Automatic identification of species with neural networks

A new automatic identification system using photographic images has been designed to recognize fish, plant, and butterfly species from Europe and South America. The automatic classification system integrates multiple image processing tools to extract the geometry, morphology, and texture of the images. Artificial neural networks (ANNs) were used as the pattern recognition method. We tested a data set that included 740 species and 11,198 individuals. Our results show that the system performed with high accuracy, reaching 91.65% of true positive fish identifications, 92.87% of plants and 93.25% of butterflies. Our results highlight how the neural networks are complementary to species identification, which is useful in today´s taxonomic crisis.

PeerJ Reviewing Manuscript

Automatic identification of species with neural networks.

Andrés Hernández-Serna [a, b] * and Luz Fernanda Jiménez-Segura [a]

[a] Grupo de Ictiología; Instituto de Biología, Universidad de Antioquia, Medellín, Colombia

[b] Department of Biology, University of Puerto Rico-Río Piedras, San Juan, Puerto Rico.

*Corresponding autor: Andres Hernández-Serna, PO Box 23360, Department of Biology, University of Puerto Rico, San Juan, PR 00931-3360, USA

e-mail: andres137@gmail.com

ABSTRACT

A new automatic identification system using photographic images has been designed to recognize fish, plant, and butterfly species from Europe and South America. The automatic classification system integrates multiple image processing tools to extract the geometry, morphology, and texture of the images. Artificial neural networks (ANNs) were used as the pattern recognition method. We tested a data set that included 740 species and 11,198 individuals. Our results show that the system performed with high accuracy, reaching 91.65% of true positive fish identifications, 92.87% of plants and 93.25% of butterflies. Our results highlight how the neural networks are complementary to species identification, which is useful in today´s taxonomic crisis.

Keywords: Fish, plant, butterflies, neural network, feature extraction, digital image, and species

INTRODUCTION

PeerJ Reviewing Manuscript

Currently, species identification is a taxonomic challenge and an integral process of all biological research, which generates important information for biodiversity conservation. Difficulty identifying species and ambiguity in the species concept, are seriously affecting our ability to estimate levels of biodiversity (Gaston & O'Neill, 2004). The Global Taxonomy Initiative highlights the knowledge gaps in our taxonomic system due to the shortage of trained taxonomists and curators; these deficiencies reduce our ability to understand, use, and conserve biological diversity. High levels of global biodiversity and a limited number of taxonomists represents significant challenges to the future of biological study and conservation. The main problem is that 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. This lack of accessibility of taxonomic knowledge to the general public has been termed the “taxonomic crisis” (Dayrat, 2005).

Recently, taxonomists have been searching for more efficient methods to meet species identification requirements, such as developing digital image processing and pattern recognition techniques. These methods automatically identify species based on extracting unique image shape information that distinguishes them by taxonomic groups. Researchers currently have recognition techniques for insects, plants, spiders, and plankton (Gaston & O'Neill, 2004). This approach can be extended even further to field-based identification of organisms such as fish (Strachan, Nesvadba & Allen, 1990; Storbeck & Daan, 2001; White, Svellingen & Strachan, 2006; Zion, Alchanatis, Ostrovsky, Barki & Karplus, 2007; Hu, Li, Duan, Han, Chen & Si, 2012), insects (Mayo & Watson, 2007; O'Neill, 2007 ; Kang, Song & Lee, 2012), zooplankton (Grosjean, Picheral, Warembourg & Gorsky, 2004) and plants (Novotny & Suk, 2013). These

PeerJ Reviewing Manuscript

methods are helpful in alleviating the "taxonomy crisis". In this research, we present a new methodology for the identification of different taxonomic groups to the species level for fish, plants, and butterflies.

We designed a simple and effective algorithm (preprocess solution) and defined a range of new features that use pattern recognition with artificial neural network designs (ANN). Our experiments are outlined, discussed, and important conclusions on automatic species image identification are summarized.

MATERIALS AND METHODS

Images

Image data in this study was taking from two sources: natural history museum records, and online databases. Analyses from each collection were done with respect to country. Ichthyology collections from Colombia were compiled from the Instituto de Investigaciones Marinas y Costeras (INVEMAR), the Colección de Referencia Biología Marina Universidad del Valle (CRBMUV), and the Coleccion Ictiologica Universidad de Antioquia (CIUA). Ichthyology collections from Brazil were found in the Museu de Zoologia da USP (MZUSP), the Instituto Nacional de Pesquisas da Amazônia Manaus (INPA), and the Museu Nacional Rio de Janeiro (MNRJ). Image data from Spain came from the Museo Nacional de Ciencias Naturales Madrid (MNCN). We tested a data set that included a total of 740 species and 11,198 individuals of fish, plants, and butterflies. Fish specimen images were taken using a Cannon EOS 6dD one-use camera with a 1280 x 960 pixel resolution. 697 total fish species, previously identified by experts, were photographed (see Fig. 1 for a subset of photographed species). Images of 32 plant species were downloaded from the Flavia database (http://flavia.sourceforge.net/) (see Fig.2).

PeerJ Reviewing Manuscript

Image data for 11 species of butterflies were downloaded from the MorphBank database (http://www.morphbank.net/) (see Fig. 3.).

System development

Based on pattern recognition theory (Marqués de Sá, 2001) and basic computer-processing pathways used in typical automated species identification systems (Gaston & O'Neill, 2004), we designed a system for automatic individual identification at the species level (Fig. 4). In a novel way, our system shares preprocess and extraction components with both the training and recognition processes. Features of training images are used to build a model of the classification progress pattern after feature extraction. These features and the trained model are then recorded in the database and incorporated in the analysis of subsequent photos. This process uses two types of data to model features of recognition files and results in better species identification results. The following sections provide implementation details for each step in Fig. 4. Due to its size, a list of features could not be included in this manuscript, but is alternatively available upon request.

Image preprocessing

Image heterogeneity in terms of orientation, size, brightness, and illumination was common (Fig. 5.1). Image background was removed with Grabcut's algorithm (Rother, Kolmogorov & Blake, 2004) (Fig. 5.2) and converted to grayscale (Fig. 5.3). Different filters were applied to improve the image by removing image noise; the filters used were smooth and median (Fig. 5.4 and 5.5), and the image was then reduced to one of two possible levels, 0 or 1 (Fig. 5.6). **Next,** the processed image was brought to a contour (Fig. 5.7) and then a skeleton (Fig. 5.8). All of these processes were performed for each taxonomic group using the image processing in MATLAB R2009b.

PeerJ Reviewing Manuscript

Feature extraction

Feature extraction greatly influences species identification from image processing. Features should represent taxonomic information and be easily acquired from data images. A series of geometrical, morphological, and texture features, unique to species, are used in our automatic identification system; these features can be efficiently extracted with image processing. Fifteen intuitive features were used in the system and are described below:

Geometrical

Geometric features contain information about form, position, size, and orientation of the region. The following are some geometric features that are commonly used in pattern recognition.

1- *Area* is the total number of pixels of the study area, and is defined as:

$$A(s) = \int_x \int_y I(x, y)\, dy\, dx$$

I (x, y) depends on the limits of the shape (see figure 5.7).

2- *Perimeter.* The number of pixels that belong to the edge of the region (see figure 5.8). In other words, it is the curve that encloses a region S, defined as:

$$P(s) = \int_t \sqrt{x^2(t) + y^2(t)}\, dt$$

3- *Diameter.* Value representing the diameter of a circle with the same area as the region.

4- *Compatibility.* The efficiency of the contour or perimeter P(s) that encloses an area A(s)

$$C(s) = \frac{4\pi\, A(s)}{P^2(s)}$$

PeerJ Reviewing Manuscript

5- *Compactness.* The efficiency with which area A(s) encloses an object is determined by P(s)

$$Co(s) = \frac{P^2(s)}{4\pi\ A(s)}$$

6- *Solidity*. The scalar specifying the proportion of the pixels in the convex hull that are also in the region. This property is supported only for 2-D input label matrices.

Solidity. The number of pixels, specified in terms of area/scalar.

Texture

Textures are important visual patterns for homogeneous description of regions. Intuitive measures provide properties such as smoothing, roughness, and regularity (Glasbey, 1996). Textures depend on the resolution of the image and can follow two approaches: statistical and frequency. We use the statistical approximation in which statistical values are analyzed first order (on the histogram) and second order (on the co-occurrence matrix).

Statistical first order is obtained from the gray level histogram of the image. Each value is divided by the total number of pixels (area) and has a new histogram representing the probability that a determined gray level is displayed in the region of interest.

Obtained properties:

7- *Median*

$$\mu = \sum_{x=1}^{n} x h(x)$$

8-*Variance*

PeerJ Reviewing Manuscript

$$\delta^{2} = \sum_{x=1}^{n} \left(x - \mu^{2}\right) h(x)$$

Statistical second order is the matrix spatial dependence of gray levels or co-occurrence matrices.

Given a vector of polar coordinates, $\delta = (r, \theta)$, one can calculate the conditional probability that two properties appear separated by a given distance δ, P_{δ} using an angle θ of -45 and a distance *r* equal to one pixel. The features that are extracted from this matrix are:

9- *Uniformity*

$$\sum_{x=1}^{n} \sum_{y=1}^{n} P_{\delta}(x, y)^{2}$$

10- *Entropy co-occurrence*

$$-\sum_{x=1}^{n} \sum_{y=1}^{n} P_{\delta}(x, y) \log P_{\delta}(x, y)$$

11- *Homogeneity*

$$\frac{\sum_{x=1}^{n} \sum_{y=1}^{n} P_{\delta}(x, y)}{1 + |x - y|}$$

12- *Inertia*

$$\sum_{x=1}^{n} \sum_{y=1}^{n} P_{\delta}(x, y)(x - y)^{2}$$

PeerJ Reviewing Manuscript

Morphological

The morphological features are those that concentrate on the organization of pixels. They perform a comprehensive description of the region of interest. They fall into two categories: two-dimensional Cartesian moments and normalized central moments.

Two-dimensional Cartesian moments are variable at minor order, and initiate at zero at higher orders. The moment of order p and q of a function I (x, y) is defined as:

$$m_{pq} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^p y^q \, I(x,y) \, dx \, dy$$

The parameters *p* and *q* denote the order of the moment. When *p* = 0 and *q* = 0, which determines the center of mass or gravity of the overall function in binary images, the center of mass or gravity of the region under study is:

$$\bar{x} = \frac{m_{10}}{m_{00}} \quad \bar{y} = \frac{m_{01}}{m_{00}}$$

The center of mass or gravity can define the central moments that are invariant to displacement or translation of the image’s region of interest defined as:

$$u_{pq} = \sum_x \sum_y (x - \bar{x})^p \, (y - \bar{y})^q \, I(x,y) \Delta A$$

Where ΔA is the area of a pixel.

Normalized central moments are invariant to scale which is defined as:

PeerJ Reviewing Manuscript

$$n_{pq} = \frac{u_{pq}}{u^{\gamma}{}_{00}}$$

$$\gamma = \frac{p+q}{2} \quad \forall\, p+q \geq 2$$

Where

The above equations can be defined by seven moments that are invariant to rotation, translation, and scale changes, known as the Hu invariant set of moments (Hu, 1962). In this study, we used the first Hu moment defined as:

13-*Hu1*

$$\varphi_1 = m_{20} + m_{02}$$

Normalized central moments can be generated by related moment invariants "AMI" (Flusser & Suk, 1993), based on the theory of algebraic invariants and invariants under general affine transformation. We used two of the four invariants associated with discriminant character moments defined as:

14-*Ami1*

$$I_1 = \frac{u_{20}\, u_{02} - u_{11}^2}{u_{00}^4}$$

15- *Ami2*

$$I_2 = \frac{u_{30}^2 u_{03}^2 - 6u_{30}u_{21}u_{12} + 4u_{30}u_{12}^3 + 4u_{21}^3 u_{03} - 3u_{21}^2 u_{12}^2}{u_{00}^{10}}$$

PeerJ Reviewing Manuscript

These moments enable a high degree of insensitivity to noise that is not altered by rotation, translation, or staggering.

The use of the above 15 features (Table 1) has two advantages. First, the features can express the structure of the individual's body, which is important for the identification at species level. Second, our features were elaborately chosen to avoid using feature optimization methods like adapted fuzzy reasoning (Lancieri & Boubchir, 2007). We designed and realized automatic extraction algorithms to compute the values of these features so that all variables and features could be calculated automatically.

Neural Network

A neural network is defined as a parallel computer model composed of a large number of adaptive processing (neural) units which communicate via interconnections with variables. A multiple layer network has one or more layers (neurons) that enable the learning of complex tasks by progressively extracting more meaningful features from the input image patterns (Wu, 1997). Compared to other machine learning methods, neural networks learn slower but predict faster and have very good models presenting nonlinear data. The simple perceptron is assigned multiple inputs but generates a single output, similar to different linear combinations that depend on input weights and generate a linear activation function (Rosenblatt, 1958). Mathematically, the neural network can be described with the following equation:

$$y = \varphi\left(\sum_{i=1}^{n} w_i * x_i + b\right)$$

$\boldsymbol{W_i}$: weight vector, $\boldsymbol{X_i}$: input vector, $\boldsymbol{b}$: bias activation function.

PeerJ Reviewing Manuscript

A multilayer perceptron consists of a set of source nodes containing one or more input layer and a set of hidden-node outputs. The input signal propagates through the network layer by layer (Zhang, Patuwo & Hu, 1998). Fig. 6 presents a diagram of the multilayer neural network.

The neural network structure is composed of N inputs $N = [N_1, N_2, ..., N_n]$, a hidden layer h and an output vector $\boldsymbol{S} = [S_1, S_2, ..., S_{\boldsymbol{m}}]$. Each S_i is assessed by a single step that transforms the vector S binary signal [0,1]. A supervised training phase, or sigmoid activation, is based on the back propagation algorithm in which the weights and biases are updated in the direction of the negative gradient of the performance and then updated in the opposite direction(Werbos, 1974; Rumelhart, Hinton & Williams, 1986; Parker, 1987; Smith & Brier, 1996;). The sigmoid activation function for the hidden layer and output layer is determined by the following equation:

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

In this study, the number of input neurons is determined by the number of descriptors that are available in each pattern, which in this case is N=15 (see variables section). The number of neurons in the hidden layer, *h,* has been experimentally determined from the error set data searching for the general training date of the ANN. The number of output neurons is determined by the number of species classified in each database.

RESULTS AND DISCUSSION

All features were extracted from images and defined according to the above mentioned methods. We tested different species from various taxonomic groups, using the developed neural network systems. The results of the main tests with different test species are listed below.

PeerJ Reviewing Manuscript

Experiments were divided into two groups: 1) images from the training group were used for building the classifications of the model; 2) images from the test group were used for the reorganization and testing of the developed model.

To determine the optimal number of neurons given a data image, the relationship between the identification success rate and the number of neurons was explored. Fig. 7 shows this relationship for the different configurations considered. We display values for neuron number for each species in the database in Table 2.

Table 3 shows the performance average of the artificial neural networks using image data and the 15 analyzing features. The data set was randomly divided into 60-70-80-90% training images, resulting in 40-30-20-10% test images. The results with the highest average accuracy for species identification were networks using 80-90% training and 20-10% test images. For these tests, the declared success rate was related to the number of species. Recognition became more difficult with increased species number, as observed in the fish result collections from MZUSP, INPA, INVEMAR, CRBMUV, and MNCN which averaged below 90% recognition.

Similar to previous findings (Strachan et al., 1990; Storbeck & Daan, 2001; White et al., 2006; Zion et al., 2007; Novotny & Suk, 2013), the neural network used classified species from image data. However, most other studies only employ databases with low levels of species richness usually spanning many different orders and families and are easily classified due to distinct differences in morphological characteristics. Our neural network builds on the work of these networks, and requires low operator expertise, costs, and response time, but also offers high reproducibility, species identification accuracy, and usability. The ANN algorithm is optimized for testing datasets with high levels of species richness, in this case 740 species (11,198 individuals) of fishes, plants and butterflies.

The predictive ability of the ANNs was affected by the high phenotypic similarity between species in the analysis, for example small fish species such as those from the family Characidos (Annex 1, Fig. 8). The magnitude of this error comes from low phenotypic differences of some species that vary only in minor details, like teeth or fin radii, which hinders classification. However, the error obtained on the neural network model has been low in other taxonomic families (Table 3). Overall performance of the system achieved high accuracy and precision, with 91.65% true positive fish identifications, 92.87% plant identifications, and 93.25% butterfly identifications. The evaluation of results appears simple at first glance: the comparison of success rates appears sufficient, however upon closer examination, the success rates in tests on closed data sets strongly depend on the number of species and the ratio of test to training image samples. The data sets with a lower species number have higher success rates, possibly explained by species with very distinct morphological characteristics.

The strength of this research is in its applicability to combat the "taxonomic crisis". In the past three decades, many promising techniques for fish identification have emerged. Many of them are based on genetics, interactive computer software, image recognition, hydro-acoustics, and morphometrics (Fischer, 2013). In our study, neural networks were tested as a possible method for species identification. However, taking advantage of the fast performance of the ANNs and the speed of modern PCs, further research should explore the applications of the ANN methodology to automate biomass estimation and real-time species classifications. This could produce useful tools for both scientific and commercial use. Fischer (2013) concludes that the image recognition methods are useful but their transferability and resolution are poor because species differ between geographic regions. This is a clear obstacle to future ANN development and network identification success. Our advances in this field in relation to species identification should be developed for specific geographic regions and translated into user-friendly

PeerJ Reviewing Manuscript

applications. We support the development of species identification methods that are globally interchangeable but also tailored to regional biodiversity composition.

ACKNOWLEDGEMENTS

We thank to Dr Paulo A. Buckup (Museu Nacional Rio de Janeiro), Dr José Luís Olivan Birindelli (Museu de Zoologia da Universidade de São Paulo, Sao Paulo-Brasil), Dr Rosseval Leite (Instituto Nacional de Pesquisas da Amazônia, Manaus-Brasil), Dra Gema Solís (Museo Nacional de Ciencias Naturales, Madrid-España), Dr Efrain Rubio Rincon (Colección de Referencia Biología Marina, Universidad del Valle, Cali-Colombia), and Dra Andrea Polanco (Instituto de Investigaciones Marinas y Costeras, Santa Marta- Colombia) to allow to photograph the fish lots, Cesar Uribe for his helping on data analysis, Jonathan Bustamante Alvarez through the photos on figure 1 (Universidad de Antioquia). Thanks are extended to Benjamin Branoff, Aaron Hogan, and Paul Furumo for reviewing the English.

REFERENCES

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

Erickson G, Jörgensen P, Jörgensen C, Riccardi G, Ronquist F, van Engelen R. 2007. Morphbank: Web Image Database Technology for Comparative Morphology and Biodiversity Research *http://www.morphbank.net/* (accessed on 18 October 2013)

Fischer J. 2013. Fish identification tools for biodiversity and fisheries assessments: review and guidance for decision-makers. *FAO, Rome ed: FAO Fisheries and Aquaculture* **107**.

Flavia. (2009). Flavia at a glance. *http://flavia.sourceforge.net/* (accessed on 28 October 2013)

Flusser J, Suk T. 1993. Pattern-recognition by affine moment invariants. *Pattern Recognition* **26:** 167-174.

PeerJ Reviewing Manuscript

PeerJ Reviewing Manuscript

Gaston KJ, O'Neill MA. 2004. Automated species identification: why not? *Royal Society Philosophical Transactions Biological Sciences* **359:** 655-667.

Glasbey C. 1996. Handbook of pattern recognition and computer vision - Chen,CH, Pau,LF, Wang,PSP. *Journal of Classification* **13:** 350-352.

Grosjean P, Picheral M, Warembourg C, Gorsky G. 2004. Enumeration, measurement, and identification of net zooplankton samples using the ZOOSCAN digital imaging system. *Ices Journal of Marine Science* **61:** 518-525.

Hu J, Li D, Duan Q, Han Y, Chen G, Si X. 2012. Fish species classification by color, texture and multi-class support vector machine using computer vision. *Computers and Electronics in Agriculture* **88:** 133-140.

Hu M-K. 1962. Visual pattern recognition by moment invariants: Information Theory, IRE Transactions. 179 - 187.

Kang S-H, Song S-H, Lee S-H. 2012. Identification of butterfly species with a single neural network system. *Journal of Asia-Pacific Entomology* **15:** 431-435.

Lancieri L, Boubchir L. 2007. Using multiple uncertain examples and adaptative fuzzy reasoning to optimize image characterization. *Knowledge-Based Systems* **20:** 266-276.

Marqués de Sá J. 2001. *Pattern Recognition: Concepts, Methods and Applications*. Springer: Springer.

Mayo M, Watson AT. 2007. Automatic species identification of live moths. *Knowledge-Based Systems* **20:** 195-202.

Novotny P, Suk T. 2013. Leaf recognition of woody species in Central Europe. *Biosystems Engineering* **115:** 444-452.

O'Neill MA. 2007. DAISY: a practical computed-based tool for semi-automated species identification. *Systematics Association Special Volume Series* **74:** 101-114.

Parker DB. 1987. Optimal Algorithms for Adaptive Networks: Second order Back Propagation, Second Order Direct Propagation, and Second Order Hebbian Learning: Proceedings of the IEEE First International Conference on Neural Networks \rm San Diego, CA. 593-600.

Rosenblatt. F. ***1958*** The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review* **65:** *386-408*

Rother C, Kolmogorov V, Blake A. 2004. "GrabCut" - Interactive foreground extraction using iterated graph cuts. *Acm Transactions on Graphics* **23:** 309-314.

Rumelhart DE, Hinton GE, Williams RJ. 1986. Learning representations by back-propagating errors. *Nature* **323:** 533-536.

Smith BP, Brier ME. 1996. Statistical approach to neural network model building for gentamicin peak predictions. *Journal of Pharmaceutical Sciences* **85:** 65-69.

Storbeck F, Daan B. 2001. Fish species recognition using computer vision and a neural network. *Fisheries Research* **51:** 11-15.

Strachan NJC, Nesvadba P, Allen AR. 1990. Fish species recognition by shape-analysis of images. *Pattern Recognition* **23:** 539-544.

Werbos PJ. 1974. Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences: *Harvard University*.

White DJ, Svellingen C, Strachan NJC. 2006. Automated measurement of species and length of fish by computer vision. *Fisheries Research* **80:** 203-210.

Wu CH. 1997. Artificial neural networks for molecular sequence analysis. *Computers & Chemistry* **21:** 237-256.

Zhang GQ, Patuwo BE, Hu MY. 1998. Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting* **14:** 35-62.

PeerJ Reviewing Manuscript

Zion B, Alchanatis V, Ostrovsky V, Barki A, Karplus I. 2007. Real-time underwater sorting of edible fish species. *Computers and Electronics in Agriculture* **56:** 34-45.

PeerJ Reviewing Manuscript

Table 1(on next page)

Table 1

Features extracted

PeerJ Reviewing Manuscript

Table 1. Features extracted

Type	Variable	Description
Geometrical	A	Area
	P	Perimeter
	D	Diameter
	C	Compatibility
	Co	Compactness
	S	Solidity
Texture	u	Median
	δ^2	Variance
	$E_{r,\theta}$	Uniformity
	$H_{r,\theta}$	Entropy co-occurrence
	$HG_{r,\theta}$	Homogeneity
	$I_{r,j}$	Inertia
Morphological	φ_1	Hu1
	I_1, I_2	Ami1-Ami2

PeerJ Reviewing Manuscript

Table 2(on next page)

table 2

FC (Fish collection); parameters used in neural network systems.

PeerJ Reviewing Manuscript

Table 2. FC (Fish collection); parameters used in neural network systems.

Data set	**Learning rate**	**Number of generations**	**Number of Hidden layers**	**Number of input layers**	**Number of output layers (# species)**
FC-MZUSP	0.2	95000	200	15	100
FC-INPA	0.15	100000	180	15	91
FC-MNRJ	0.25	78000	60	15	14
FC-INVEMAR	0.3	84000	250	15	189
FC-CIUA	0.12	90000	60	15	33
FC-CRBMUV	0.35	140000	300	15	172
FC-MNCN	0.2	110000	250	15	98
FLAVIA	0.1	50000	60	15	32
BUTTERFLIES	0. 5	50000	35	15	11

PeerJ Reviewing Manuscript

Table 3(on next page)

Table 3

FC (Fish collection); results of ANN tests with species for 15 features

PeerJ Reviewing Manuscript

Table 3. FC (Fish collection); results of ANN tests with species tests for 15 features

			Average Percentage of images (Training / test)			
Data set	**Species**	**Images**	**60/40**	**70/30**	**80/20**	**90/10**
FC-MZUSP	100	1718	76.67	81.34	83.34	88.31
FC-INPA	91	1640	76.29	78.94	84.44	89.93
FC-MNRJ	14	422	82.62	87.18	90.56	91.65
FC-INVEMAR	189	1703	76.72	84.03	86.45	88.08
FC-CIUA	33	472	83.08	86.99	90.19	91.77
FC-CRBMUV	172	2392	77.36	85.21	87.29	88.85
FC-MNCN	98	959	72.34	86.21	88.15	89.11
FLAVIA	32	1800	68.79	88.48	91.61	92.87
BUTTERFLIES	11	92	73.62	80.43	88.83	93.25

PeerJ Reviewing Manuscript

PeerJ Reviewing Manuscript

Figure 1

Figure 1

Samples of some species data set : 1)*Curimata mivartii 2*)*Leporinus striatus*3)*Ctecolucius hujeta*4)*Cinopotamus magdalenae*5)*Astyanax magdalenae*6)*Roeboides occidentalis*7)*Genycharax tarpon*8)*Cyphocharax magdalenae*9)*Hemibrycon decurrens* 10)*Brycon medemi*11)*Lebiasina multimaculata*12)*Hemibrycon dentatus*13)*Triporheus magdalenae*14)*Characidium phoxocephalum*15)*Leporinus muyscorum*16)*Hemibrycon boquiae*17)*Brycon henni*r18*Characidium caucanum*19)*Roeboides dayi*20)*Astyanax fasciatus*21)*Argopleura magdalenensis*22)*Apteronotus eschemeyeri*23)*Eigenmannia virescens.*

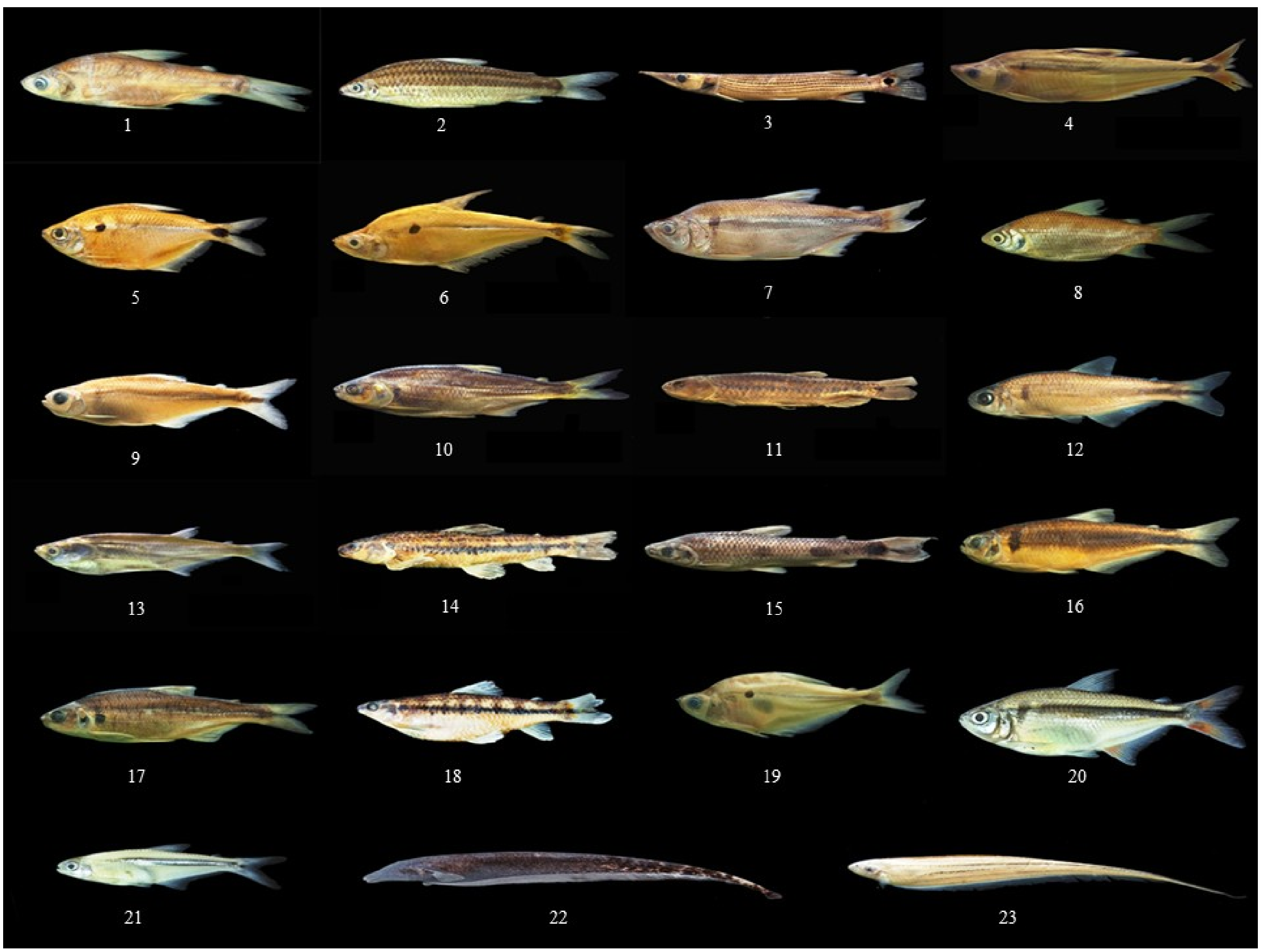
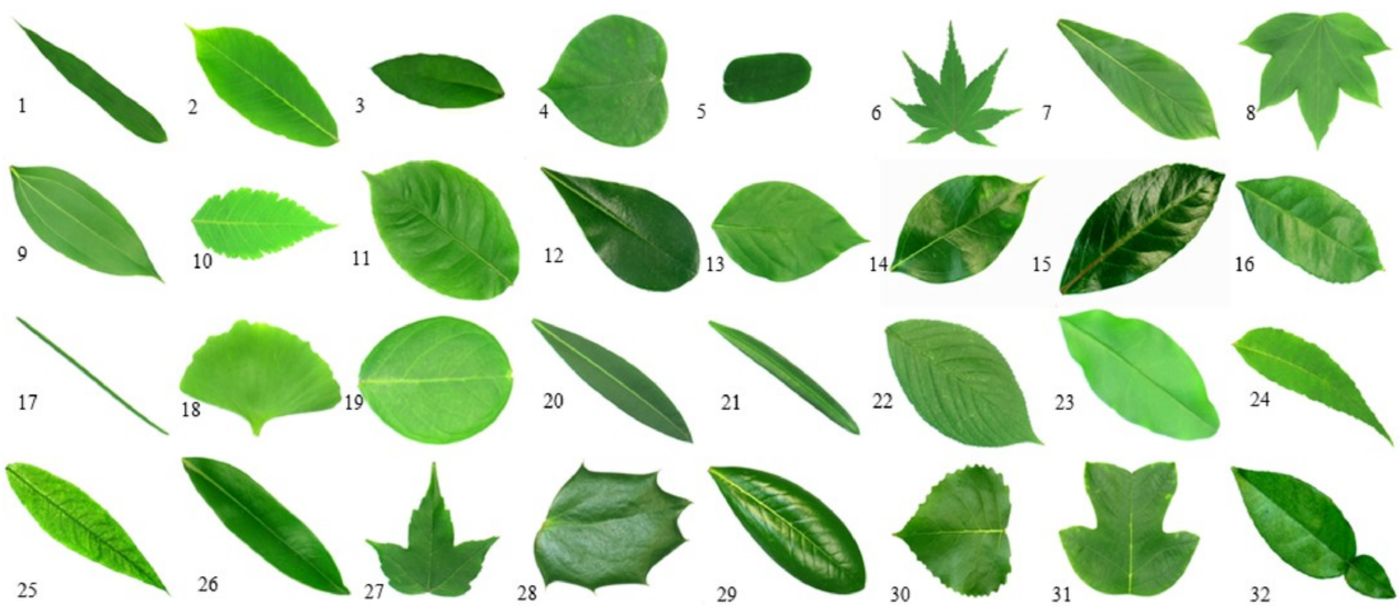


Figure 2

Figure 2

Samples of our data set: 1)*Phyllostachys edulis*2)*Aesculus chinensis*3)*Berberis anhweiensis*4)*Cercis chinensis*5)*Indigofera tinctoria*6)*Acer Dalmatum*7)*Phoebe zhennan*8)*Kalopanax septemlobus*9)*Cinnamomum japonicum*10)*Koelreuteria paniculata*11)*Ilex macrocarpa*12)*Pittosporum tobira*13)*Chimonanthus praecox*14)*Cinnamomum camphora*15)*Viburnum awabuki*16)*Osmanthus fragrans*17)*Cedrus deodara*18)*Ginkgo biloba*19)*Lagerstroemia indica*20)*Nerium oleander*21)*Podocarpus macrophyllus*22)*Prunus yedoensis*23)*Ligustrum lucidum*24)*Tonna sinensis*25)*Prunus persica*26)*Manglietia fordiana*27)*Acer buergerianum*28)*Mahonia bealei*29)*Magnolia grandiflora*30)*Populus Canadensis*31)*Liriodendron chinense*32)*Citrus reticulate*.



PeerJ Reviewing Manuscript

PeerJ Reviewing Manuscript

Figure 3

Figure 3

Samples of our data set: 1) *Agraulis vanillae* 2) *Anthocharis midea* 3) *Ascia monuste* 4) *Danaus gilippus* 5) *Danaus plexippus* 6) *Dryas iulia* 7) *Enodia portlandia* 8) *Glutophrissa Drusilla* 9) *Heliconius charithonia* 10) *Pieres rapae* 11) *Pontia protodice*.

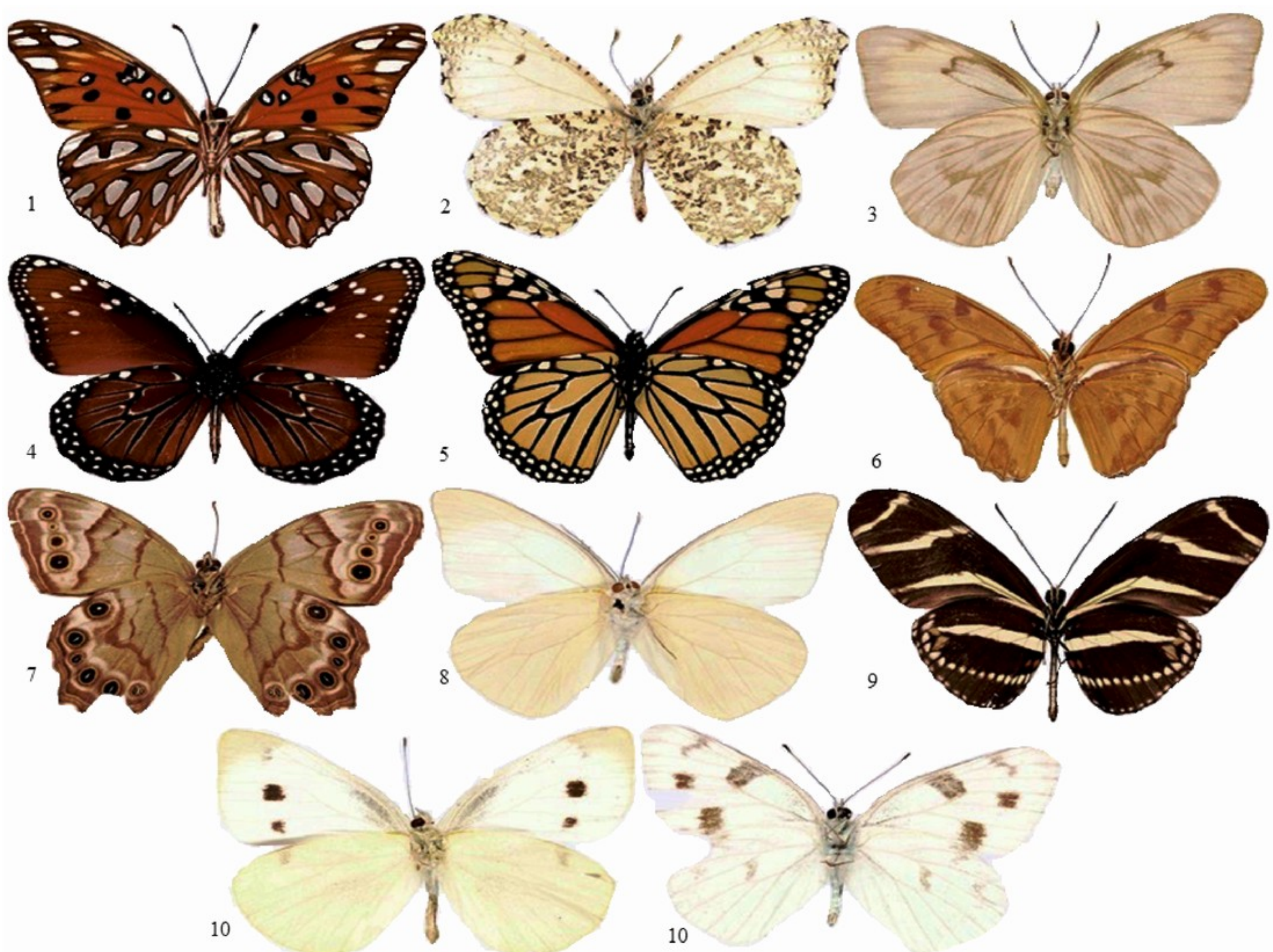
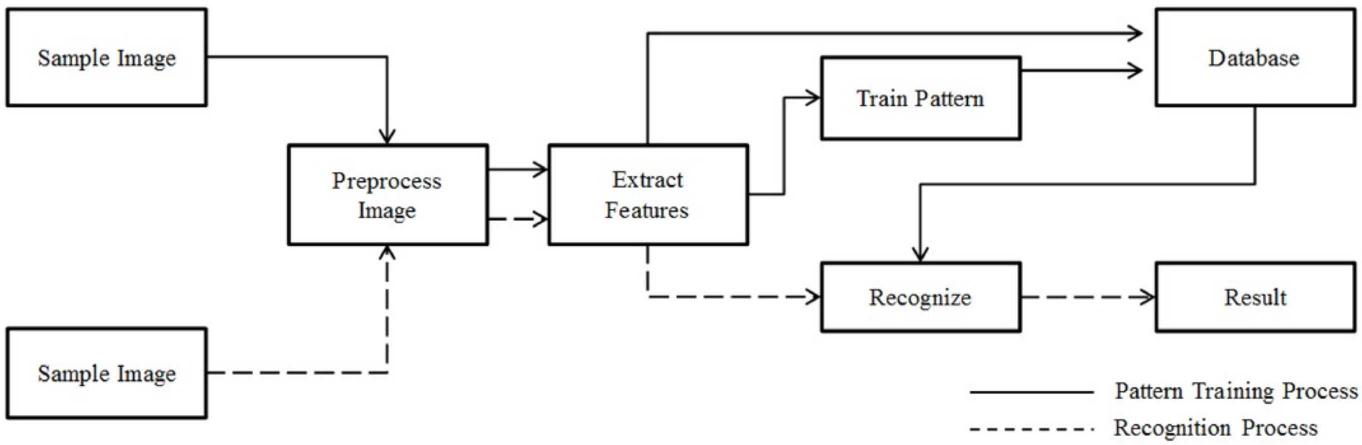


Figure 4

Figure 4

System architecture

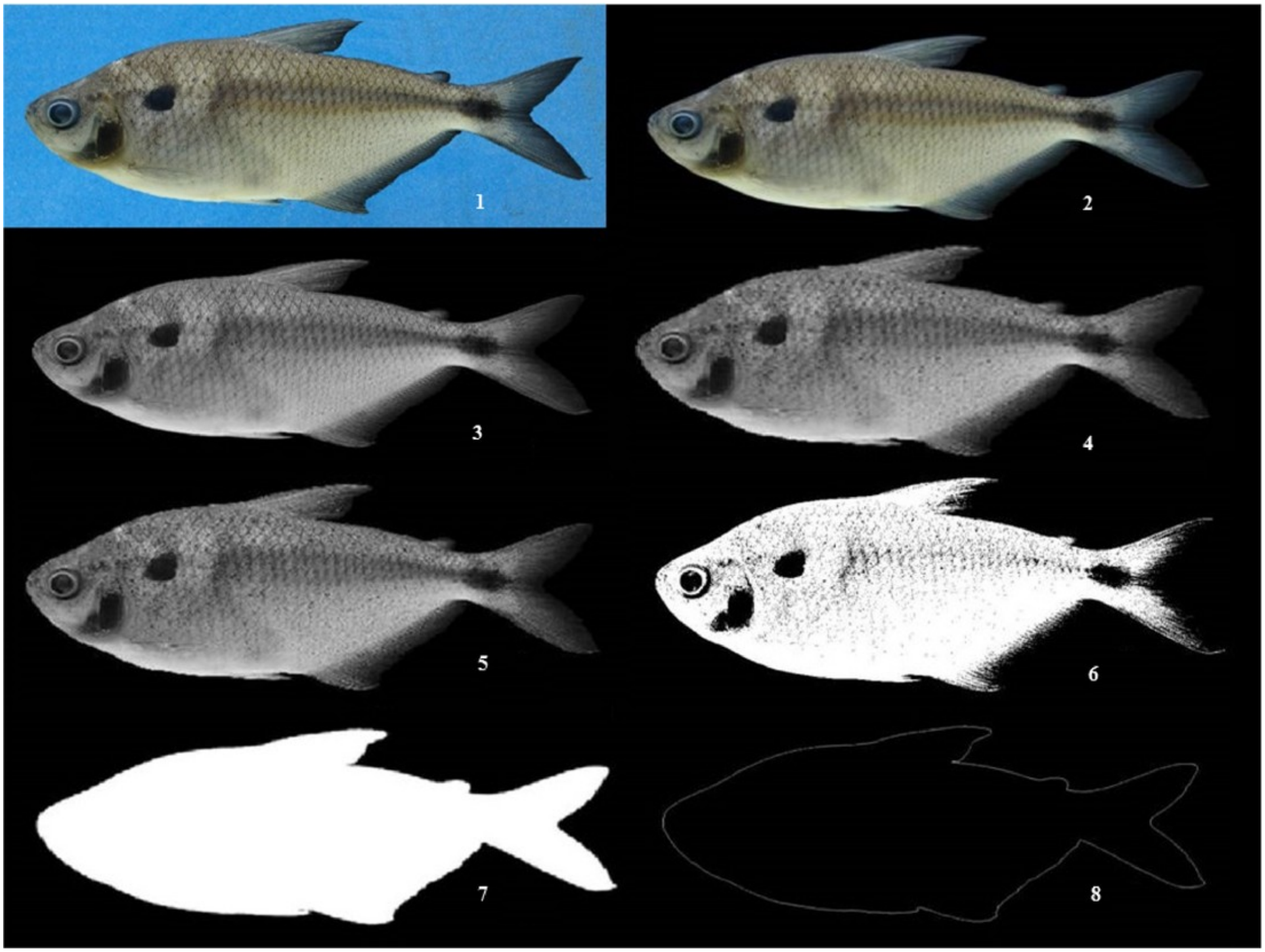


PeerJ Reviewing Manuscript

Figure 5

Figure 5

Image processing 1) jpg image, 2) Image background is removed, 3) grayscale image, 4) smoothing filter, 5) median filter, 6) binarized image, 7) contour image 8) skeletonized image.

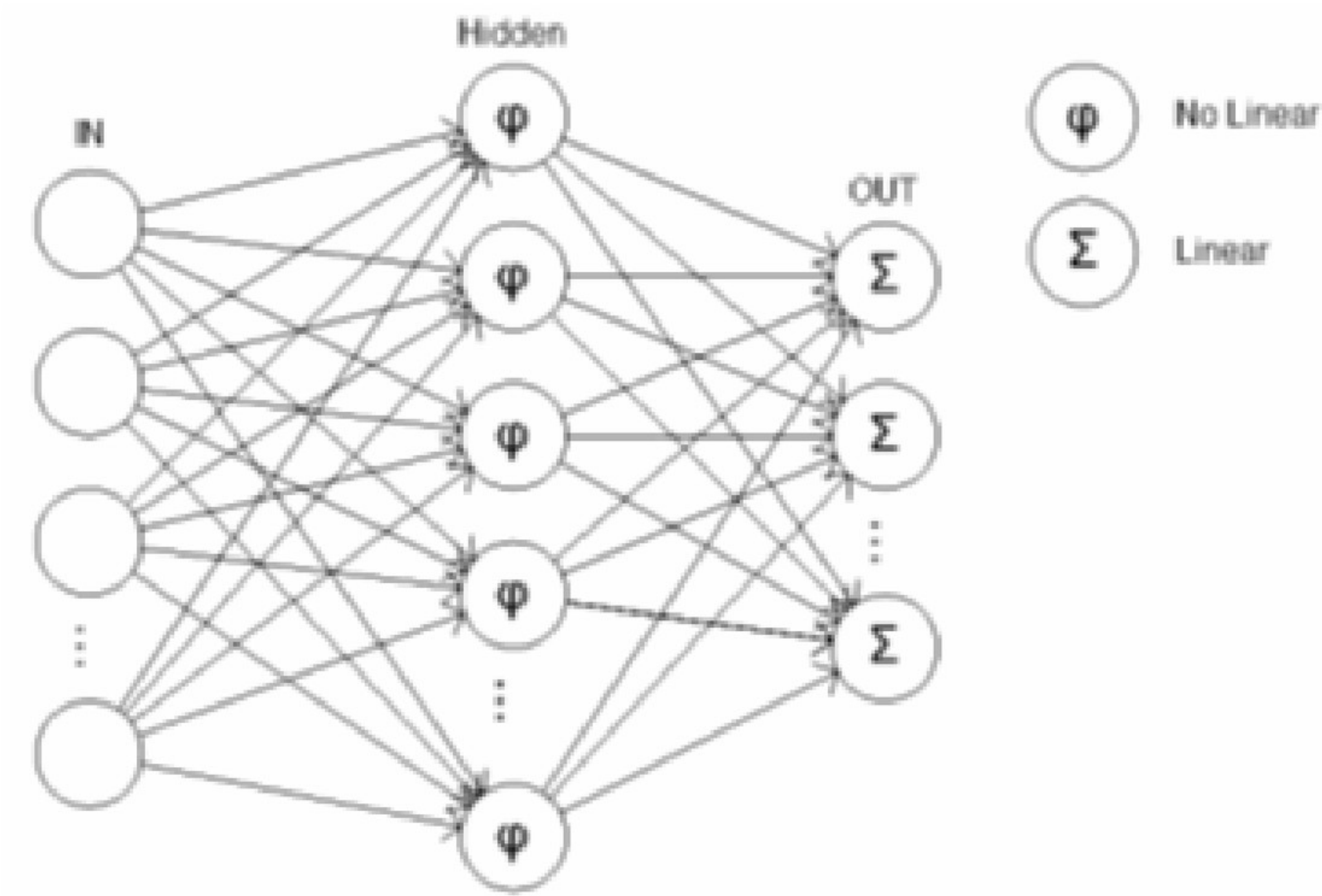


PeerJ Reviewing Manuscript

Figure 6

Figure 6

General architecture of a multilayer perceptron.

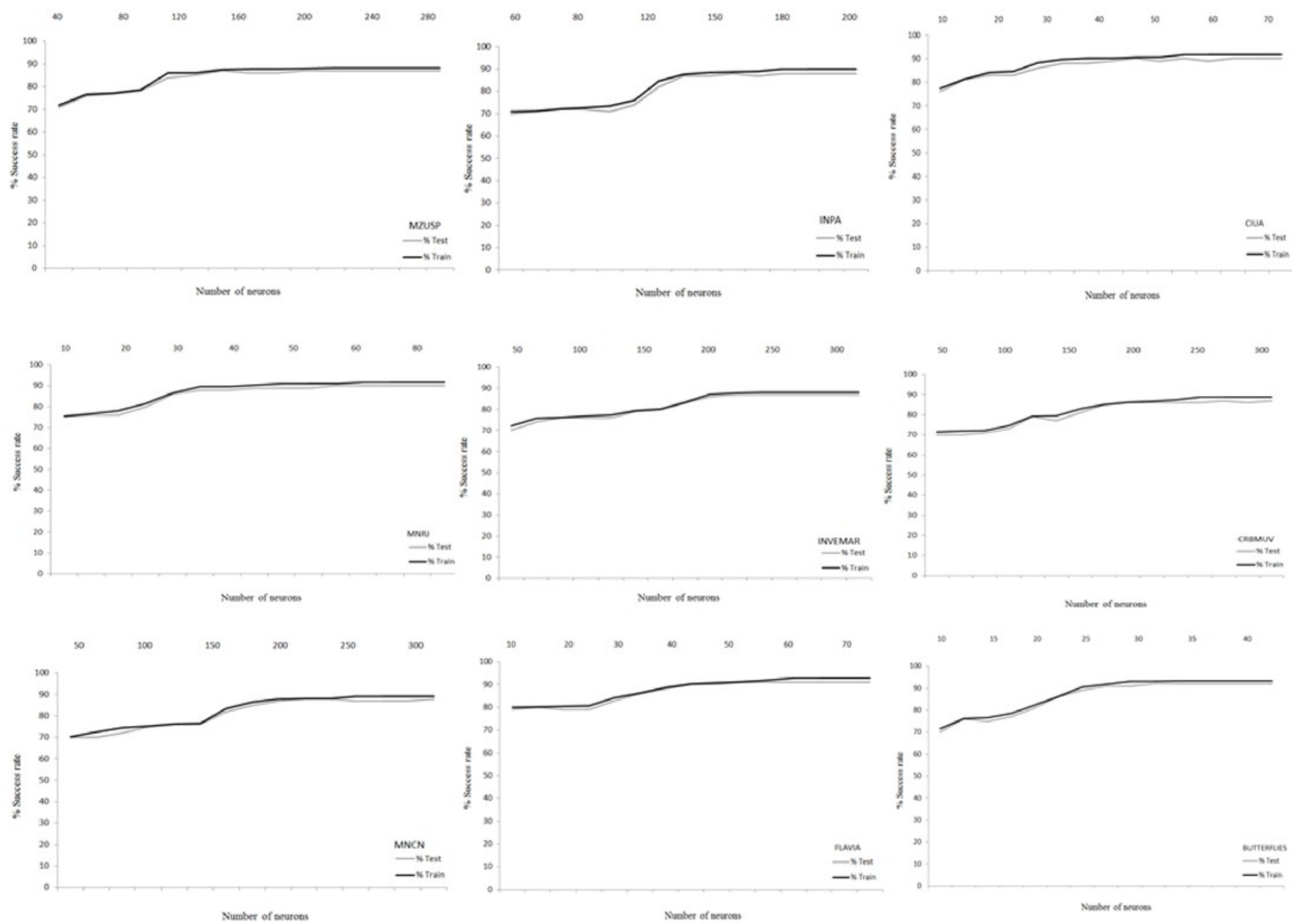


PeerJ Reviewing Manuscript

Figure 7

Figure 7

Relationship between the success rate and the number of neurons for each neural network.



PeerJ Reviewing Manuscript

Figure 8

Figure 8

An example of species confusion in the genus Astyanax 1)*Astyanax magdalenae*, 2)*Astyanax caucanus*, 3)*Astyanax fasciatus*, and4) *Astyanax microlepis.*

PeerJ Reviewing Manuscript