Forage Res., **50**(4) : pp. 436-440 (2025) **http://forageresearch.in**

CLASSIFICATION OF FORAGE SORGHUM GENOTYPES USING DISCRIMINANT ANALYSIS

PRATIBHA BHARTI[1]*, SARITA RANI[1], AJAY SHARMA[1], PUMMY KUMARI[2]

[1]Department of Mathematics and Statistics, CCS Haryana Agricultural University, Hisar-125 004 (Haryana), India
[2]Deparment of G&PB (Forage section), CCS Haryana Agricultural University, Hisar-125 004 (Haryana), India
**(e-mail: pratibha595@gmail.com)*

(Received : 07 March 2025; Accepted : 30 March 2025)

SUMMARY

The present study was carried out for classifying and predicting classes for yield and protein content of multicut forage sorghum genotypes using discriminant analysis, based on performance measures derived from a confusion matrix. Secondary data of 117 genotypes of multicut forage sorghum, along with two checks measured for 15 traits, was used in this study. The genotypes were grouped into two categories, G1 (low) and G2 (high), under two grouping schemes (GS I and GS II) across two datasets: 1st cut and 2nd cut. Classification and prediction results were obtained for both training and testing datasets. A confusion matrix was generated from the testing data to classify and predict classes based on fodder yield and protein content. The highest accuracy percentage (85.7%) was achieved in grouping scheme GS I for green fodder yield (GFY) in the testing dataset of the 1st cut, demonstrating the effectiveness of discriminant analysis in accurate classification and prediction.

Key words: Discriminant analysis, performance measures, green fodder yield, accuracy, confusion matrix

Sorghum [*Sorghum bicolor* (L.) Moench] belonging to family *Poaceae* is a versatile crop with multiple uses like food, fodder, feed, biofuel and other industrial uses. It is an important climate resilient crop grown in semi-arid tropics due to its drought tolerant nature. In India, it is used mainly as food and fodder crop due to its high grain yield and biomass yield with good nutritional composition of fodder. Livestock security mainly depends on the availability of quality green, dry fodder and concentrates. Although, India is home to 535.78 million livestock (Anonymous, 2019), but the country is deficit in fodder availability. The green and dry fodder is deficit to the tune of 11.24 and 23.40 percent respectively (Roy *et al.*, 2021). The major constraints for low production and productivity of fodder in India is the scarcity of improved varieties of forage crops with good fodder yield and quality to the farmers and also less acreage is allotted to forage crops. Sorghum is an important fodder crop in India. In forage sorghum, absence of improved genotype resulted in 39 % losses in the productivity as compared to full package of practices (Satpal *et al.*, 2021). This means availability of an improved genotype plays a very critical role in ensuring high fodder yields.

Discriminant analysis constructs a classification rule based on a training sample to assign new observations to one of several predefined classes. A widely used and fundamental method in this domain is Linear Discriminant Analysis (LDA), which is commonly implemented using Fisher's Linear Discriminant Function (LDF). This approach utilizes the sample means and covariance matrices of different classes to derive an optimal discriminant function. Originally introduced for binary classification problems (Fisher, 1936), Fisher's method was later generalized to multiple classes through Multiple Discriminant Analysis (Rao, 1948). Despite the development of more advanced nonlinear classification techniques, Fisher's method remains widely employed due to its simplicity, interpretability, and robust performance across various datasets (Croux *et al.*, 2008). Moreover, the Fisher LDF, as a linear combination of measured variables, facilitates straightforward interpretation while maintaining competitive classification accuracy in many applications. Elfadl and Abdallah (2017) even applied discriminant analysis to classify and predict the fertility status of Friesian cattle, demonstrating its effectiveness in handling biological classification problems. The study showed that discriminant analysis could provide accurate and interpretable results in livestock fertility assessment also.

MATERIALS AND METHODS

The present study for the classification of forage sorghum genotypes using discriminant analysis was conducted at the Department of Mathematics & Statistics, CCS Haryana Agricultural University, Hisar (Haryana), India during 2022-24 using the secondary data. For the study, the secondary data of 117 genotypes of multicut forage sorghum was obtained from the experiment conducted by the Forage Section of Department of Genetics & Plant Breeding, CCSHAU, Hisar during *kharif* season. Classifying and predicting classes for yield and protein content of multicut forage sorghum genotypes using discriminant analysis, based on performance measures derived from a confusion matrix.The data were obtained for the following 15 characters *viz*., Early vigour score (EV), Plant height (cm) (PH), Number of tillers per plant (TP), Number of leaves per plant (LP), Leaf length (cm) (LL), Leaf breadth (cm) (LB), Stem girth (cm) (SG), Leaf stem ratio (LSR), Plant population per meter row length (PP), Regeneration score (RG), Green fodder yield(q/ha) (GFY) , Dry fodder yield (q/ha) (DFY) , HCN content on fresh weight basis (ppm) (HCN) , Crude protein (%) (PC), *In-vitro* dry matter digestibility (%) (DMD). The 117 genotypes of multicut forage sorghum (1st cut and 2nd cut dataset) were divided into two groups (G1 and G2) according to the two schemes, *viz*. GS I and GS II.

Grouping scheme I

G1 (Low yielding group): Fodder yield < [mean-(standard deviation/2)]
G2 (high yielding group): Fodder yield ≥ [mean+(standard deviation/2)]

Grouping scheme II

G1 (Low protein group): Protein content < [mean-(standard deviation/2)]
G2 (High protein group): Protein content ≥ [mean+(standard deviation/2)]

Pre-processing of data

The analysis began with the construction of correlation matrix for 1st cut and 2nd cut dataset of forage sorghum to check the association between different morphological variables. The problem of testing whether a sample comes from a normal population has been studied by much generation of statisticians. Kolmogorov-Smirnov (KS) test (Kolmogorov, 1933) is widely used and based on the maximum difference between the observed distribution and expected cumulative normal distribution. The null hypothesis assumes that the dataset follows a normal distribution. The K-S test was performed for each variable under study at 0.05% level of significance. Now equality of group means was tested using D-square test and the assumption of homogeneity of variance- covariance matrices was tested using Box's M test. This test evaluates whether the variance-covariance matrices of the groups under study are equal.

Discriminant Analysis (DA)

Fisher's Linear Discriminant Analysis (LDA) was employed as a classification technique to distinguish among predefined groups of forage sorghum genotypes. LDA constructs a linear combination of predictor variables that best separates the classes by maximizing the ratio of between-group variance to within-group variance. The analysis was performed using standardized input features to ensure comparability. The classification rule was derived from a training dataset, utilizing group means and a pooled within-group covariance matrix to compute discriminant functions. These functions were then

TABLE 1
Mean, standard deviation and coefficient of variation for multicut forage sorghum variables

Variable	1st cut			2nd cut		
	MEAN	SD	CV	MEAN	SD	CV
EV	3.35	0.85	25.49	-	-	-
PH	142.29	21.62	15.19	70.63	28.21	39.95
TP	3.42	1.55	45.20	1.84	0.69	37.43
LP	23.45	8.69	37.07	16.76	5.33	31.80
LL	75.62	9.19	12.15	58.19	8.51	14.62
LB	5.68	1.59	27.93	4.31	0.69	16.01
SG	1.56	0.57	36.23	1.80	0.43	23.89
LSR	0.42	0.10	23.40	0.36	0.07	19.46
PP	10.15	2.50	24.67	2.88	1.90	65.69
RG	-	-	-	2.04	0.99	48.37
GFY	667.75	185.19	27.73	45.51	37.13	81.58
DFY	139.44	42.19	30.26	9.52	8.71	91.51
HCN	139.71	70.37	50.37	142.33	73.57	51.69
PC	10.50	0.85	8.08	9.28	0.80	8.58
DMD	51.07	3.54	6.93	46.27	5.21	11.25

TABLE 2
Standardized coefficients obtained from discriminant analysis

Scheme	Variables	Discriminant function
1st cut		
GS I GFY	TP, LP, LL, PP	D=0.323TP-0.050LP-0.006LL+0.479PP
GS II PC	PH, LL, LB, SG	D=-0.019PH-0.009LL+0.271LB-1.735SG
2nd cut		
GS I GFY	PH, TP, LP, LL, LB, SG, PP, RG, PC	D=0.051PH+0.138TP-0.025LP-0.013LL-0.075LB-0.387SG+0.968PP-0.01 RG+0.003PC
GS II PC	TP, LP, SG, LSR, HCN	D=-0.172TP+0.041LP-1.795SG+12.009LSR+0.006HCN

applied to classify observations in the test set. The performance of the model was evaluated using classification accuracy and confusion matrices.

Discriminant analysis functional form as follows:

$$DF=V_1X_1+V_2X_2+ V_3X_3+---V_iX_i$$

Where V_1, V_2...V_i are diagnosis coefficients and X_1, X_2...X_i are independent variables. Discriminant analysis is a multivariate technique which focuses on association between categorical dependent variables and multiple independent variables. Simplest form of discriminant analysis is when dependent variable is dichotomous, in this case discriminant function use to classify genotypes into two groups.

RESULTS AND DISCUSSION

Secondary data on 117 genotypes of multicut forage sorghum were divided into two groups (G1 and G2) according to the three schemes, *viz*. scheme-I, II. The D-square (D^2) test was used in discriminant analysis for evaluating the difference between two groups mean based on multiple variables. Significance value found less than 0.05 level of significance for both grouping schemes and two dataset so the null hypothesis was rejected at the 5% level of significance for the equality of mean vectors and concludes that the mean vectors of two groups were not equal. Then significant variables

TABLE 3
Performance of DA for classification of forage sorghum genotypes on the basis of grouping schemes

Dataset	Grouping Scheme		Training		Testing	
			Accuracy	Kappa	Accuracy	Kappa
1st cut	GS I	GFY	0.855	0.853	0.857	0.848
	GS II	PC	0.844	0.842	0.850	0.843
2nd cut	GS I	GFY	0.745	0.741	0.769	0.755
	GS II	PC	0.852	0.850	0.875	0.868

TABLE 4
Classification and prediction of classes for yield and protein content of forage sorghum genotypes using discriminant analysis with confusion matrix and performance measures for 1st and 2nd cut dataset

Performance statistics		1st cut dataset				2nd cut dataset			
		Predicted							
		GS I		GS II		GS I		GS II	
		GFY		PC		GFY		PC	
		L	H	L	H	L	H	L	H
Observed	L	4	1	8	2	5	2	9	1
	H	1	8	1	9	1	5	1	5
n*		14		20		13		16	
Sensitivity		0.889		**0.900**		0.833		0.833	
Specificity		0.800		0.800		0.714		0.900	
Positive predictive values		**0.889**		0.818		0.714		0.833	
Negative predictive values		0.800		**0.889**		0.833		0.900	
Balanced Accuracy		0.844		0.850		0.774		0.867	
F-measure		**0.889**		0.857		0.769		0.833	

were selected using independent sample t-test for applying discriminant analysis technique. In GS I, based on univariate t-test, variables tillers per plant, leaf per plant, leaf length and plant population were found to have significant differences in the two group means and early vigour, plant height, leaf breadth, stem girth, leaf stem ratio, HCN, crude protein and dry matter digestibility are found least discriminatory variables. In GS II, variables plant height, leaf length, leaf breadth and stem girth were found to have significant differences in the two group means and early vigour, tillers per plant, leaf per plant, leaf stem ratio, plant population, green fodder yield, dry fodder yield, HCN, dry matter digestibility were found least discriminatory variables.

After application of discriminant analysis technique using R software, for 1st cut dataset in GSI, leaf per plant and leaf length for green fodder yield showed negative standardized coefficient while rest of the variables tiller per plant and plant population showed the positive standardized coefficient. In GS II, plant height, leaf length, stem girth showed the negative standardized coefficient and leaf breadth showed the positive standardized coefficient (Table 2). Confusion matrix obtained for both grouping schemes and datasets are in Table 3 and 4.

The results presented in Table 3 indicate that the performance of Discriminant Analysis (DA) varies with the choice of grouping scheme and the cut stage of the forage sorghum. For the 1st cut, both grouping schemes-based on green fodder yield (GFY) and protein content (PC)-demonstrated comparable classification accuracy and kappa values, with a slight advantage observed for the GFY-based scheme (Training Accuracy = 0.855; Testing Accuracy = 0.857). However, in the 2nd cut, a marked improvement in performance was noted when using PC-based grouping (Training Accuracy = 0.852; Testing Accuracy = 0.875), as compared to the GFY-based scheme (Training Accuracy = 0.745; Testing Accuracy = 0.769). According to Table 4 the discriminant analysis showed satisfactory classification performance for both 1st and 2nd cut datasets. For the 1st cut, both grouping schemes (GFY and PC) achieved balanced accuracy above 0.84, with slightly better sensitivity and F-measure under the GFY-based grouping. In the 2nd cut, the PC-based grouping outperformed GFY-based grouping across most metrics, particularly in specificity, negative predictive value, and balanced accuracy (0.867). Overall, the results indicate that grouping based on protein content (PC) is more effective in the later cut stage, while both schemes perform comparably in the early stage. Twinkle (2022) also classified 310 Indian mustard genotypes using Discriminant Analysis (DA) and achieved 85.00 and 86.11 per cent accuracy for seed yield and days to maturity, respectively. Bishnoi et al. (2022) also used linear discriminant analysis for classification of 452 cotton genotypes and achieved competitive accuracy.

CONCLUSION

These findings suggest that the effectiveness of DA was influenced not only by the variable used for grouping but also by the developmental stage of the crop, with PC-based grouping providing superior classification performance in the later stage. Overall, DA shows strong potential for genotype classification, particularly when appropriate grouping criteria are selected. For the 1st cut of forage sorghum, both grouping schemes (GFY and PC) achieved balanced accuracy above 0.84, with slightly better sensitivity and F-measure under the GFY-based grouping. In the 2nd cut, the PC-based grouping outperformed GFY-based grouping across most metrics, particularly in specificity, negative predictive value, and balanced accuracy (0.867). Overall, the results indicated that grouping based on protein content (PC) was more effective in the later cut 2nd stage, while both schemes perform comparably in the early stage (1st cut).

ACKNOWLEDGEMENTS

The authors are thankful to The Head, Department of Mathematics & Statistics, College of Basic Sciences & Humanities, CCS Haryana Agricultural University, Hisar, India and The Head, Forage Section, Department of Genetics & Plant Breeding, CCS Haryana Agricultural University, Hisar, India for providing the required facilities used in the study.

REFERENCES

Anderson, T. W., 1984 : *An introduction to multivariate statistical analysis* (2nd Edn.). Wiley.

Anonymous 2018 : https://www.fao.org/news/archive/news-by-date/2018/en/.

Anonymous 2019 : 20th Livestock Census - 2019. http://dahd.nic.in/about-us/divisions/statistics.

Bishnoi, S., N. Al-Ansari, M. Khan, S. Heddam, A. Malik, 2022 : Classification of cotton genotypes with mixed continuous and categorical variables:

application of machine learning models. *Sustainability,* **14**: 13685.

Croux, C., P. Filzmoser and K. Joossens, 2008: Classification efficiencies for robust linear discriminant analysis. *Statistica Sinica,* **18**(2): 581-599.

Das, B., R. N. Sahoo, A. Biswas, S. Pargal, G. Krishna, R. Verma, V. Chinnusamy, V. K. Sehgal, and V. K. Gupta, 2018 : Discrimination of rice genotype using field spectroradiometry. *Geocarto International,* **35**(1): 64-77.

Elfadl, E. A. A. and F. D. M. Abdallah, 2017 : Using Discriminant Analysis and Artificial Neural Network Models for Classification and prediction of fertility status of friesian cattle. *American Journal of Applied Mathematics and Statistics,* **5**(3): 90-94.

Fisher, R. A., 1936: The use of multiple measurements in taxonomic problems. *Annals of Eugenics,* **7**(2): 179-188. *https://doi.org/10.1111/j.1469-1809. 1936.t*

Godara, P., S. Verma, S. Kumari and S. Kumar, 2022 : Importance of variable using gini index and discriminant score in Indian mustard genotypes. *Journal of Agriculture Research and Technology,* **Special Issue (1)**: 100-105.

Rao, C. R., 1948 : The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society: Series B (Methodological),* **10**(2): 159-203.

Roy, A. K., R. K. Agrawal, N. R. Bhardwa, A. K. Misra, and S. K. Mahanta, 2021 : Indian Forage Scenario – Region Wise Availability and Deficit. IGFRI, Jhansi, India.

Satpal, S. Kumar, A. Kumar, B. Gangaiah, K.K. Bhardwaj, and Neelam, 2021 : Evaluation of energy efficiency and optimum resource management in forage sorghum [*Sorghum bicolor* (L.) Moench] under semi-arid tropics. *Forage Res.*, **47**(3): 308-312.