Maramba, J. S. D. Mendoza and P. S. Suman, "Characterization and Classification of Coffee Bean Types using Support Vector Machine," 2020 IEEE 12th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM), 2020, pp. 1-6, doi: 10.1109/HNICEM51456.2020.9400144.

[12] V. Kukreja and D. Kumar, "Automatic Classification of Wheat Rust Diseases Using Deep Convolutional Neural Networks," 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2021, pp. 1-6, doi: 10.1109/ICRITO51393.2021.9596133.

[13] P. Mekha and N. Teeyasuksaet, "Image Classification of Rice Leaf Diseases Using Random Forest Algorithm," 2021 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunication Engineering, 2021, pp. 165-169, doi: 10.1109/ECTIDAMTNCON51128.2021.9425696.

[14] J. Wongbongkotpaisan and S. Phumeechanya, "Plant Leaf Disease Classification using Local-Based Image Augmentation and Convolutional Neural Network," 2021 18th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2021, pp. 1023-1027, doi: 10.1109/ECTI-CON51831.2021.9454672.

[15] Koklu, M. and Ozkan, I.A., (2020), Multiclass Classification of Dry Beans Using Computer Vision and Machine Learning Techniques. Computers and Electronics in Agriculture, 174, 105507

[16] C. H. Mendigoria et al., "Seed Architectural Phenes Prediction and Variety Classification of Dry Beans (Phaseolus vulgaris) Using Machine Learning Algorithms," 2021 IEEE 9th Region 10 Humanitarian Technology Conference (R10-HTC), 2021, pp. 01-06, doi: 10.1109/R10-HTC53172.2021.9641554.

[17] G. Shobana and S. N. Bushra, "Classification of Myopia in Children using Machine Learning Models with Tree Based Feature Selection," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020, pp. 1599-1605, doi: 10.1109/ICECA49313.2020.9297623.

[18] G Shobana, N Priya, A New Multi-Phase Feature Selection Framework for The Prediction of Breast Cancer Drug Using Machine Learning Techniques, Journal of Algebraic Statistics 13 (2), 300-312(2022).



eISSN 1303-5150