[1]Urvashi L. Solanki
[2]Sahistabanu S. Machchhar
[3]Hardi S. Sanghavi

Grain Classification Using Different Machine Learning (ML) Classifier along with Feature Extraction Techniques PCA & LDA

Journal of Electrical Systems

Abstract: - The paper presents a model which does identification of multi grains. To design and test such model, datasets of raisin, dry-beans and rice are used. For the said purpose commonly preferred classifiers Decision Tree, Naïve Bayes & Support Vector Machine (SVM) are opted. The classifiers may accompanied by the feature extraction/reduction techniques (like PCA & LDA) to have better result. This possibility is explored in this paper. The performance of the model designed with Machine Learning (ML) classifiers only and along with feature reduction techniques are analyzed. It is observed that Decision Tree classifier serves better amongst opted ML classifiers but involving PCA/LDA in addition to ML classifier gives the best result. Additionally, the simulation results advocate the proposed model for high efficiency, accuracy and requirement of less computation power.

Keywords: SVM, Decision Tree, Naïve Bayes, PCA, LDA, Feature extraction, Feature reduction

I. INTRODUCTION

Now a day's most of the things are working on a large scale so it's difficult to go manually and complete the task on time that's the reason to make automated model which does the assigned task accurately and in less amount of time. There are number of application of machine learning, one of them is classification. Classification application of ML like disease classification-identify the type of disease in fruits or in leaves [1], breast cancer malignant or not [2], text classification-whether text is spam or not [3], vegetables and fruits classification [1,4], rice classification [5] and so many. Table 1 show literature of few research paper where they have used different classifier for different applications with their performance parameters like Accuracy, Precision, Recall, F1-score etc. This paper aids to applications of machine learning classifiers for the identification of particular class of rice, raisin and dry beans. From UCI machine learning library datasets are used [11-13].

Table 1: Literature Review

Title	Classifier	Remarks	Year
A novel method for vegetable and fruit classification based on using diffusion maps and machine learning [4]	Decision Tree, Naïve bayes, Linear Discriminant Analysis, Support vector machine, Bagging	• Diffusion maps used for the redundant information reduction • SVM classifier achieves 96.25%	2024
Applying Machine Learning Algorithms for the Classification of Sleep Disorders [6]	Support vector machine, K-nearest neighbors, Decision Tree, Random Forest, Artificial Neural Network	• Used Genetic algorithm with ML classifier to improve the accuracy • ANN algorithm get higher accuracy than other	2024
A CNN-Based Hybrid Approach to Classification of Raisin Grains [7]	KNN, Ridge Classifier, XGBoost, SVC and LDA	• Hybrid model created by combining CNN and classifier • Hybrid model increased the success in result	2023

[1]*Asst. Professor, Government Engineering College, Bhavnagar, Gujarat, India. urvashi.solanki1@gujgov.edu.in

[2,3] Asst. Professor, Government Engineering College, Bhavnagar, Gujarat, India.

Analytics of machine learning-based algorithms for text classification [8]	Support vector machine, K-nearest neighbors, Multinomial naïve bayes(MNB), Logistic Regression, Random Forest	• Used two datasetsIMDB dataset SVM and Logistic Regression have 85.5 and 85.9% accuracies, • Spam datasetsupport vector machine outperforms and accuracy is 95.5%	2022
Multiclass classification of dry beans using computer vision and machine learning techniques [9]	Multi-Layer Perceptron, Support vector machine, Decision Tree, K-nearest neighbors	• SVM give higher accuracy 93.13% • 10 fold cross validation used	2020
Rice Grain Classification using Image Processing& Machine Learning Techniques [5]	Logistic Regression, Decision Tree Classifiers, Random Forest Classifier, SVM Algorithm	• extract different parameters of each rice grains • Show the results of LR and SVM	2020
Classification of Raisin Grains Using Machine Vision and Artificial Intelligence Methods [10]	Logistic Regression Multilayer Perceptron Support Vector Machine	• Image preprocessing done to extract the 7 features • SVM give good accuracy 86.44%	2020

II. General Flow Of Classification

First of all images are captured, then preprocessing is used to extract the features from the images. The extracted feature will be given to classifier for training purpose and the model is trained. The trained model is tested for the datasets and results are checked. If results are not promising then different preprocessing techniques are tried or classifier is changed to reach the desired result.

III. ML Classifier & Feature Extraction Techniques

In classification preprocessed features give as input and train the model for classification. Whenever new testing data will come it's give to the trained model and get the target value/result for new data.Amongst many, here three most commonly preferred three classifiers are opted.

A. Decision Tree

Decision tree classifier works on the hierarchical tree kind of architecture. Where tree is created from the given dataset. Attributes are checked to divide dataset into other nodes/sub datasets based on criteria like Gini index, Information gain or entropy. For every created new nodes again the attribute is calculated and this process is repeated unless stopping criteria like either decided depth is reached or all leaf node have same class value is met. Corresponding values for particular node is checked and each node have one and more values base on the value jump to the next node and like this reach at the end node which node contain any one value from the possible classification labels. Used entropy as criteria for classification. Entropy shows the impurity of data.

Entropy(S) = $\sum_{i=1}^{c} -p_i log_2 p_{i+1}$ ..... (01)

B. Support Vector Machine

It uses hyper plane to classify the data. Hyper plan is one kind of plan that separate two classes. The same concept is applied for multiclass in which multiple hyper plans are used to separate the classes. If there's two classes in dataset then required one hyper plan for separation and if 'n' classes then required (n-1) hyper plans for separation. Hyper plan is chosen in such a way that they make more separation between two different classes and classify the data perfectly. Every time, it is not possible to divide the classes linearly so in this situation used kernel function to convert non liner plan to linear plan. Gaussian RBF kernel is used for SVM classification. It is in the form as below.

$K(x_1, x_2) = e^{\frac{-||x_1 - x_2||^2}{2\sigma^2}}$ .....(02)

Where 'σ' is the variance,x_1, x_2 are two data points and $||x_1 - x_2||$ show Euclidean distance

C. Naïve Bayes

This classifier works on probability concepts. Naïve means each feature have their independent probability and Bayes means its use Bayes theorem. Bayes theorem is based on the prior probability, calculates the posterior probability. In naïve bayes classifier, firstly it determines the prior probability for the specified class labels then determines the likelihood probability for each class using each attribute. At last uses the Bayes Formula to enter these values and determine the posterior probability then see which class give higher probability and that is the final class.

The Naive Bayes algorithm is normally used to distribute data which is known as Gaussian Naive Bayes. When working with continuous data, an assumption often taken is that the continuous values associated with each class are distributed according to a normal (or Gaussian) distribution. The likelihood P of the features is assumed to be for $x_i$within y

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right)} \quad \text{..... (03)}$$

Whenever a new data point x is received, the algorithm determines the maximum posterior probability associated with each class and allocates the data point to that class.

The feature extraction/reduction techniques opted are Principal Component Analysis (PCA) & Linear Discriminant Analysis (LDA).

A. PCA

PCA stands for Principal Component Analysis. It is used to reduce the dimension of features by identify 'n' principal components. PCA is unsupervised dimensionality reduction technique because it does not considered class label for process. PCA create orthogonal axes (principal components) shows variance. Total variance of dataset is similar to the combined variance of all principal components.

Steps of PCA are as below:

- Standardization of data. Means standardize the continuous initial variable ranges so that each contributes equally to the analysis.

$$z = \frac{X-\mu}{\sigma} \quad \text{.....(04)}$$

Where μ is mean and σ is standard deviation

- To identify the correlation, calculate covariance matrix. Covariance matrix represent two or more variable correlated with each other.

$$\text{Cov (x,y)} = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{n} \quad \text{.....(05)}$$

Where x_i and y_i data value of x and y respectively, $\bar{x}$ and $\bar{y}$ mean value of x and y respectively

- Calculate eigenvalue and eigenvector from covariance matrix.

$$Av = \lambda v \quad \text{.....(06)}$$

λ shows eigen value and v shows eigen vector

- Create new feature set by selecting first n principal components (Eigen vectors). The first component (high variance) give more information than second one and second one give more information than third one and so on.

B. LDA

LDA stands for Linear Discriminant Analysis. It reduces the number of features from the labels provided to them. It is a supervised learning dimensionality reduction algorithm. LDA forms a new axis from existing one. To maximize the separation between the classes, LDA projects the data into a space with a lower dimension. To do this, a collection of linear discriminants that maximizes the variance ratio within a class to variance between classes is identified. It determines which feature space directions best divide the various data classes. The goal of LDA is to minimize overlap across each class and identify the best possible straight line or plane for dividing these groups.

Steps of LDA as below:

- Calculate the mean vectors for the individual classes.

$$m = \frac{1}{N}\Sigma_{i=1}^{n} x_i \quad \text{.....(07)}$$

Where m and x_i represents mean and individual data respectively

- Calculate intra-class and inter-class scatter matrices. Where S_w represents within (intra) class scatter and S_b represents between (inter) class scatter and $S_k$represents scatter for each classes.

$S_k = \Sigma_{x(n)\epsilon C_i}(x(n) - m_i)(x(n) - m_i)^T$…….(08)

$S_w = \Sigma_{i=1}^{C} S_k$……..(09)

$S_b = \Sigma_{i=1}^{C} N_i(m_i - m)(m_i - m)^T$ *……(10)*

- Calculate eigenvectors(W) for intra-class and inter-class scatter

$W = eig(S_w^{-1}S_b)$…….(11)

- The desired eigenvectors are finally obtained from the associated eigenvalues by Eigen decomposition of$S_w^{-1}S_b$. The maximum value of total eigenvalues is C-1.

IV. PROPOSED METHOD

PCA and LDA can be used for feature extraction as well as dimensionality reduction purpose. When more number of features exist at that time its required to convert them into less number because that features will be trained by model and if number of features are more than model take more time to train the model. Another advantage of feature extraction is that it gives important features for consideration. Hence it is recommended to use feature reduction technique along with ML classifier for the better result. Fig. 1 represents the proposed model which classifies the grains.

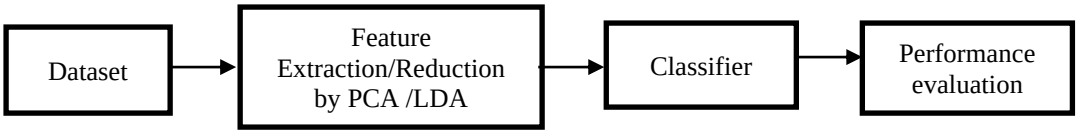


Fig. 1: Proposed model for Performance improvement

V. SIMULATION AND RESULTS

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper.

The models that does classification of grains, using ML classifiers alone and along with feature reduction techniques are simulated in Python. The model can be designed for the following three possibilities.

1. *Classifier Results without PCA/LDA*
2. *Classifier Results with Feature extraction/reduction by PCA*
3. *Classifier Results with Feature extraction/reduction by LDA*

The required datasets are fetched from UCI machine learning library. Table 1 shows the detail of all three dataset. All dataset feature type is either integer or real, so there is no need to encode the features.

Table 1: Datasets

	Instances	**Features**	**Target Value**	**Ref. Link**
Raisin	900	7	2	[11]
Rice(Cammeo and Osmancik)	3810	7	2	[12]
Dry Beans	13611	16	7	[13]

For classification, first step is to divide the dataset into the two part - train and test. Here it is 80 % and 20 % of the dataset, respectively.

To check the performance of model in term of accuracy, the formula used is as below:

Accuracy= $\frac{TP+TN}{P+N}$ ……..(12)

Where

TP: true positive represents number of instances correctly classified,

TN: True Negative represents number of instances correctly negatively classified,

P: Actual positive labeled instances, N: Actual negative labeled instances

A. Classifier Results Without PCA/LDA

The accuracy of the models designed using the opted ML classifiers separately are listed in Table 2.

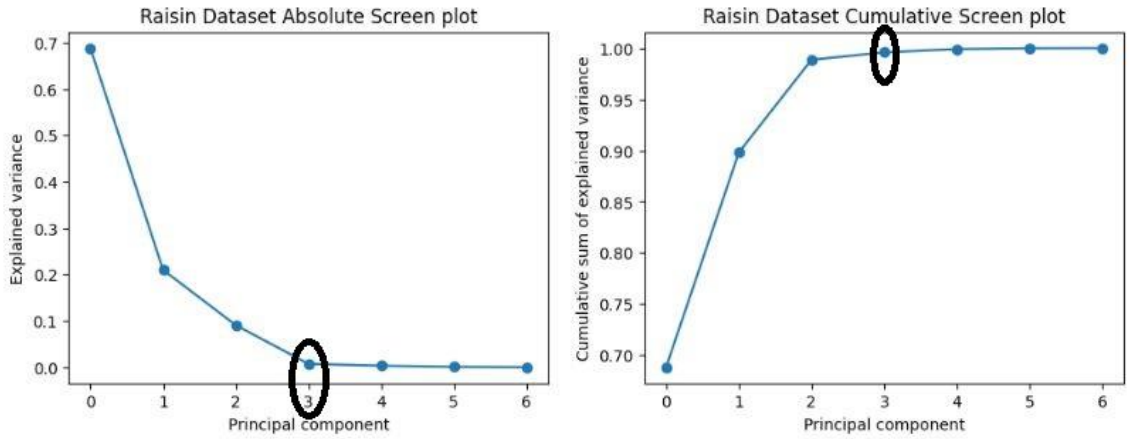
Table 2: Accuracy (%) of classifier without feature extraction/reduction

Dataset	Support Vector Machine (SVM)	Decision Tree	Naïve Bayes
Raisin	49.26	81.85	81.11
Rice	56.61	89.23	91.86
Dry beans	25.88	89.42	76.52

Analyzing the accuracies, it is found that the performance of SVM is deteriorated. Decision Tree performs steady and Naïve Bayes is well for the datasets of raisin and rice.

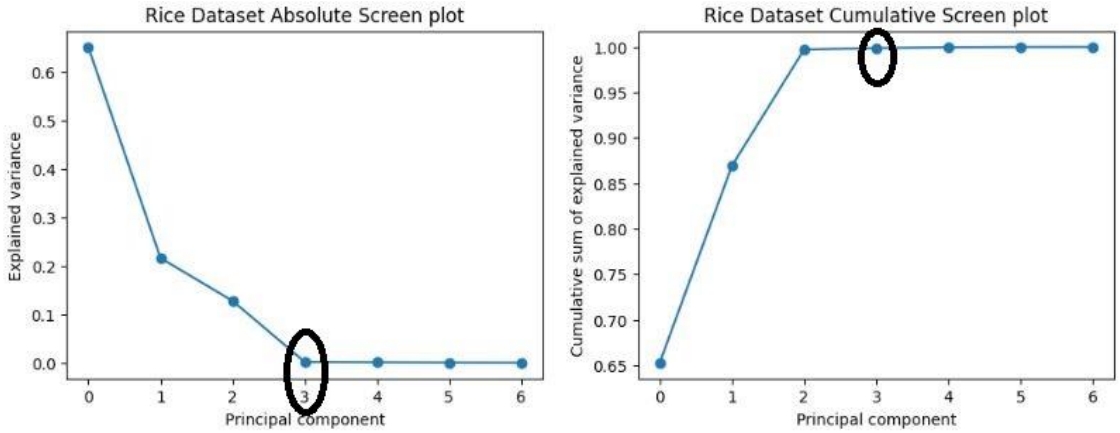
B. Classifier Results with Feature extraction/reduction by PCA

Fig.2, 3 & 4 presents selection of principal components for Raisin, Rice and Dry beans datasets, respectively. Fig. 2, 3 & 4 comprises of absolute screen plot and cumulative screen plot of variance for the datasets. Absolute Screen Plot show individual component variance, while Cumulative screen plot shows summation of variance at particular component. Using these plots value of principal component 'n' can be decided. For example, in Raisin dataset, total number of features is 7. Referring to fig.2 (a), first principal component gives 70% variance, second component gives 20% and third component gives 10% and cumulatively its 100%, refer fig. 2(b). It means it covers whole dataset in only 3 reduced feature components. Hence, the principal component value could be taken 3 (n = 3). Similarly for Rice and dry beans datasets it gets n=3 and n=6, respectively.
Accuracy of the model designed by using the numbers of features, supplied by the PCA, and by the opted ML classifiers is calculated and listed in Table 3.



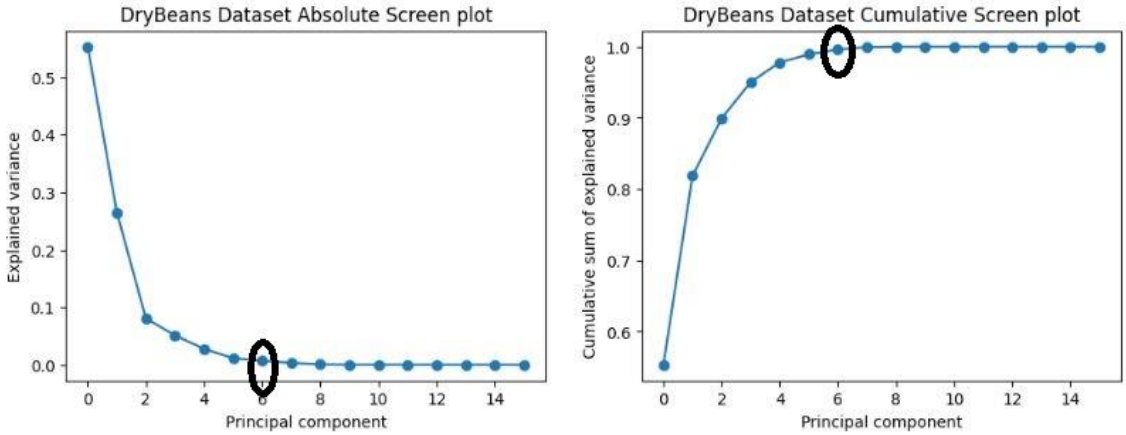
(a) Raisin Dataset Absolute Screen Plot (b) Raisin Dataset Cumulative Screen Plot

Fig. 2: Principal component selection for Raisin dataset



(a) Rice Dataset Absolute Screen Plot (b) Rice Dataset Cumulative Screen Plot

Fig. 3: Principal component selection for Rice dataset



(a) Drybeans Dataset Absolute Screen Plot (b) Drybeans Dataset Cumulative Screen Plot

Fig. 4: Principal component selection for Dry beans dataset

Table 3: Classifier Accuracy (%) when Feature extraction/reduction by PCA

Dataset	PCA 'n' components	Support Vector Machine	Decision Tree	Naïve Bayes
raisin	n=3	90.00	79.44	87.22
rice	n=3	93.31	89.63	93.04
Dry beans	n=6	92.25	89.39	90.08

C. Classifier Results with Feature extraction/reduction by LDA

In LDA number of 'n' component is chosen based on eigenvalue. At most possible eigenvalue is C-1, where 'C' is the target value. Target value for Raisin, Rice and Dry bean datasets are 2, 2, & 7, respectively. Therefore, the possible eigenvalue and hence number of 'n' component for the said datasets could be 1, 1 and 6. Considering the number of 'n' component supplied by LDA, the model is designed using ML classifier and their efficiency for the datasets are listed in Table 4.

Table 4: Classifier Accuracy (%) with Feature extraction/reduction by LDA

Dataset	LDA 'n' components	Support Vector Machine	Decision Tree	Naïve Bayes
Raisin	n=1	90.56	86.11	90.00
Rice	n=1	93.83	89.63	93.96
Dry beans	n=6	91.88	89.97	91.48

Table 5: Best Result for three datasets

Dataset	Feature Extraction	Classifier	Accuracy (%)
Raisin	LDA	Support Vector Machine	90.56
Rice	LDA	Naïve Bayes	93.96
Dry beans	PCA	Support Vector Machine	92.25

From Table 2, 3 & 4 best possible result for individual database is picked out and showcase in Table 5. Analyzing accuracy listed in table 2, 3, 4 and 5 it is noticed that the model designed using ML classifiers only are possessed reasonable accuracy but when feature extraction techniques are incorporated with ML classifier the accuracy of the model is hinged above 90 %.

VI. CONCLUSION

It is observed the Decision Tree and Naïve Bayes classifier performs remarkable but the performance of the model is boosted when PCA/LDA feature extraction/reduction technique is inculcated with ML classifiers. PCA/LDA extract important features from dataset and reduce the dimensionality of dataset, hence may also assure less computational power requirement. Therefore, for such type of applications it is advisable to involve feature reduction techniques along with ML classifiers for better results. The proposed model can be extended to the medical field like classification of diseases and many more.

REFERENCES

[1] Solanki, U., Jaliya, U. K., & Thakore, D. G. (2015). A survey on detection of disease and fruit grading. International Journal of Innovative and Emerging Research in Engineering, 2(2), 109-114.

[2] S. Ara, A. Das and A. Dey, "Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms," 2021 International Conference on Artificial Intelligence (ICAI), Islamabad, Pakistan, 2021, pp. 97-101, doi: 10.1109/ICAI52203.2021.9445249.

[3] A. Sinha, M. N. B. J. Naskar, M. Pandey and S. S. Rautaray, "Text Classification Using Machine Learning Techniques: Comparative Analysis," 2022 OITS International Conference on Information Technology (OCIT), Bhubaneswar, India, 2022, pp. 102-107, doi: 10.1109/OCIT56763.2022.00029.

[4] Wenbo Wang, Aimin Zhu, Hongjiang Wei, Lijuan Yu,A novel method for vegetable and fruit classification based on using diffusion maps and machine learning,Current Research in Food Science,Volume 8,2024,100737,ISSN 2665-9271,https://doi.org/10.1016/j.crfs.2024.100737.

[5] B. Arora, N. Bhagat, L. R. Saritha and S. Arcot, "Rice Grain Classification using Image Processing & Machine Learning Techniques," 2020 International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2020, pp. 205-208, doi: 10.1109/ICICT48043.2020.9112418.

[6] T. S. Alshammari, "Applying Machine Learning Algorithms for the Classification of Sleep Disorders," in IEEE Access, vol. 12, pp. 36110-36121, 2024, doi: 10.1109/ACCESS.2024.3374408.

[7] Yılmaz, E. K., Oğuz, T., & Adem, K. (2023). A CNN-Based Hybrid Approach to Classification of Raisin Grains. International Conference on Frontiers in Academic Research, 1, 460–467. Retrieved from https://as-proceeding. Com /index.php /icfar /article /view /147.

[8] Sayar Ul Hassan, Jameel Ahamed, Khaleel Ahmad,Analytics of machine learning-based algorithms for text classification, Sustainable Operations and Computers, Volume 3, 2022,Pages 238-248,ISSN 2666-4127,https://doi.org/10.1016/j.susoc.2022.03.001.

[9] Murat Koklu, Ilker Ali Ozkan,Multiclass classification of dry beans using computer vision and machine learning techniques,Computers and Electronics in Agriculture,Volume 174, 2020, 105507, ISSN 0168-1699, https://doi.org/10.1016/j.compag.2020.105507.

[10] Cinar, Ilkay & Koklu, Murat & Tasdemir, Sakir. (2020). Classification of Raisin Grains Using Machine Vision and Artificial Intelligence Methods.10.30855/gmbd.2020.03.03.

[11] Raisin dataset: https://archive.ics.uci.edu/dataset/850/raisin.

[12] Rice dataset https://archive.ics.uci.edu /dataset /545 /rice +cammeo +and + osmancik.

[13] Drybean dataset: https://archive.ics.uci.edu/dataset/602/dry+bean+dataset