Scientia Agricola
http://dx.doi.org/10.1590/1678-992X-2015-0451

Automation in accession classification of Brazilian *Capsicum* germplasm through artificial neural networks

Mariane Gonçalves Ferreira[1], Alcinei Mistico Azevedo[2]*, Luhan Isaac Siman[1], Gustavo Henrique da Silva[1], Clebson dos Santos Carneiro[1], Flávia Maria Alves[1], Fábio Teixeira Delazari[1], Derly José Henriques da Silva[1], Carlos Nick[1]

[1]Federal University of Viçosa – Dept. of Crop Science, Av. Peter Henry Rolfs, s/n – 36570-900 – Viçosa, MG – Brazil.
[2]Federal University of Minas Gerais/Institute of Agricultural Sciences, Av. Universitária, 1000 – 39404-547 – Montes Claros, MG – Brazil.
*Corresponding author <alcineimistico@hotmail.com>

Edited by: Leonardo Oliveira Medici

Received November 25, 2015
Accepted July 11, 2016

ABSTRACT: Germplasm classification by species requires specific knowledge on/of the culture of interest. Therefore, efforts aimed at automation of this process are necessary for the efficient management of collections. Automation of germplasm classification through artificial neural networks may be a viable and less laborious strategy. The aims of this study were to verify the classification potential of *Capsicum* accessions regarding/ the species based on morphological descriptors and artificial neural networks, and to establish the most important descriptors and the best network architecture for this purpose. Five hundred and sixty-four plants from 47 Brazilian *Capsicum* accessions were evaluated. Neural networks of multilayer perceptron type were used in order to automate the species identification through 17 morphological descriptors. Six network architectures were evaluated, and the number of neurons in the hidden layer ranged from 1 to 6. The relative importance of morphological descriptors in the classification process was established by Garson's method. Corolla color, corolla spot color, calyx annular constriction, fruit shape at pedicel attachment, and fruit color at mature stage were the most important descriptors. The network architecture with 6 neurons in the hidden layer is the most appropriate in this study. The possibility of classifying *Capsicum* plants regarding/ the species through artificial neural networks with 100 % accuracy was verified.

Keywords: *Capsicum* spp., Garson's method, artificial intelligence, taxonomy, germplasm bank

Introduction

Capsicum is a genus of the highly diverse Solanaceae family with origins in South and Central America (Nicolai et al., 2013). Cultivated forms of *Capsicum* represent one of the most economically important vegetable crops worldwide (Albrecht et al., 2012), used as fresh vegetables and spices (Ibiza et al., 2012). Five domesticated *Capsicum* species are recognized and include *C. annuum*, *C. baccatum* L., *C. chinense* Jacq., *C. frutescens* L. and *C. pubescens* Ruiz et Pav. Twenty-five additional wild *Capsicum* species are documented (Djian-Caporalino et al., 2007). These species have been introduced from South America and *C. baccatum* var. *pendulum* (Willd.) Eshbaugh is endemic to the south-south-eastern region of Brazil (Albrecht et al., 2012).

High levels of global biodiversity and a limited number of taxonomists represent significant challenges to the future of biological study and conservation. The main problem is almost all taxonomic information exists in languages and formats not easily understood or shared without a high level of specialized knowledge and vocabularies. Thus, taxonomic knowledge is localized within limited geographical areas and among a limited number of taxonomists. Furthermore, an expert on one species or family may be unfamiliar with another (Cope et al., 2012). This lack of accessibility of taxonomic knowledge to the general public has been termed the "taxonomic crisis" (Dayrat, 2005). This has led to an increasing interest in automating the process of species identification and related tasks (Cope et al., 2012).

Artificial neural networks (ANN), as a pattern recognition tool, have been used for modeling complex systems (Azevedo et al., 2015). Its main advantages are the fact that it is non-parametric, it enables nonlinear solutions, and considers several explanatory variables simultaneously (Niska et al., 2010). This technique has been successfully used in the identification of *Banksia integrifólia* genotypes (Pandolfi et al., 2009), *Camellia* species (Lu et al., 2012), weed species (Li et al., 2009) and wheat plants among weed herbs (Gomez-Casero et al., 2010).

The aim of this study was to classify Brazilian pepper germplasm accessions regarding/ the species through artificial neural networks, through the use of morphological descriptors; to establish those of greater importance; and to determine the best network architecture.

Materials and Methods

Plant materials

The experiment was carried out at Viçosa, Minas Gerais, Brazil (20°45′14″ S; 42°52′53″ W; at 648.74 m asl), from Apr 2014 to Feb 2015. The climate is Cwb type, mesothermal humid, with rainy summers and dry winters, according to the Koppen climate classification (Alvares et al., 2013).

Forty-seven *Capsicum* accessions of the Vegetable Germplasm Bank of the Federal University of Viçosa (Table 1) were evaluated. Out of these accessions, 17 belong to the *C. annuum* var. *annuum* species; 20 belong to the *C. baccatum* var. *pendulum* species; one belongs to the *C. baccatum* var. *baccatum* species; seven belong to the *C. chinense* species; and two belong to the *C. frutescens* species. Seedlings were produced in expanded polystyrene trays of 128 cells. Transplanting to the field was carried out when seedlings presented 5 or 6 true leaves. Plants were spaced 1.0 m between rows, and 0.6 m between plants. Twelve plants were evaluated for each one of the 47 accessions, making 564 plants in total.

Table 1 – Identification and origin of *Capsicum* accessions.

Accessions	Scientific name	Origin
BGH-135	*C. annuum* var. *annuum*	Aracaju - SE
BGH-145	*C. annuum* var. *annuum*	Aracaju - SE
BGH-147	*C. chinense*	Aracaju - SE
BGH-169	*C. baccatum* var. *pendulum*	Maceió - AL
BGH-177	*C. annuum* var. *annuum*	Vitória de Santo Antão - PE
BGH-303	*C. annuum* var. *annuum*	Vitória de Santo Antão - PE
BGH-824	*C. baccatum* var. *pendulum*	Juiz de Fora - MG
BGH-853	*C. annuum* var. *annuum*	São José do Rio Pardo - SP
BGH-957	*C. chinense*	Campinas - SP
BGH-958	*C. annuum* var. *annuum*	Campinas - SP
BGH-1009	*C. annuum* var. *annuum*	Timbó - SC
BGH-1038	*C. annuum* var. *annuum*	Petrópolis - RJ
BGH-1039	*C. annuum* var. *annuum*	Guanabara - RJ
BGH-1258	*C. baccatum* var. *pendulum*	Marretes - PR
BGH-1275	*C. baccatum* var. *pendulum*	Raul Soares - MG
BGH-1276	*C. baccatum* var. *pendulum*	Raul Soares - MG
BGH-1611	*C. baccatum* var. *pendulum*	Pelotas, colônia - RS
BGH-1650	*C. baccatum* var. *pendulum*	Viçosa - MG
BGH-1652	*C. annuum* var. *annuum*	Currais Novos - RN
BGH-1661	*C. baccatum* var. baccatum	General Sampaio - CE
BGH-1680	*C. baccatum* var. *pendulum*	São Domingos do Prata - MG
BGH-1751	*C. chinense*	Ipeacs, km 47 - RJ
BGH-1770	*C. baccatum* var. *pendulum*	Cruz Alta - RS
BGH-1787	*C. chinense*	Maceió - AL
BGH-4169	*C. baccatum* var. *pendulum*	Curitiba - PR
BGH-4211	*C. chinense*	Marabá - PA
BGH-4562	*C. chinense*	Igarapé - MG
BGH-4563	*C. annuum* var. *annuum*	Igarapé - MG
BGH-4703	*C. annuum* var. *annuum*	Igarapé - MG
BGH-5385	*C. annuum* var. *annuum*	Dourados - MT
BGH-6011	*C. chinense*	Belém - PA
BGH-6016	*C. annuum* var. *annuum*	Belém - PA
BGH-6026	*C. baccatum* var. *pendulum*	São Paulo
BGH-6027	*C. baccatum* var. *pendulum*	São Paulo
BGH-6147	*C. annuum* var. *annuum*	Viçosa - MG
BGH-6267	*C. frutescens*	Univercity Purdue, USA
BGH-6272	*C. baccatum* var. *pendulum*	Embrapa, Cenargem, Brasília - DF
BGH-6649	*C. baccatum* var. *pendulum*	Correntina - BA
BGH-7174	*C. annuum* var. *annuum*	UENF
BGH-7178	*C. baccatum* var. *pendulum*	UENF
BGH-7179	*C. baccatum* var. *pendulum*	UENF
BGH-7184	*C. baccatum* var. *pendulum*	UENF
BGH-7190	*C. baccatum* var. *pendulum*	UENF
BGH-7278	*C. frutescens*	Natal - RN
BGH-7280	*C. baccatum* var. *pendulum*	Natal - RN
BGH-7281	*C. baccatum* var. *pendulum*	Brasília - DF
BGH-7282	*C. annuum* var. *annuum*	Brasília - DF

Morphological Traits

Seventeen descriptors proposed by IPGRI (1995) for the *Capsicum* were used for germplasm characterization:

1) Stem color: 1 = Green; 2 = Green with purple stripes; 3 = Purple; 4 = Other.

2) Leaf shape: 1 = Deltoid; 2 = Ovate; 3 = Lanceolate.

3) Days to flowering: Number of days from sowing/transplanting until 50 % of plants have at least one flower open.

4) Number of flowers per axil: 1 = One; 2 = Two; 3 = Three or more; 4 = Many flowers in bunches but each in individual axil (fasciculate growth); 5 Other (cultivars with two flowers in the first axil and with only one in the other).

5) Flower position: 1 = Pendant; 2 = Intermediate; 3 = Erect.

6) Corolla colour: 1 = White; 2 = Light yellow; 3 = Yellow; 4 = Yellow-green; 5 = Purple with White base; 6 = White with purple base; 7 = White with purple margin; 8 = purple; 9 = Other.

7) Corrolla spot colour: 1 = White; 2 = Yellow; 3 = Yellow-green; 4 = Green; 5 = Purple; 6 = Other.

8) Anther colour: 1 = White; 2 = Yellow; 3 = Pale blue; 4 = Blue; 5 = Purple; 6 = Other.

9) Calyx annular constriction: At junction of calyx and pedicel. Observed at mature stage. 1 = Absent; 2 = Present.

10) Fruit colour at intermediate stage: 1 = White; 2 = Yellow; 3 = Green; 4 = Orange; 5 = Purple; 6 = Deep purple; 7 = Other.

11) Fruit colour at mature stage: 1 = White; 2 = Lemon-yellow; 3 = Pale Orange-yellow; 4 = Orange-yellow; 5 = Pale orange; 6- Orange; 7 = Light red; 8 = Red; 9 = Dark red; 10 = Purple; 11 = Brown; 12 = Black; 7 = Other.

12) Fruit shape: 1 = Elongate; 2 = Almost round; 3 = Triangular; 4 = Campanulate; 5 = Blocky; 6 = Other.
13) Fruit length: Average fruit length of 10 ripe fruits.

14) Fruit width: Measured at the widest point. Average fruit with of 10 ripe fruits.

15) Fruit weight: Average fruit with of 10 ripe fruits.

16) Fruit shape at pedicel attachment: 1 = Acute; 2 = Obtuse; 3 = Truncate; 4 = Cordate; 5 = Lobate.

17) Fruit shape at blossom end Fruit weight (FWE): 1 = Pointed; 2 = Blunt; 3 = Sunken; 4 = Sunken and pointed; 5 = Other.

Training of neural networks and selection of the best topologies

For the development of MLP networks (Multi-Layer-Perceptron), the Neural Network Toolbox from Matlab software (version 8.1.0.604) was used with back-propagation algorithm and Levenberg-Marquadt optimization.

In the process of MLP training, 17 morphological descriptors were used as input data. In the output layer, the scientific name of each plant was identified represented by numbers from 1 to 5 (1 = *C. annuum* var. *annuum*; 2 = *C. baccatum* var. *pendulum*; 3 = *C. baccatum* var. *baccatum*; 4 = *C. chinense*; 5 = *C. frutescens*).

Information from twelve plants was used for each accession, making a total of 564 plants. From this set, plant information was randomly used as follows: 282 plants for training (50 %), 141 plants for cross-validation - early stopping (25 %), and 141 plants for testing (25 %).

For better network efficiency, before training, input data were normalized for an interval between -1 and 1. The maximum number of training epochs was set as 1000; the minimum mean squared error (MSE) for stopping was set as 1.0×10^{-7}, and the maximum number of successive failures (early stopping) was set as 6.

All trained networks had a neuron in the output layer, and a single hidden layer. In order to identify the best network architecture, 1-6 neurons were tested in the hidden layer, making a total of 6 architectures. Logistic activation function was used in the hidden layer, and a linear activation function was used in the output layer. At the beginning of the training, the synaptic weights are randomly generated, which influences the final result. Thus, 1000 trainings were carried out for each network architecture. Network efficiency was presented in bar graphs for accuracy rate (percentage of correct classifications), accompanied by the respective standard deviations. Hyperbolic tangent function activation was used for neurons in the hidden layers. For the output layer, the linear function was used.

Determining variable importance

In order to reduce the required number of traits for plant classification, the most important traits were determined by Garson's method (1991). The relative contribution (%) was presented by bar graph, with the respective standard deviations. Subsequently, new trainings (1000 trainings) were carried out for each of the six network architectures, considering only the most important traits as input.

Results and Discussion

High efficiency of artificial neural networks was found in the classification of Brazilians *Capsicum* accessions regarding the species. When six neurons were used in the hidden layer, the accuracy rate was 100 % in 1000 trainings (Figure 1). This accuracy rate decreased with the reduction of the number of neurons in the hidden layer. When only one neuron was used in the hidden layer, in the 1000 trainings, there was a mean accuracy rate of 78 %. Li et al. (2009), studying weeds classification, also observed a decrease in the efficiency of neural networks with a reduced number of neurons in the hidden layer.

Figure 2 shows the network architecture with six neurons in the hidden layer, which was more efficient (Figure 1). In a similar study, Li et al. (2009) used 60 neurons in the hidden layer, and found a 78 % accuracy rate. The greater efficiency found in the present study (100 %) may be explained by the greater number of plants evaluated (564 plants), since for more efficient network training, it is important the availability of large data sets (Azevedo et al., 2015). Furthermore, the explanatory variables used in this study may have been more appropriate than those used by Li et al. (2009).

Although the 17 descriptors used in this study are easy to measure, their evaluation may be unfeasible if there is a very large number of plants to be classified. In this context, the study of the most important traits in prediction through ANN becomes necessary, which makes it possible to reduce computational effort and the use of labor (Paliwal and Kumar, 2011).

According to Garson's method (1991), the descriptors with higher relative contribution were: Corolla color (16 %), calyx annular constriction (9 %), corolla spot color (8 %), fruit color at mature stage (8 %), and fruit shape at pedicel attachment (7 %) (Figure 3). Garson's method is widely used in the literature and may be used for the application of low importance explanatory variables (Olden et al., 2004).

After applying the descriptors with a lower relative contribution, one hundred new trainings were carried out, and a high level of efficiency in plant classification was confirmed (Figure 4). When using six neurons in

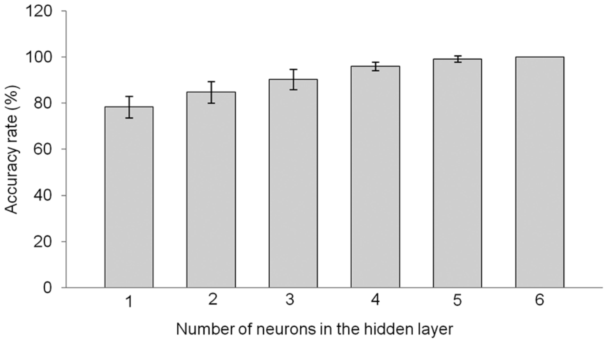

Figure 1 – Accuracy rate and its standard deviation regarding the classification of *Capsicum* species through artificial neural networks of multilayer perceptron type, using 17 morphological descriptors.

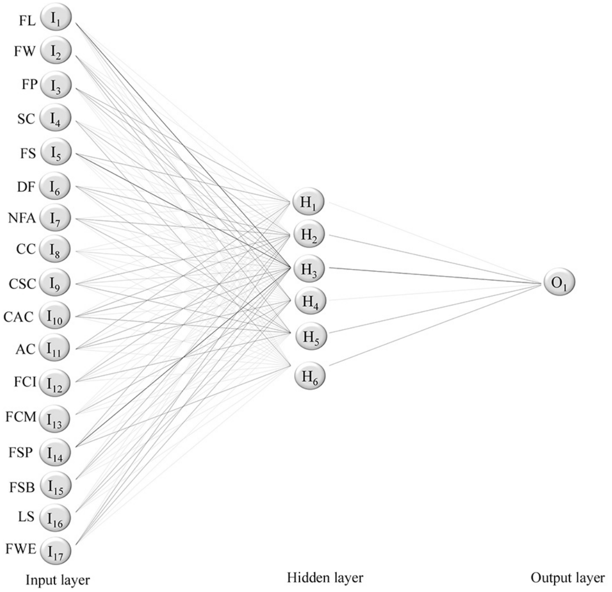

Figure 2 – Architecture of the best network evaluated for classification of *Capsicum* plants by scientific name, considering 17 morphological descriptors in the input layer. Fruit length (FL); Fruit width (FW); Flower position (FP); Stem color (SC); Fruit shape (FS); Days to flowering (DF); Number of flowers per axil (NFA); Corolla color (CC); Corrolla spot color (CSC); Calyx annular constriction (CAC); Anther color (AC); Fruit color at intermediate stage (FCI); Fruit color at mature stage (FCM); Fruit shape at pedicel attachment (FSP); Fruit shape at blossom end (FSB); Leaf shape (LS) and Fruit weight (FWE).

the hidden layer, a 100 % accuracy rate was found. Very close results were also found for network architectures with four and five neurons in the hidden layer. In this way, the feasibility of reducing the number of descriptors to five was verified. The morphological descriptor with the highest relative contribution when considering six neurons in the hidden layer was corolla color (56 %), followed by corolla spot color (18 %), calyx annular constriction (11 %), fruit shape at pedicel attachment (8 %), and fruit color at the mature stage (7 %) (Figure 5).

The three most important traits (color corolla, corolla spot color, and calyx annular constriction) are associated with flowers. Traits associated with flowers are also considered as important in the classification of pepper varieties by Sudré et al., 2010. According to these authors, the different species and varieties of peppers can be differentiated by morphological traits, especially in flowers, such as the position of the flower and the pedicel, the presence or absence of spots in the petal lobes, the edge of the calyx, and the number of flowers per internode.

The possibility of automation through ANNs for species classification is important, since it eliminates the necessity of extensive knowledge on taxonomy (Cope et al., 2012). Furthermore, it enables the evaluation of a large number of plants in an easy, efficient and less laborious way (Husin et al., 2012).

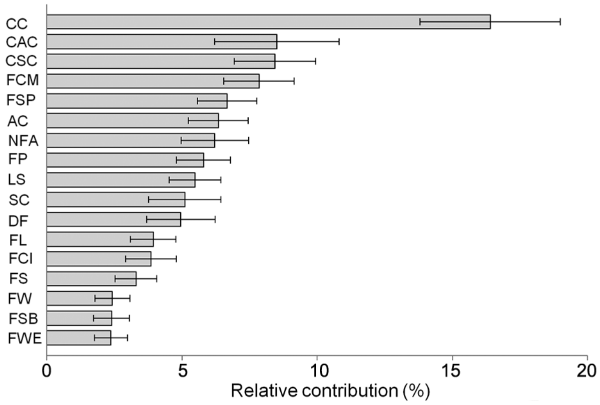

Figure 3 – Relative contribution estimated by Garson's method (1991) of 17 morphological descriptors used in the classification of *Capsicum* plants through artificial neural networks. Corolla color (CC); Calyx annular constriction (CAC); Corrolla spot color (CSC); Fruit color at mature stage (FCM); Fruit shape at pedicel attachment (FSP); Anther color (AC); Number of flowers per axil (NFA); Flower position (FP); Leaf shape (LS); Stem color (SC); Days to flowering (DF); Fruit length (FL); Fruit color at intermediate stage (FCI); Fruit shape (FS); Fruit width (FW); Fruit shape at blossom end (FSB); Fruit weight (FWE).

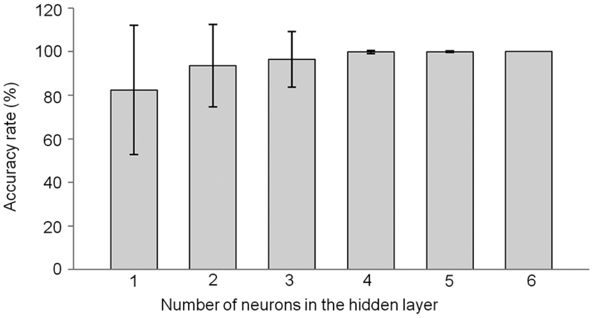

Figure 4 – Accuracy rate and its standard deviation in terms of the classification of pepper species through artificial neural networks of multilayer perceptron type, using 5 morphological descriptors.

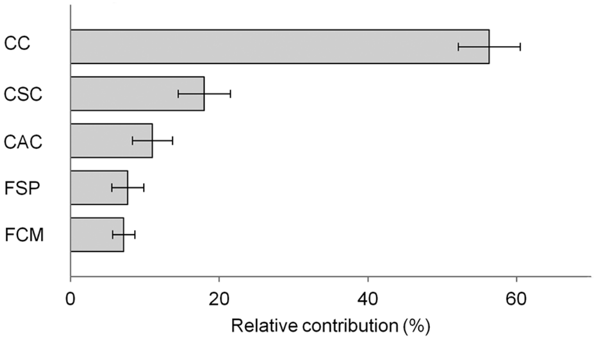

Figure 5 – Relative contribution estimated by Garson's method (1991) of 5 morphological descriptors used in the classification of *Capsicum* plants through artificial neural networks. Corolla color (CC); Corrolla spot color (CSC); Calyx annular constriction (CAC); Fruit shape at pedicel attachment (FSP); Fruit color at mature stage (FCM).

Conclusion

The morphological descriptors corolla color, corolla sport color, calyx annular constriction, fruit shape at pedicel attachment, and fruit color at mature stage are the most important in this study. Network architecture with six neurons in the hidden layer is the most appropriate. It is possible to classify plants of the *Capsicum* genus by species through artificial neural networks, considering morphological descriptors with 100 % accuracy.

Acknowledgments

The authors thank the Brazilian National Council for Scientific and Technological Development (CNPq), the Coordination for the Improvement of Higher Level Personnel (CAPES) and the Minas Gerais State Foundation for Research Support (FAPEMIG) for providing scholarships and resources to develop this study.

References

Albrecht, E.; Zhang, D.; Saftner, R.A.; Stommel, J.R. 2012. Genetic diversity and population structure of *Capsicum baccatum* genetic resources. Genetic Resources and Crop Evolution 59: 517-538.

Alvares, C.A.; Stape, J.L.; Sentelhas, P.C.; Gonçalves, J.L.M.; Sparovek, G. 2013 Köppen's climate classification map for Brazil. Meteorologische Zeitschrift 22: 711-728.

Azevedo, A.M.; Andrade Júnior, V.C.; Pedrosa, C.E.; Oliveira, C.M.; Dornas, M.F.S.; Cruz, C.D.; Valadares, N.R. 2015. Application of artificial neural networks in indirect selection: a case study on the breeding of lettuce. Bragantia 4: 387-393.

Cope, J.S.; Corney, D.; Clark, J.Y.; Remagnino, P.; Wilkin, P. 2012. Plant species identification using digital morphometrics: a review. Expert Systems with Applications 39: 7562-7573.

Dayrat, B. 2005. Towards integrative taxonomy. Biological Journal of the Linnean Society 85: 407-415.

Djian-Caporalino, C.; Lefebvre, V.; Sage-Daube`ze, A.M.; Palloix, A. 2007. *Capsicum*. p. 186-245. In: Singh, R.J.; Jauhar, P.P., eds. Genetic resources, chromosome engineering and crop improvement. CRC Press, Boca Raton, FL, USA.

Garson, G.D. 1991. Interpreting neural network connection weights. International Journal of Artificial Intelligence and Expert Systems 6: 47-51.

Gomez-Casero, M.T.; Castillejo-Gonzalez, I.L.; Garcia-Ferrer, A.; Pena-Barragan, J.M.; Jurado-Exposito, M.; Garcia-Torres, L.; Lopez-Granados, F.L. 2010. Spectral discrimination of wild oat and canary grass in wheat fields for less herbicide application. Agronomy for Sustainable Development 30: 689-699.

Husin, Z.; Shakaff, A.Y.M.; Aziz, A.H.A.; Farook, R.S.M.; Jaafar, M.N.; Hashim, U.; Harun, A. 2012. Embedded portable device for herb leaves recognition using image processing techniques and neural network algorithm. Computers and Electronics in Agriculture 89: 18-29.

Ibiza, V.P.; Blanca, J.; Canizares, J.; Nuez, F. 2012. Taxonomy and genetic diversity of domesticated *Capsicum* species in the Andean region. Genetic Resources and Crop Evolution 59: 1077-1088.

International Plant Genetic Resources Institute [IPGRI]. 1995. Descriptors for *Capsicum* (*Capsium* spp.). FAO/IPGRI, Rome, Italy.

Li, Z.; An, Q.; Ji, C. 2009. Classification of weed species using artificial neural networks based on color leaf texture feature. Computer and Computing Technologies in Agriculture II 2: 1217-1225.

Lu, H.; Jiang, W.; Ghiassi, M.; Lee, S.; Nitin, M. 2012. Classification of *Camellia* (Theaceae) species using leaf architecture variations and pattern recognition techniques. PLoS One 7: e29704.

Nicolai, M.; Cantet, M.; Lefevre, V.; Sage-Palloix, A.M.; Polloix, A. 2013. Genotyping a large collection of pepper (*Capsicum* spp.) with SSR loci brings new evidence for the wild origin of cultivated *C. annuum* and the structuring of genetic diversity by human selection of cultivar types. Genetic Resources and Crop Evolution 60: 2375-2390.

Niska, H.; Skon, J.P.; Packalén, P.; Tokola, T.; Maltamo, M.; Kolehmainen, M. 2010. Neural networks for the prediction of species-specific plot volumes using airborne laser scanning and aerial photographs. IEEE Transactions on Geoscience and Remote Sensing 48: 1076-1085.

Olden, J.D.; Joy, M.K.; Death, R.G. 2004. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. Ecological Modelling 178: 389-397.

Paliwal, M.; Kumar, U. 2011. Assessing the contribution of variables in feed forward neural network. Applied Soft Computing 11: 3690-3696.

Pandolfi, C.; Messina, G.; Mugnai, S.; Azzarello, E.; Masi, E.; Dixon, K.; Mancuso, S. 2009. Discrimination and identification of morphotypes of *Banksia integrifolia* (Proteaceae) by an Artificial Neural Network (ANN), based on morphological and fractal parameters of leaves and flowers. Taxon 58: 925-933.

Sudré, C.P.; Gonçalves, L.S.A.; Rodrigues, R.; Amaral Júnior, A.T.; Riva-Souza, E.M.; Bento, C.S. 2010. Genetic variability in domesticated *Capsicum* spp. as assessed by morphological and agronomic data in mixed statistical analysis. Genetics and Molecular Research 9: 283-294.