IECN 2024 Conference
The 4th International Electronic Conference on Nutrients
16–18 October 2024 | Online
MDPI

Machine learning and taxonomy of *Silene* L. species

***Anton Soria – Lopez*[1], *María García – Martí*[2], *Juan C. Mejuto*[1], *Jesus Simal – Gandara*[2], *Gonzalo Astray*[1]**

1. Universidade de Vigo, Departamento de Química Física, Facultade de Ciencias, 32004 Ourense, España

2. Universidade de Vigo, Departamento de Química Analítica y Alimentaria, Facultade de Ciencias, 32004 Ourense, España

INTRODUCTION & AIM

The identification of key morphological features, especially seeds, remains very useful in *Silene* taxonomy, since some *Silene* species have been considered an interesting source of nutraceutical compounds due to their medicinal properties including anti-enzymatic, antioxidant, and antimicrobial effects [1]. In fact, recognition of taxonomic groups based on seed morphology contributes to a better understanding of the diversity and identification of these species. Machine learning allows computer systems to learn automatically from experience without having to be explicitly programmed [2]. These kind of approaches could be interesting to evaluate their performance on identifying *Silene* L. species using seeds' morphological data obtained from literature.

Therefore, the aim of this study is to use three different machine learning algorithms, random forest (RF), support vector machine (SVM) and artificial neural network (ANN), such as innovative methodologies to evaluate their effectiveness to identify *Silene* L. seeds using the data reported in the research carried out by Martín-Gómez *et al.* [3].

METHOD

The database from Martín-Gómez *et al.* (2022) [3] contains 8 input variables based on the geometric seeds data (perimeter, length, width, area, aspect ratio, circularity, solidity and roundness) for the lateral view and another 8 for the dorsal view of the seeds. Therefore, 16 input variables were used to identify the seed's specie (variable to predict).

The database was randomly divided into three groups: 50% (for training, T), 30% (for validation, V) and 20% (for testing, Z). Data used were normalized to a range [-1,1].

RF models were created using the following hyperparameter combinations: number of trees (ranging from 1 to 100 using 99 steps), maximal depth (ranging from 1 to 100 using 99 steps), criterion (gini index, information gain, gain ratio and accuracy), pruning (true and false), pre-pruning (true and false) and voting strategy (majority and confidence). In the case of SVM models, the hyperparameters were SVM types (C-SVC and nu-SVC), *gamma* (between around $3 \cdot 10^{-5}$ and 8 in 18 steps), *C* (between around $3 \cdot 10^{-2}$ and 32768 in 20 steps) and nu (between 0.1 and 0.2 in 4 steps). Regarding ANN models, two hyperparameters were used: training cycles (ranging from 1 to 131,072 in 17 steps) and decay (true or false). The hidden neurons were stablished in the interval 1 to 2*n*+1 (being *n* the input variables).

The statistical parameters accuracy and kappa were employed to evaluate the performance of the models developed with the different algorithms.

All machine learning models were created using the RapidMiner Studio Educational 10.2.000 version (RapidMiner GmbH).

RESULTS & DISCUSSION

Different models, above 300,000, have been developed using the three algorithms with different hyperparameter combinations. The criterion used to select the best model for each algorithm is the highest accuracy value for the validation phase. Accordingly, the table shows the best models for each algorithm.

	T		V		Z	
Model	**Accuracy**	**Kappa**	**Accuracy**	**Kappa**	**Accuracy**	**Kappa**
RF	1.000	1.000	0.847	0.842	0.765	0.758
SVM	0.966	0.965	0.834	0.829	0.805	0.799
ANN	0.954	0.952	0.790	0.784	0.755	0.747

According to the statistical data shown in the table above, it can be said that the RF model presented the highest accuracy and kappa values in the validation and training phase, presenting an accuracy of 76.5% for the validation phase. The SVM model follows closely with high values in these two phases but showing the best accuracy for the testing phase (80.5%). The ANN model presents the worst adjustments in terms of accuracy in all phases.

It can be concluded that the three normalized models were suitable to determinate the species of *Silene* L., nevertheless, it is the SVM model the approach that presents a better accuracy for previously unseen samples. Therefore, these kind of machine learning models could be a necessary tool in taxonomy of *Silene* L. seeds.

CONCLUSION

Three different machine learning-based algorithms were developed to evaluate their performance to identify *Silene* L. seeds.

Regarding the obtained results, it can be said that the models developed presented good performance with the internal data, with accuracy values ranging between 0.790 and 1.000. In addition, these models also showed a good generalization capacity with the external data, showed accuracy values between 0.755 and 0.805.

Future work could attempt to improve these results by using other combinations of hyperparameters or using different models.

REFERENCES

[1] G. Zengin et al., "Functional constituents of six wild edible *Silene* species: A focus on their phytochemical profiles and bioactive properties," Food Biosci., vol. 23, pp. 75–82, 2018, doi: https://doi.org/10.1016/j.fbio.2018.03.010.

[2] H. H. Rashidi, N. K. Tran, E. V. Betts, L. P. Howell, and R. Green, "Artificial Intelligence and Machine Learning in Pathology: The Present Landscape of Supervised Methods," Acad. Pathol., vol. 6, 2374289519873088, 2019, doi: https://doi.org/10.1177/2374289519873088.

[3] J. J. Martín-Gómez, J. L. Rodríguez-Lorenzo, A. Juan, Á. Tocino, B. Janousek, and E. Cervantes, "Seed Morphological Properties Related to Taxonomy in *Silene* L. Species," Taxonomy, vol. 2, no. 3, pp. 298–323, 2022, doi: 10.3390/taxonomy2030024.