XIII Argentine Congress of Bioinformatics and Computational Biology

XIII International Conference of the Iberoamerican Society of Bioinformatics

III Annual Meeting of the Ibero-American Artificial Intelligence Network for Big BioData

NEW SET OF CLASSES FOR FRUIT SHAPE CLASSIFICATION IN TOMATO BASED ON MACHINE LEARNING

Vazquez, Dana V.[1,2]; Spetale, Flavio E.[3,4]; Tapia, Elizabeth [3,4]; Rodríguez, Gustavo R.[1, 2]

[1]Cátedra de Genética. Facultad de Ciencias Agrarias [2]IICAR-CONICET-UNR. [3]Facultad de Ciencias Exactas, Ingeniería y Agrimensura, UNR, Rosario, Argentina. UNR, Zavalla, Argentina. [4] CIFASIS-CONICET-UNR.

INTRODUCTION

- Tomato (*Solanum lycopersicum* L.) is the second most consumed global vegetable. Fruit shape significantly impacts on yield, quality, consumer preference, and commercial usage.
- Despite of the digital advancements in precision agriculture, the determination of fruit shape still relies predominantly on visual assessment, and there are no standardized approaches.
- Classification criteria often vary among experts, and exist four for tomato: "Rodriguez2011," "Visa2014," "UPOV," and "IPGRI". They define eight, nine, ten, and eight classes, respectively. Unfortunately, these classifications do not present consensus which limits the genetic understanding of shape determinants.

OBJECTIVE

To develop a machine-learning model for automated tomato shape classification and establish a "gold standard"

MATERIALS AND METHODS

PLANT MATERIAL

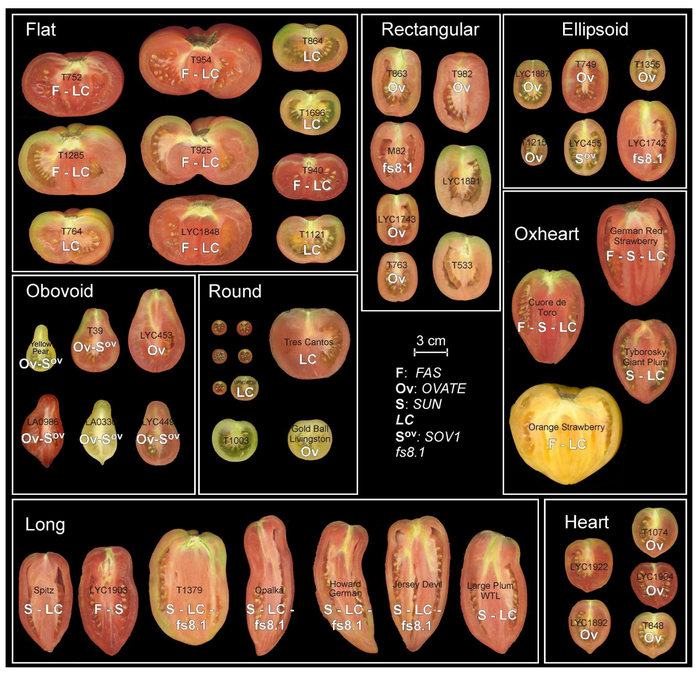

Figure 1: Representative longitudinal-section fruit images from Rodríguez et al. (2011b). Each fruit is identified by the name of the accesion and the presence of mutation in the main shape genes. Scale: 1cm

Longitudinal-section cultivar images (Figure 1) corresponding to 368 tomato accessions were downloaded from https://solgenomics.net and split into individual fruit images (n=1124). The images were evaluated using the Tomato Analyzer 3.0 program to assess fruit morphological traits, resulting in a total of 41 traits. The images were visually categorized into different morphological classes based on four available systems: UPOV, IPGRI, "Rodriguez2011," and "Visa2014" (Figure 2). Additionally, a new set of classes was introduced, where the rectangular class from the Rodriguez2011 method was merged with the ellipsoid class, creating a new class called "newclass," resulting in a total of five datasets

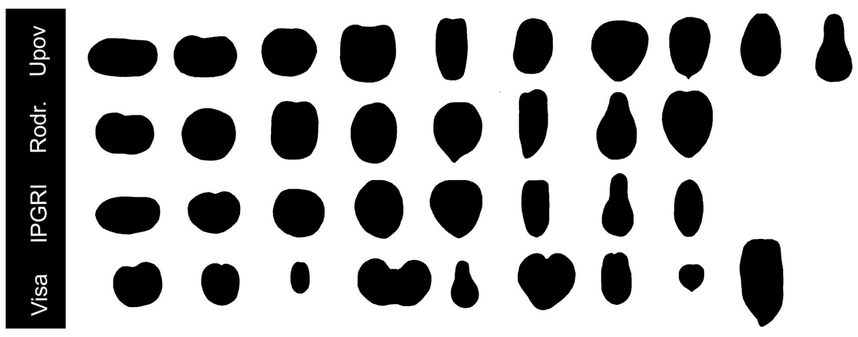

Figure 2 Representative scheme of fruit shape classes on longitudinal section for the different systems available

METHODS

The numeric variables were normalized using a z-score approach, and highly correlated variables (above 0.95) were removed. The data was split into 80% for training and 20% for testing. We applied Recursive Feature Elimination (RFE) as a feature selection technique using the Support Vector Machine model.

The supervised classification methods employed included multinomial logistic regression (MLR), random forest (RF), and support vector machine (SVM). A five-fold cross-validation was performed, considering different data sets for training and testing. Quality metrics for the various models and classifications were assessed. Significant differences between methods and models were evaluated after cross-validation using the Kruskal-Wallis test, and the mean differences between pairs of samples were compared with the Wilcoxon-Mann-Whitney (WMW) test.

RESULTS

Four highly correlated (>0.95) variables were removed. By RFE method 12 variables that were common across all methods were kept.

The mean accuracy values in our study ranged from 0.69 to 0.85, while the standard deviation (SD) values varied between 0.01 and 0.03 (Table 1). The lowest accuracy value was observed in the case of the MLR model applied to the UPOV dataset, while the highest mean accuracy was achieved by the SVM model with the new set of classes. When considering the different datasets, we did not detect significant differences in mean accuracy across the models (Figure 3A, Table 1) at a 5% significance level. However, substantial differences were observed among the datasets ($p < 0.01$) for all models.

Table 1: Mean and standard desviation (sd) for accuracy values of differents machine learning models.

	Multinomial Logistic Regresion	Support Vector Machine	Random Forest	
	mean ± sd	mean ± sd	mean ± sd	pvalue(F)
Rodriguez2011	0.78 ± 0.03	0.80 ± 0.03	0.79 ± 0.02	0.23
Visa2014	0.79 ± 0.03	0.79 ± 0.02	0.79 ± 0.02	0.76
UPOV	0.69 ± 0.03	0.72 ± 0.02	0.70 ± 0.02	0.37
IPGRI	0.72 ± 0.02	0.72 ± 0.02	0.73 ± 0.02	0.65
newclas	0.83 ± 0.02	0.85 ± 0.01	0.85 ± 0.01	0.22
pvalue(F)	<0.001	<0.001	<0.001	

pvalue(F): significance level according to Kruskal-Wallis test

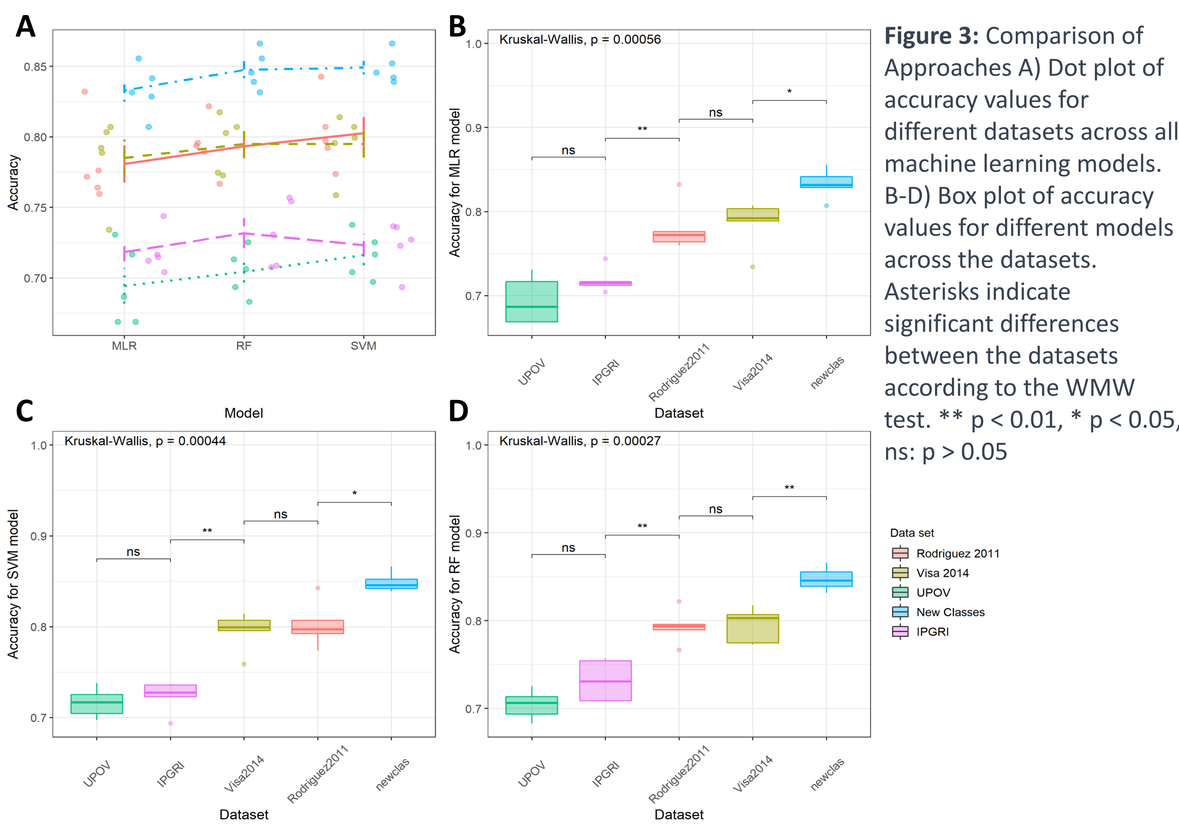

Figure 3: Comparison of Approaches A) Dot plot of accuracy values for different datasets across all machine learning models. B-D) Box plot of accuracy values for different models across the datasets. Asterisks indicate significant differences between the datasets according to the WMW test. ** $p < 0.01$, * $p < 0.05$, ns: $p > 0.05$

WMW test showed that mean accuracy for UPOV and IPGRI was not significantly different and exhibited the lowest values, Rodriguez2011 and Visa2014 showed no significant differences for accuracy and intermediate values, and the novel set of classes yielded the highest mean accuracy values across all four models, i.e., 85% (Frigure 3, Table 1)

CONCLUSION

A new classification system for fruit shapes, including seven categories (flat, round, ellipsoid, heart, oxheart, obovoid, and long), has significantly improved accuracy, achieving an 85% success rate. This "gold standard" for fruit shape facilitates precise tomato cultivar description and consensus among researchers, aiding genetic understanding.

REFERENCES

IPGRI. (1996). Descriptors for Tomato (Lycopersicon spp.). International Plant Genetic Resources Institute.

Rodríguez, G. R et al. (2011). Distribution of SUN, OVATE, LC, and FAS in the Tomato Germplasm and the Relationship to Fruit Shape Diversity. Plant Physiology, 156(1), 275–285.

UPOV. (2001). Guidelines for the conduct of tests for distinctness, uniformity and stability (Tomato).

Visa, S. et al. (2014). Modeling of tomato fruits into nine shape categories using elliptic fourier shape modeling and Bayesian classification of contour morphometric data. Euphytica, 200(3), 429–439.