Euphytica (2009) 166:411–421
DOI 10.1007/s10681-008-9828-9

Artificial neural networks as a tool for plant identification: a case study on Vietnamese tea accessions

Camilla Pandolfi · Sergio Mugnai · Elisa Azzarello · Silvia Bergamasco · Elisa Masi · Stefano Mancuso

Received: 11 September 2007 / Accepted: 30 September 2008 / Published online: 12 October 2008

Abstract Seventeen tea accessions belonging to Chinese (*Camellia sinensis*), Assamic (*C. sinensis* var. *assamica*), and Shan tea (*C. sinensis* var. *pubilimba*) groups, which are either commercially planted or new promising tea germplasm, were morphologically described at Phu Tho province (Viet Nam) and assessed for their diversity. Fourteen phyllometric parameters were qualitatively and quantitatively investigated using digital image analysis. The accessions were then discriminated by a dedicated artificial neural network for univocal plant identification and a hierarchical cluster analysis was performed in order to build a dendrogram reporting the relationships among them. Results proved the diversity of investigated tea morphotypes from Phu Tho province based on a morphological screening. More, the artificial neural network was able to perform a correct identification for almost all the accessions using simple dedicated instruments.

Keywords *Camellia sinensis* var. *assamica* · *C. sinensis* var. *pubilimba* · *C. sinensis* · Image analysis · Morphotypes · Phyllometric parameters

C. Pandolfi · S. Mugnai (✉) · E. Azzarello · S. Bergamasco · E. Masi · S. Mancuso
Department of Horticulture, University of Florence, viale delle Idee 30, 50019 Sesto Fiorentino, FI, Italy
e-mail: sergio.mugnai@unifi.it

Abbreviations

ANN	Artificial neural network
BPNN	Back-propagation neural network

Introduction

The tea plant is classically classified as *Camellia sinensis* (L.) O. Kuntze belonging to the family *Theaceae*. It is indigenous to the area which includes the Chinese provinces of Yunnan and Sichuan (Chen and Yamaguchi 2005), and spontaneously grows widely from tropical to temperate Asian regions. Tea has been a source of revenue for almost all of the producing countries and has contributed significantly to the local rural economies (Paul et al. 1997). Viet Nam is considered a leading tea producer and exporter, with Phu Tho as one of the most important provinces for tea production, located between the North Vietnam plains and midlands.

The classification of tea plant had been controversial for many years (Chen et al. 2006a, b). In 1919, Cohen-Stuart proposed only one species, *C. sinensis*, with three varieties: *C. sinensis* var. *bohea* (Chinese tea), *C. sinensis* var. *shan* (Shan tea) and *C. sinensis* var. *assamica* (Assam tea) (Yamamoto et al. 1997). Sealy (1958) proposed a new classification based on leaf and growth characteristics in which two distinct *taxa* were described: *C. sinensis* var. *sinensis* from China, with small leaves, dwarf habitus and slow

growth, and *C. sinensis* var. *assamica* (Masters) Kitamura from the Assam region in India, with large leaves, tall habitus and quick growth. This classification was revised by Wight (1962) who relied mainly on reproductive structures and assigned specific status to var. *sinensis* and var. *assamica* and recognized a Southern or Cambodian form of *C. assamica* (Masters) Wight classified as *C. assamica* ssp. *lasiocalyx* (Planchon ex. Watt) Wight. Nowadays, tea classification is mainly based on Chang's (1981, 1984) taxonomic system, which usually involves one main species (*C. sinensis*), three varieties (*C. sinensis* var. *assamica*, *C. sinensis* var. *pubilimba*, named 'Shan tea' in Viet Nam, and *C. sinensis* var. *kucha*) and numerous botanical *formae* (i.e., *C. sinensis* fo. *macrophylla*).

Tea is an allogamous plant with a massive freely interbreeding. For this reason, tea plants are highly diverse and heterozygous, with many overlapping morphological, biochemical and physiological attributes (Willson and Clifford 1992). Indeed, because of the extreme hybridization, existence of the pure archetypes of tea is doubtful (Willson and Clifford 1992). The correct classification of tea genotypes is further complicated by the great number of ecotypes not yet registered as cultivars, but locally well-known and cultivated.

With this background, accurate but rapid tea cultivar identification is important and mandatory for both practical breeding purposes and proprietary rights protections. Understanding the genetic background will also greatly help in selecting parents for current and long-term success of tea breeding programs.

Recently, numerous studies focused on the identification and distinction among the different tea genotypes. Interesting perspectives came from the use of isoenzymatic markers (Lu et al. 1992) or from the molecular characterization by randomly amplified polymorphic DNA (RAPD) (Kaundun et al. 2000; Chen and Yamaguchi 2005). More, the simple sequence repeat anchored polymerase chain reaction (SSR-anchored PCR) has been used in *C. sinensis* to determinate parentage and genetic diversity by the development of microsatellite markers (Ueno et al. 1999; Kaundun and Matsumoto 2002). The genetic diversity was also assessed analyzing secondary metabolites such as leaf catechins and polyphenols (Saravanan et al. 2005). Although these methods are effective, they are also resource and labor intensive, and require a skilled and experienced technical staff to be effectively exploited. Therefore, we assessed the use of artificial neural networks (ANN) as a possible alternative for the discrimination and identification of tea genotypes from morphological parameters in those situations in which the use of molecular methods is not possible for technical and/or economic reasons.

Artificial Neural Networks (ANN) are powerful computational tools that "learn" with training examples and have the capability for extrapolating their "knowledge" to new situations related to problems of classification, modeling, mapping and association types. ANN are an attempt to emulate (very roughly) the basic functions of the mammalian brain to perform complex functions that computer systems are incapable of doing. Though one of the acknowledged advantages of the neural networks is the capacity to overcome the need for a sample statistically representative of a population, they also have the capability for generalization beyond the training data, to produce approximately correct results for new cases that were not used in training (Pandolfi et al. 2006). The most utilized type of network for plant identification is the supervised back-propagation neural network (Mancuso and Nicese 1999; Mugnai et al. 2008), which is a particular kind of multilayer feed-forward network, or multilayer perceptron (MLP). Briefly, a BPNN in its basic form has a layered structure, with its architectural layout basically composed by some layers of neurons: the input layer, one or more hidden layers and the output layer. Each layer receives its input from the previous layer or from the network input, while the output of each neuron feeds the next layer or the output of the network. Particular nodes were also used to shift the neuron transfer function and to improve the network performance, thanks to the back-propagation of errors (Rumelhart et al. 1986). Further details on the construction of a dedicated BPNN for plant identification are available in Mugnai et al. (2008). In horticulture the applications of ANNs are just at the beginning despite their skill and speed in the recognition of patterns in complex, non-linear data, such as those derived from many experimental area of horticulture (Pandolfi et al. 2006). However, neural networks have been recently and successfully applied to the identification of *C. japonica* L. varieties (Mugnai et al. 2008).

In this study, the main scope was to build, train and test a back-propagation neural network (BPNN) to morphologically differentiate and univocally discriminate 17 accessions of tea, selected among the most broadly diffused genotypes in Phu Tho province (Viet Nam), which have been collected in the Tea Research Institute of Vietnam (TRI). The assessment of morphological diversity of the existing tea resources should help: (1) to improve the choice of varieties for agronomically important characters; (2) to preserve the intellectual property rights of tea breeders; (3) to quickly identify individual tea varieties in a given environment by making a "fingerprint" passport; (4) to permit a preliminary classification of tea morphotypes based on leaf morphological parameters.

Materials and methods

Plant material

The plant material was collected from the living collection located in Thanh Ba District, Phu Tho Province, Viet Nam (Lat: 21°27′N; Long: 105°14′E). All the selected accessions belonged to the genus *Camellia* (Table 1). Leaves for the morphological characterization were picked from at least three plants per accession, chosen for their good edaphic conditions and solar radiation exposition. Any sample of leaves was composed of 40 specimen, a number considered optimum for this experimental method (Mancuso 1999).

Image acquisition and determination of morphometric parameters

An optical scanner, set at 300 × 300 dpi, 16 million colors, was used to acquire leaves images (Fig. 1). Fourteen morphometric parameters (Table 2) were determined for each image through an image analysis software (UTHSCSA Image Tool 3.0) performed on a personal computer.

Construction of the BPNN

In this study, 14 image analysis parameters were used as input layers, and the 17 tea accessions represented the output. To optimize the neural network activity, the number of hidden neurons and the number of iterations was modified. Concerning the hidden layer many factors such as learning scheme, numbers of nodes of the output and input and connections between them, play an important role for the

Table 1 The 17 selected tea accessions collected from the Tea Research Institute, Phu Tho, Viet Nam

Abbr.	Name	Species	Type	Origin
BAT	Bat Tien	*C. sinensis*	Clone	Taiwan
CHAT	Chat Tien	*C. sinensis* var. *pubilimba*	Clone	Ha Giang, Vietnam
CU	Cu De Phung	*C. sinensis* var. *pubilimba*	Clone	Ha Giang, Vietnam
GIA	Gia Vai	*C. sinensis* var. *pubilimba*	Clone	Ha Giang, Vietnam
HUNG	Hung Ding Bach	*C. sinensis*	Clone	China
KEO	Keo Am Tich	*C. sinensis*	Clone	China
KIM	Kim Tuyen	*C. sinensis* × *C. sinensis* var. *assamica*	Clone	Phu Tho, Vietnam
LDP2	LDP2	*C. sinensis* × *C. sinensis* var. *assamica*	Clone	Phu Tho, Vietnam
LDP1	LDP1	*C. sinensis* × *C. sinensis* var. *assamica*	Clone	Phu Tho, Vietnam
NAM	Nam Ngat	*C. sinensis* var. *pubilimba*	Clone	Ha Giang, Vietnam
PH1	PH1	*C. sinensis* var. *assamica*	Seed	Assam, India
PHUC	Phuc Van Tien	*C. sinensis*	Clone	China
PT95	PT 95	*C. sinensis*	Clone	China
TAM	Tam Ve	*C. sinensis* var. *pubilimba*	Clone	Ha Giang, Vietnam
THUY	Thuy Ngoc	*C. sinensis* × *C. sinensis* var. *assamica*	Clone	Phu Tho, Vietnam
TRI	TRI 777	*C. sinensis* var. *assamica*	Clone	Sri Lanka
TRU	Trung Du	*C. sinensis* fo. *macrophylla*	Seed	Phu Tho, Vietnam

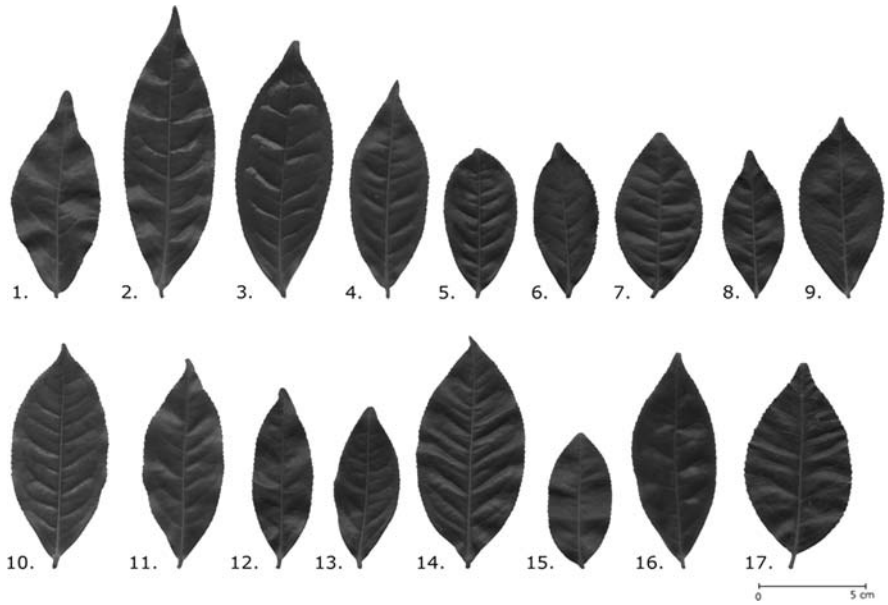


Fig. 1 Leaf images acquired by an optical scanner 300 × 300 dpi. Leaves were picked from at least three plants per accession, chosen for their good edaphic conditions and solar radiation exposition. (1) Bat Tien; (2) Chat Tien; (3) Cu de Phung; (4) Gia Vai; (5) Hung Ding Bach; (6) Keo Am Tich; (7) Kim Tuyen; (8) LDP 1; (9) LDP 2; (10) Nam Ngat 2; (11) PH 1; (12) Phuc Van Tien; (13) PT 95; (14) Tham Ve; (15) TRI 777; (16) Thuy Ngoc; (17) Trung Du

determination of the best configuration (Zurada and Malinowski 1994). In our case, the minimum error was reached with a network composed of 50 hidden neurons, positioned on one level, with the hidden layer activated by a logistic sigmoid activation function:

$$f(x) = \frac{1}{(1 + e^{-x})} \quad (1)$$

The function of a node's activation (except for the input layer, which uses the input themselves) controls the output signal strength for the unit. These sigmoid functions set the output signal strength between 0 and 1. The sigmoid function acts like an output gate that can be opened (1) or closed (0). As the function is continuous, it is also possible for the gate to be partially opened (i.e., a value between 0 and 1). In an ideal case only a class of output, representing an accession, would show an average value of 1 (correct identification) while all the other classes would show the value 0 (incorrect identification). In practice this happens occasionally, so a value closer to zero is considered as 'wrong', while 'right' is a value as close as possible to 1. The learning phase was protracted until the root mean square (RMS) error was less than 0.06 and the difference between the RMS in two consecutive periods was less than 0.0001. The BPNN outputs can be represented by a *XY*-graph for each accession, with the accession names on the *x*-axis, and the *y*-axis representing the output. Each graph aims to show how the BPNN was able to discriminate the selected accession in comparison with the others.

Assessment of performance

A misidentification matrix was produced showing the values of identification for each species. All identification attempts performed by the network were averaged to produce the results in the table. On the bottom row, the matrix also shows the confidence of correct identification (%Conf). This is identical to the confidence of correct classification used by Morgan et al. (1998), and is a measure of the likelihood that a species identification is correct, given that the network truly and effectively identified an unknown specimen as that taxon. It is calculated by expressing as a percentage the proportion of correct identifications with respect to the total number of identifications, including wrong identifications (Eq. 2).

$$\%\text{Conf} = \frac{\text{correct}}{\text{correct} + \text{incorrect}} \times 100 \quad (2)$$

This represent a method of analyzing outputs that weights the correct output in comparison with the incorrect ones. It is just a mathematic evaluation and it is not able to percept the difference between an

Table 2 Leaf morphological inputs determined by the image analysis software

	Parameter	Definition
1	Area	The area of the leaf
2	Perimeter	The perimeter of the leaf
3	Major axis length	The length of the longest line that can be drawn trough the leaf
4	Minor axis length	The length of the longest line that can be drawn trough the leaf perpendicular to the major axis
5	Roundness	Computed as: (4 × π × area)/perimeter 2
6	Elongation	The ratio of the length of the major axis to the length of the minor axis
7	Feret diameter	The diameter of a circle having the same area of the leaf
8	Compactness	Computed as: sqrt (4 × area/π)/major axis length
9	Integrated density	Computed as the product of the mean gray level and the number of pixels in the image of the leaf
10	Min gray level	Minimum gray level of the leaf
11	Mean gray level	Mean gray level of the leaf
12	Median gray level	Median gray level of the leaf
13	Mode gray level	Mode gray level of the leaf
14	Max gray level	Maximum gray level of the leaf

unique and significant peak in an incorrect class, and the presence of diffuse peaks in different incorrect accessions. For this reason, %Conf was used for a preliminary screening of the accessions before the direct analysis of each output value and its related graph.

Statistical analysis

Leaf morphometric parameters were subjected to one-way ANOVA and their means separated by Tukey's Multiple Comparison Test ($n = 40$, $P < 0.05$). NTSYS 2.1 was used to investigate neural network outputs performing a cluster analysis using the Unweighted Pair Group Method Analysis (UPGMA) based on the similarity matrix calculated using the cosine function (Eq. 3).

$$\mathrm{COSINE}_{(x,y)} = \frac{\sum_i (x_i y_i)}{\sqrt{\left(\sum_i x_i^2\right) \times \left(\sum_i y_i^2\right)}} \quad (3)$$

Results

The first four morphometric parameters listed in Table 2 permitted a preliminary discrimination among the accessions with the creation of groups of similarity on the basis of leaf size and morphology (Table 3). For example, while the highest leaf area was found in the accession CHAT, seven accessions showed the smallest (HUNG, KEO, KIM, LDP2, PHUC, PT95, THUY). Leaf perimeter almost followed the area behavior. On the contrary, the other two graphs (leaf major axis and leaf minor axis) focused on the different leaf shapes. CHAT had both the highest major and minor axis, while other accessions showed different leaf shapes, more lanceolate (PHUC) or more obovate (TRU) than CHAT. By the way, the similarities among the accessions were better expressed and deeper appreciated by the construction of a dedicated ANN. The species-based misidentification matrix is shown in Table 4, where the rows refer to the species in the test set. Similarly, the columns report the species to which the test plants are referred by the neural network. Identification average values are shown relative to the total samples of the row test species that are identified as belonging to the corresponding column species, whereas correct identifications are shown in bold. The network almost completely and univocally discriminated among the relative accessions with the exceptions of CU, GIA, HUNG, LDP1 and TRU (Table 4; Fig. 2). The average outputs for the correct identification ranged between 0.31 (NAM) and 0.76 (CHAT), while those for the incorrect identification ranged between 0.19 (HUNG) and 0.28 (CU). These results underline that the immediate analysis of output data can not lead to

Table 3 Principal leaf morphological parameters calculated through the image analysis software (area, perimeter, leaf major axis, leaf minor axis)

	Leaf area (cm^2)	Perimeter (cm)	Major axis length (cm)	Minor axis length (cm)
BAT	24.83 d	24.78 de	10.12 d	3.89 ef
CHAT	39.19 a	32.90 a	13.59 a	4.35 abc
CU	30.42 c	27.59 b	11.15 bc	4.14 cd
GIA	32.14 bc	28.27 b	11.62 b	4.29 abcd
HUNG	17.37 f	20.12 g	8.02 gh	3.12 h
KEO	16.30 f	19.82 g	7.83 gh	3.18 h
KIM	17.23 f	18.97 gh	7.34 hi	3.43 gh
LDP2	15.85 f	19.94 g	7.80 gh	3.16 h
LDP1	21.45 de	22.87 f	9.09 f	3.69 fg
NAM	31.38 bc	26.59 bc	10.54 cd	4.51 ab
PH1	23.97 d	23.20 ef	9.13 ef	3.97 de
PHUC	15.21 f	19.57 g	7.96 gh	2.79 i
PT95	18.09 ef	20.40 g	8.46 fg	3.21 h
TAM	34.05 b	28.33 b	11.34 bc	4.50 a
THUY	15.40 f	17.54 h	6.71 i	3.27 h
TRI	30.73 bc	26.67 bc	11.02 c	4.18 bcd
TRU	28.93 c	25.44 cd	9.95 de	4.45 ab

Data were subjected to one-way ANOVA and their means separated by Tukey's Multiple Comparison Test. Different letters refer to a significant difference for $P < 0.05$ ($n = 40$)

fix an unique threshold value for a correct and successful discrimination. On the contrary, the discrimination process must be carefully performed at each moment because the absence of other significant peaks in correspondence to other accessions is much more determinant for a correct identification than high values in the correct output. For example, accessions NAM, PT95 and PH1 reported a low average output between 0.31 and 0.32, but these results can be still accepted as successful discriminations because the output graph reported no other significant peaks in correspondence to other accessions (Fig. 2). On the contrary, accessions CU and GIA associated a low output value to a concurrent and almost equal significant peak corresponding to the other accession. In fact, the average output of CU was 0.28, with a significant peak corresponding to GIA (0.18), while the average output of GIA was 0.26 with a significant peak in CU (0.19). The last row of the table refers to the confidence of correct identification (%Conf). The value of this coefficient range between 84.92% (CHAT) and 21.91% (LDP2). As a general behavior, we can assess that a %Conf lower than 30% should refer to an incorrect identification but, as previously asserted, this value alone is not able to distinguish between an unique uncorrected peak or a sum of small uncorrected peaks, or rumors. For example, even if TRU showed a %Conf of 31.72, the network was not able to univocally distinguish this accession from the others. The network outputs were also analyzed using the UPGMA method for the construction of a dendrogram (Fig. 3). The dendrogram can be split into 2 principal clusters (A and B). Cluster A contains mainly Shan Tea genotypes, with a sub-cluster formed by all the Shan tea varieties where four of them are strictly grouped together (CU, GIA, TAM, NAM). CU and GIA confirmed the highest level of similarity (coefficient of similarity 0.60) previously noticed, followed by TAM and NAM (0.44 and 0.29, respectively). On the contrary, Cluster B includes the majority of the Chinese varieties and the hybrids between Chinese and Assamic genotypes. HUNG, KEO, LDP1, LDP2, PHUC and PT95 constituted the first sub-cluster, whereas KIM, PH1 and THUY belonged to the second.

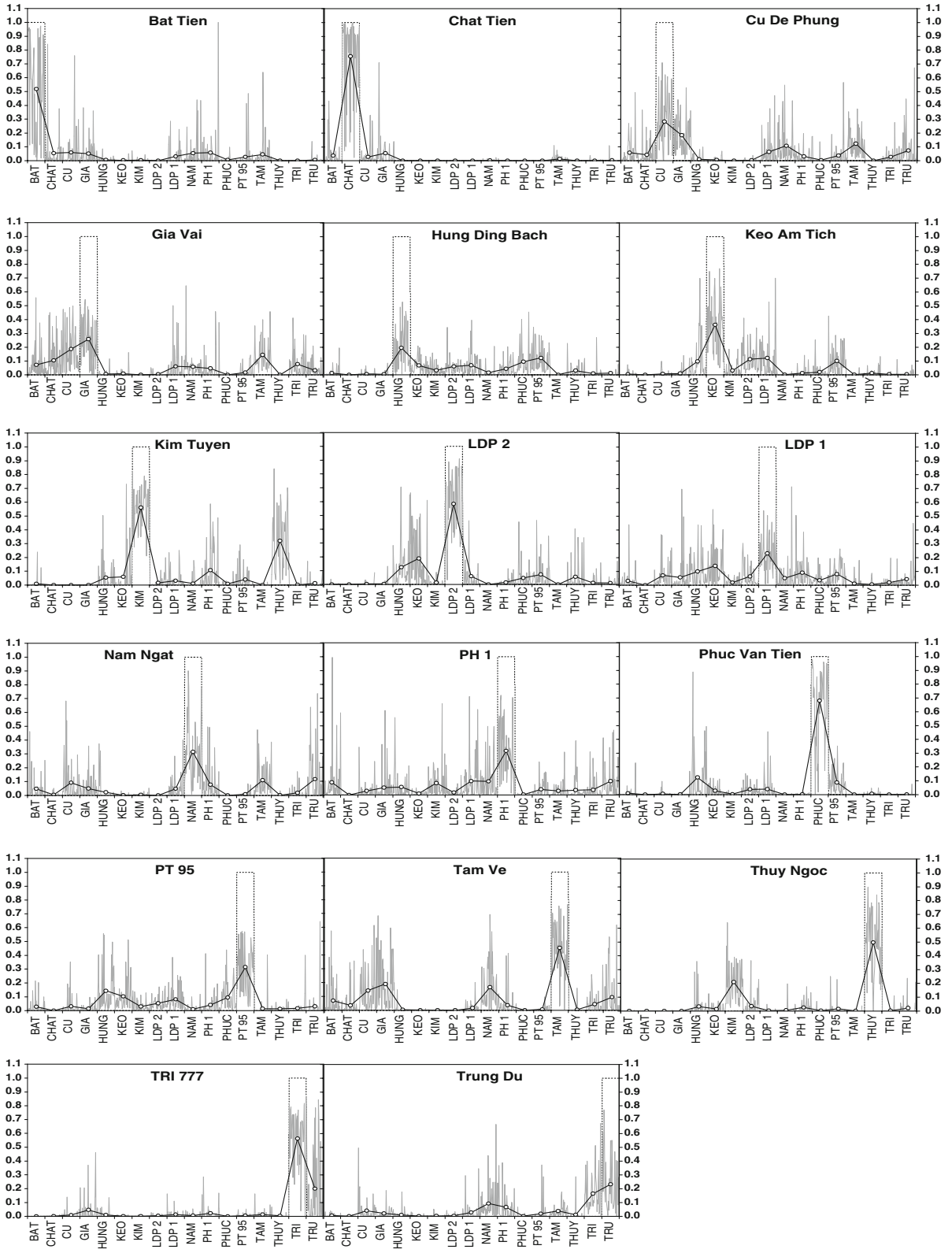
Discussion

From our results, the construction of a dedicated BPNN can be effective to discriminate among different tea varieties through the image analysis of leaves, as almost all the tested accessions were

Table 4 Misidentification matrix produced with the values of identification for each species

	BAT	CHAT	CU	GIA	HUNG	KEO	KIM	LDP2	LDP1	NAM	PH1	PHUC	PT95	TAM	THUY	TRI	TRU
BAT	**0.52**	0.04	0.06	0.07	0.01	0.01	0.01		0.03	0.05	0.09	0.01	0.03	0.07			0.01
CHAT	0.05	**0.76**	0.04	0.1			0							0.04			
CU	0.06	0.03	**0.28**	0.19	0.01	0.01			0.07	0.09	0.03	0.01	0.03	0.14		0.01	0.04
GIA	0.05	0.06	0.18	**0.26**		0.01			0.06	0.05	0.05		0.01	0.19		0.05	0.02
HUNG	0.01		0.01	0.01	**0.19**	0.1	0.05	0.13	0.1	0.02	0.06	0.12	0.14		0.03	0.01	0.01
KEO			0.01	0.01	0.07	**0.36**	0.06	0.19	0.14		0.01	0.03	0.1		0.01		
KIM					0.03	0.03	**0.56**	0.01	0.02		0.09		0.03		0.21		
LDP2					0.06	0.11	0.02	**0.58**	0.06		0.02	0.04	0.05		0.04		
LDP1	0.03		0.06	0.06	0.07	0.12	0.03	0.06	**0.23**	0.05	0.1	0.04	0.08	0.02		0.01	0.03
NAM	0.05		0.11	0.06	0.02		0.01		0.05	**0.31**	0.1		0.01	0.17	0.01	0.01	0.09
PH1	0.06		0.03	0.05	0.04	0.01	0.11	0.02	0.09	0.08	**0.32**		0.04	0.04	0.03	0.02	0.07
PHUC	0.01				0.1	0.02	0.01	0.05	0.04			**0.68**	0.1	0	0		
PT95	0.03		0.04	0.02	0.12	0.1	0.04	0.07	0.08	0.01	0.04	0.09	**0.32**	0.01	0.02	0.01	0.02
TAM	0.05	0.01	0.12	0.14	0	0	0		0.01	0.11	0.03		0.01	**0.45**		0.01	0.04
THUY					0.03	0.01	0.32	0.06	0.01		0.04	0.01	0.01		**0.5**		0.01
TRI			0.03	0.08	0.01			0.01	0.02	0.02	0.04	0	0.02	0.04		**0.56**	0.16
TRU	0.01		0.07	0.03	0.01		0.01	0.01	0.04	0.12	0.1	0	0.03	0.09	0.02	0.2	**0.23**
%Conf	56.61	84.92	26.99	24.02	25.05	40.38	45.43	49.19	21.91	34.63	28.68	66.55	30.73	35.98	57.71	63.55	31.72

The rows refer to the species in the test set. Similarly, the columns report the species to which the test plants are referred by the neural network. Identification average values are shown relative to the total samples of the row test species that are identified as belonging to the corresponding column species. Correct identifications are shown in bold. On the bottom row, the matrix shows the confidence of correct identification (%Conf)


Bat Tien
Chat Tien
Cu De Phung
Gia Vai
Hung Ding Bach
Keo Am Tich
Kim Tuyen
LDP 2
LDP 1
Nam Ngat
PH 1
Phuc Van Tien
PT 95
Tam Ve
Thuy Ngoc
TRI 777
Trung Du
BAT CHAT CU GIA HUNG KEO KIM LDP 2 LDP 1 NAM PH 1 PHUC PT 95 TAM THUY TRI TRU
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1

◀ **Fig. 2** Output *graphs* obtained by the BPNN. Each frame is dedicated to a specific accession and shows the BPNN output for the input represented by the *phyllometric parameters* of 40 leaves. *Reported lines* show the averaged output data

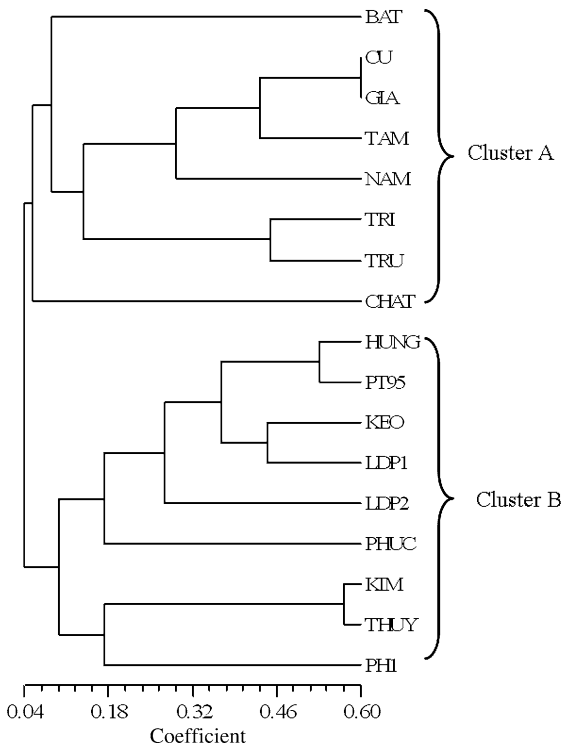


Fig. 3 The dendrogram obtained from the UPGMA cluster analysis of the 17 tea accessions. The coefficients of similarity were calculated through the NTSYS 2.1 software

univocally differentiated by the network. Plant identification by a BPNN was previously performed in olive by Mancuso and Nicese (1999), in chestnut by Mancuso et al. (1999), in *Rollinia* by Mariño and Tressens (2001), in grapevine by Mancuso (1999), in *Tilia* spp. by Clark (2004) and in *C. japonica* by Mugnai et al. (2008). The BPNN also proved to be a powerful tool in order to detect the similarities among the morphotypes through the construction of a dendrogram by the UPGMA method, so the relationships among the 17 varieties have been better clarified (Fig. 3). Cluster A contains mainly Shan Tea plants with a sub-cluster formed by all the Shan tea varieties. This is an important result, due to the fact that the taxonomic position of Shan tea is still controversial and under debate. In fact, some authors considered Shan teas as varieties of *C. sinensis* (*C. sinensis* var. *shan*, Cohen-Stuart (1919) cited by Yamamoto et al. 1997), some others a subspecies (*C. sinensis* var. *pubilimba*; Chang 1981, 1984) as happens in Viet Nam (TRI, personal communication). In the cluster A, CHAT is positioned quite far from the other accessions (CU, GIA, TAM and NAM); in this cluster the dendrogram also positioned TRI, an Assamic tea, next to TRU which belong to the Chinese teas. This apparent discrepancy can be explained by the fact that TRU is a Vietnamese local variety, traditionally propagated by seed, method that lead to high heterogeneous morphological characters. Cluster B includes the majority of the Chinese morphotypes and the hybrids between Chinese and Assamic types. HUNG and PT95 are very similar (0.53): these varieties both belong to the so-called Chinese 'small leaf' type (*C. sinensis* fo. *parvifolia*) and were first selected in China, then imported in Viet Nam. LDP2 and LDP1 had a coefficient of 0.27, and seem to be the offspring of the same parents, a Chinese cultivar (DBT) and PH1. Also KIM and THUY denoted a good similarity (0.57). In fact, these varieties are hybrids between unknown parents (probably Chinese ones) both created for Oolong teas which are characterized by some common morphological feature. Lastly, PH1 is an Assamic variety, known as the first created by the Viet Nam Tea Research Institute more than 20 years ago; it has been propagated by seed during the years, causing a massive hybridization with some correspondence in morphological features with Chinese type.

In general, we can assess that both the morphological analysis and the creation of a neural network were able to characterize, univocally recognize and associate most of the accessions in the correct clusters without a molecular analysis, as previously done by Mondal (2002) and Yao et al. (2008). For example, Mondal (2002) analyzed some cultivars belonging to the three main groups of tea (Chinese, Assamic and Cambodian) through the simple sequence repeat anchored PCR (SSR-anchored PCR). The obtained dendrogram showed a division of the selected accessions in three main clusters, one for each tea type: a marked genetic separation between the different geographic groups was found, partially confirming our results.

The limitations of the BPNN method are largely the same as those of a human expert, namely that success depends on the quantity, validity, and

accuracy of training data. It is well-known that neural networks train best and learn to generalize best when presented data rich in variation. In our case, the creation of an artificial neural network based on leaves morphometric parameters can lead to an effective genotype recognition, even though particular care must be directed to the choice of the plant material, which must be healthy and well-developed. Moreover, further studies must be directed in order to evaluate the effect of the environment on the morphological plasticity of the plants, as the current study is essentially devoted to the characterization of tea varieties in the Phu Tao environment. Concluding, in the present work the application of a BPNN is proposed as a complementary method of botanical identification, being capable to separate almost all the tested tea accessions and to create good associations between the accessions with the same origin. More, this technique represents a economic alternative to the genetic methods commonly used for cultivar discrimination: the need of very simple dedicated instruments, such as an optical scanner and a personal computer, could be a leading reason in order to spread this method in the developing countries.

Acknowledgments The Authors would like to thank Mr. Nguyen Huu La, Head of the Department for Genetic Resources of the Tea Research Institute of Viet Nam (TRI), and Mr. Nguyen Van Tao, Director of for the Tea Research Institute of Viet Nam (TRI), for their technical support and assistance.

References

Chang HT (1981) A taxonomy of the genus *Camellia*. Acta Sientiarum Naturalium Universitatis Sunyatseni 1–124

Chang HT (1984) A revision of the tea resource plants. Acta Sientiarum Naturalium Universitatis Sunyatseni 1–12

Chen L, Yamaguchi S (2005) RAPD markers for discriminating tea germplasm at the inter-specific level in China. Plant Breed 124:404–409. doi:10.1111/j.1439-0523.2005.01100.x

Chen L, Yao MZ, Yang YJ, Yu YJ (2006a) Collection, conservation, evaluation and utilization of tea genetic resources (*Camellia* spp.) in China. Floriculture, ornamental and plant biotechnology: advances and topical issues. Glob Sci Books UK 1:578–583

Chen L, Yao MZ, Zhao LP, Wang XC (2006b) Recent research progresses on molecular biology of tea plant (*Camellia sinensis*). Floriculture, ornamental and plant biotechnology: advances and topical issues. Glob Sci Books UK 4:425–436

Clark JY (2004) Identification of botanical specimens using artificial neural networks. Proceedings of the 2004 IEEE symposium on computational intelligence in bioinformatics and computational biology, La Jolla, USA, 7th–8th October 2004, pp 87–94

Kaundun SS, Matsumoto S (2002) Heterologous nuclear and chloroplast microsatellite amplification and variation in tea, *Camellia sinensis*. Genome 45:1041–1048. doi:10.1139/g02-070

Kaundun SS, Zhyvoloup A, Park YG (2000) Evaluation of the genetic diversity among elite tea (*Camellia sinensis* var. *sinensis*) genotypes using RAPD markers. Euphytica 115:7–16. doi:10.1023/A:1003939120048

Lu CY, Liu WH, Li MJ (1992) Relationship between the evolutionary relatives and the variation of esterase isozymes in tea plant. J Tea Sci 12:15–20

Mancuso S (1999) Fractal geometry-based image analysis of grapevine leaves using the box counting algorithm. Vitis 38:97–100

Mancuso S, Nicese FP (1999) Identifying olive (*Olea europaea* L.) cultivars using artificial neural networks. J Am Soc Hortic Sci 124:527–531

Mancuso S, Ferrini F, Nicese FP (1999) Chestnut (*Castanea sativa* L.) genotype identification: an artificial neural network approach. J Hortic Sci Biotechnol 74:777–784

Mariño SI, Tressens SG (2001) Artificial neural networks application in the identification of three species of *Rollinia* (*Annonaceae*). Ann Bot Fenn 38:215–224

Mondal TK (2002) Assessment of genetic diversity of tea (*Camellia sinensis* (L.) O. Kuntze) by inter-simple sequence repeat polymerase chain reaction. Euphytica 128:307–315. doi:10.1023/A:1021212419811

Morgan A, Boddy L, Mordue JEM, Morris CW (1998) Evaluation of artificial neural networks for fungal identification employing morphometric data from spores of *Pestalotiopsis* species. Mycol Res 102:975–984. doi:10.1017/S0953756297005947

Mugnai S, Pandolfi C, Azzarello E, Masi E, Mancuso S (2008) *Camellia japonica* L. genotypes identified by an artificial neural network based on phyllometric and fractal parameters. Plant Syst Evol 270:95–108. doi:10.1007/s00606-007-0601-7

Pandolfi C, Mugnai S, Azzarello E, Masi E, Mancuso S (2006) Fractal geometry and neural networks for the identification and characterization of ornamental plants. In: Teixeira da Silva J (ed) Floriculture, ornamental and plant biotechnology: advances and topical issues, vol IV. GlobalScience Books, Kyoto, pp 213–225

Paul S, Wachira FN, Powell W, Waugh R (1997) Diversity and genetic differentiation among populations of Indian and Kenyan tea (*Camellia sinensis* (L.) O. Kuntze) revealed by AFLP markers. Theor Appl Genet 94:255–263. doi:10.1007/s001220050408

Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back propagating errors. Nature 323:533–536. doi:10.1038/323533a0

Saravanan M, Maria John KM, Raj Kumar R, Pius PK, Sasikumar R (2005) Genetic diversity of UPASI tea cones (*Camellia sinensis* (L.) O. Kuntze) on the basis of total catechins and their fractions. Phytochemistry 66:561–565. doi:10.1016/j.phytochem.2004.06.024

Sealy J (1958) A revision of the genus *Camellia*. Royal Hort Soc, London

Ueno S, Yoshimaru H, Tomaru N, Yamamoto S (1999) Development and characterization of microsatellite markers in *Camellia japonica* L. Mol Ecol 8:335–336. doi:10.1046/j.1365-294X.1999.00534.x

Wight W (1962) Tea classification revised. Curr Sci 31: 298–299

Willson KC, Clifford MN (1992) Tea: cultivation to consumption. Chapman & Hall, London

Yamamoto T, Juneja LR, Chu DC, Kim M (1997) Chemistry and application of green tea. CRC Press, Boca Raton, FL

Yao MZ, Chen L, Liang YR (2008) Genetic diversity among tea cultivars from China, Japan and Kenya revealed by ISSR markers and its implication for parental selection in tea breeding programmes. Plant Breed 127:166–172. doi: 10.1111/j.1439-0523.2007.01448.x

Zurada JM, Malinowski A (1994) Multilayer perceptron networks: selected aspects of training optimization. Appl Math Comp Sci 4:281–307