Neural Comput & Applic
DOI 10.1007/s00521-016-2591-2

ENGINEERING APPLICATIONS OF NEURAL NETWORKS

Detecting invasive species with a bio-inspired semi-supervised neurocomputing approach: the case of *Lagocephalus sceleratus*

Konstantinos Demertzis[1] · **Lazaros Iliadis**[1]

Received: 30 January 2016 / Accepted: 6 September 2016

Abstract The need to protect the environment and biodiversity and to safeguard public health require the development of timely and reliable methods for the identification of particularly dangerous invasive species, before they become regulators of ecosystems. These species appear to be morphologically similar, despite their strong biological differences, something that complicates their identification process. Additionally, the localization of the broader space of dispersion and the development of invasive species are considered to be of critical importance in the effort to take proper management measures. The aim of this research is to create an advanced computational intelligence system for the automatic recognition, of invasive or another unknown species. The identification is performed based on the analysis of environmental DNA by employing machine learning methods. More specifically, this research effort proposes a hybrid bio-inspired computational intelligence detection approach. It employs extreme learning machines combined with an evolving Izhikevich spiking neuron model for the automated identification of the invasive fish species "*Lagocephalus sceleratus*" extremely dangerous for human health.

Keywords eDNA · Semi-supervised learning · Semi-supervised ELM · Izhikevich neuron model · Invasive species · *Lagocephalus sceleratus*

✉ Konstantinos Demertzis
kdemertz@fmenr.duth.gr

Lazaros Iliadis
liliadis@fmenr.duth.gr

[1] Department of Forestry and Management of the Environment and Natural Resources, Democritus University of Thrace, 193 Pandazidou St., 68200 Orestiada, Greece

1 Introduction

1.1 Invasive species early detection

Invasive species, as a potential impact of climate change, pose a serious and rapidly worsening threat to natural biodiversity and ecological balance of the planet, particularly regarding marine species [1]. Although not all alien and invasive species are harmful, the precautionary principle dictates that all incomers need to be detected and that the competent bodies are obliged to be ready to respond quickly and deal with any problems that may arise. Therefore, early detection of these species is a critical process, which can slow the uncontrolled expansion of the problem, increase the likelihood of eliminating the phenomenon before it is widely established and ultimately avoid the need for costly and long-term control efforts.

The identification and classification of invasive species using exclusively phenotypic markers is an extremely difficult and uncertain process, as neither the big differences in morphology nor the significant similarities reflect the level of affinity between the organizations (species problem) [2]. The effort of species identification using genetic methods, such as DNA barcoding or by performing comparisons of biochemical or molecular markers, are the best choice for studies of intraspecific populations and subspecies. This is because high levels of polymorphism can be used to describe the genetic diversity, assessing the degree of genetic differentiation between populations [3].

The *Lagocephalus sceleratus* is common in the tropical waters of the Indian and Pacific oceans. It is a characteristic case of invasive species whose presence in the Mediterranean Sea causes serious problems. Its uncontrolled

Published online: 16 September 2016

invasion and its reproduction threatens the marine environment with an irreparable imbalance. Its presence causes an intense competition with the native fish regarding the available food. Moreover, it is extremely poisonous if eaten because it contains tetrodotoxin in its ovaries and to a lesser extent in its skin muscles and liver, which protects it from voracious predators. It becomes toxic as it eats bacteria that contain the toxin. This deadly substance causes paralysis of voluntary muscles, which may cause its victims to stop breathing or induce heart failure [4].

1.2 Environmental DNA (eDNA)

The environmental DNA (eDNA) is recovered from an environmental sample such as soil or water, rather than a single body. This technique relies on the fact that all the animals leave, in the area driven, DNA residues via feces, urine and skin. Taking samples (e.g., water) and analysis of finding eDNA, it is possible to demonstrate the presence of species without actually having this species to be caught or seen. Such samples can be analyzed by high-performance methods of DNA sequencing determination, for the rapid measurement and monitoring of biodiversity. The process of analyzing these samples called metagenomics requires specialized equipment and personnel in specialist laboratories and is quite expensive [5, 6].

1.3 Species detection by eDNA

The methodology used involves a fairly complex process in which specific primers are used in the first stage (species specific primers—SSP) [7]. Primer is a short, synthesized oligonucleotide which is used in molecular search. It is designed to recognize the precise sequence of DNA nucleotides, which is afterward used as a model for PCR and amplifies the specific part of the strand. One of the most important factors for successful DNA amplification is the proper design of primers that are species specific. The starters they interact only with the DNA of the target species sought. Then, the typically quite small amount of DNA of the target species that is detected in the eDNA (if any) is amplified by the process of polymerase chain reaction (polymerase chain reaction—PCR).

This fact formalizes the existence and identification of the target species. For this method, there is a compromise between the numbers of species that can be detected on the basis of the available primers that may be used. Also when primers are targeted at too many species (Multi specific Approach), rare species may be ignored, which imposes focused search to a particular group or species family [5, 6].

1.4 DNA-based identification

The procedure described in the paper starts by taking a random sample from the environment (eDNA), which contains material from different species, maybe thousands. The target is the identification of the genetic material of fish to then identify the genetic material of *L. sceleratus*.

To accomplish this, we use the respective sequence-specific primers (SSP) with genetic material from the groups (Algae, Cnidaria, Fishes, Mammals) which are marine species and have a similar genetic form. The aim is to use them in the training of the semi-supervised ELM model, to isolate the desired groups. The reason for using four SSP is to create a realistic and highly complex dataset. The SSP serves as reagents which are activated as soon as the corresponding DNA has been found. In this way, we isolate the genetic material of the fish of interest.

Since we complete the first stage and the DNA is grouped into four classes (algae, Cnidaria, fishes, mammals), the second phase of the proposed algorithmic approach follows. In this stage, the class "fishes" obtained from the previous process is considered as the initial dataset and thus pattern recognition is performed based on the Izhikevich spiking neuron model. This process manages to achieve the final goal which is the detection of the *L. sceleratus* DNA.

1.5 Literature review

Valentini et al. [8] tested if an eDNA metabarcoding approach, using water samples, can be used for addressing significant questions in ecology and conservation. Two key aquatic vertebrate groups were targeted: amphibians and bony fish. The reliability of this method was cautiously validated in silico, in vitro, and in situ. When compared with traditional surveys or historical data, eDNA metabarcoding showed a much better detection probability overall. For amphibians, the detection probability with eDNA metabarcoding was 0.97 (CI 0.90–0.99) versus 0.58 (CI 0.50–0.63) for traditional surveys. For fish, in 89 % of the studied sites, the number of taxa detected using the eDNA metabarcoding approach was higher or identical to the number detected using traditional methods.

Research by Herder et al. [9] has shown that in this method it is possible to detect species without actually seeing or catching them. The method uses DNA-based identification, to detect species from extracellular DNA, or cell debris, that species leave behind in the environment. Dejean et al. [10] compare the sensitivity of traditional field methods, based on auditory and visual encounter surveys, with an eDNA survey for the detection of the American bullfrog *Rana catesbeiana* = *Lithobates catesbeianus*, which is invasive in south-western France. They

demonstrate that the eDNA method is valuable for species detection and surpasses traditional amphibian survey methods in terms of sensitivity and sampling effort. The bullfrog was detected in 38 sites using the molecular method, compared with seven sites using the diurnal and nocturnal surveys, suggesting that traditional field surveys have strongly underestimated the distribution of the American bullfrog. Dejean et al. [11] estimated the time of DNA detection taking into account aquatic environment conditions and DNA concentrations. Experimentation was performed on two different species: the American bullfrog (*Rana catesbeiana* = *Lithobates catesbeianus*) and the Siberian sturgeon (*Acipenser baerii*).

On the other hand, in [12], Pan Yi discusses the use of machine learning methods with various advanced encoding schemes and classifiers to improve the accuracy of protein structure prediction. Also, in [13] a machine learning method is proposed for classifying DNA-binding proteins from non-binding proteins based on sequence information. Finally, paper [14] introduces three ensemble machine learning methods for analysis of biological DNA binding by transcription factors (TFs). The goal is to identify both TF target genes and their binding motifs. Subspace-valued weak learners (formed from an ensemble of different motif finding algorithms) combine candidate motifs as probability weight matrices (PWM), which are then translated into subspaces of a DNA k-mer (string) feature space. Assessing and then integrating highly informative subspaces by machine methods gives more reliable target classification and motif prediction.

2 Innovation of this research

The most important innovation proposed by this research is the use of machine learning methods to analyze and detect an invasive species through eDNA analysis. Although there are several related analytical studies that make use of the eDNA [6, 8–11] (to the best of our knowledge), it is the first attempt in the literature that employs a spiking neural networks machine learning approach.

Also, an important innovation is the proposal of incorporation of artificial intelligence, in digital machines that can identify invasive or rare species based on their genetic material, easily quickly and at minimal cost [3]. This will greatly enhance the planning and development of innovative biosecurity programs for the European Union [15] and other countries [16]. Also, by adding machine learning algorithms in DNA identification systems, the process is simplified, and the time required to export the results of identification is reduced and minimized for the reason that a usual system can manage one sample at a time and generate the profile within 90 min [17]. Another innovative aspect of this research is related to the collection and selection of the data, which emerged after extensive comparisons between the primers based on the FASTA algorithm [18]. These data vectors were the training samples in the learning process. Finally, the innovation is enhanced further by the development and use of a hybrid machine learning model (HMLM). The method proposed herein combines the semi-supervised classification (SSC) ELM algorithm with a sophisticated classification approach that employs the Izhikevich neuron model, whose performance is optimized with the differential evolution algorithm (DEA). The HMLM combines for the first time two very fast and highly accurate algorithms of biologically inspired machine learning, to solve a multidimensional and complex genetic identification problem.

3 Methodologies

3.1 Semi-supervised learning

The main drawback of classical learning methods with full supervision is that they need a large number of labeled training examples to construct a model with acceptable accuracy. The training is usually done manually by the instructor, which is a tedious and time-consuming process. A key feature of learning with partial supervision (PSL) is the use of pre-classified and at the same time unsorted cases (in the training process) to produce the final model. PSL uses first time seen examples, selected from the allocation followed in the real world, to enhance the efficiency of the learning process, using as few manually preclassified data vectors as possible. Self-training, mixture models, graph-based methods, co-training and multiview learning are characteristic examples of PSL [19]. It should be emphasized that the success of learning with partial supervision depends on some basic assumptions imposed by each model or algorithm.

3.2 Semi-supervised ELM classification

The ELMs are characterized by the possibility to establish the parameters of hidden nodes randomly before they see the training data vectors; they are extremely fast and efficient and can handle a multitude of trigger functions without problems such as stopping criterion, learning rate and learning epochs [20]. The semi-supervised classification ELM approach works provided that the input patterns with and without data tags come from the same marginal distribution or follow a common classes structure. The unclassified data vectors provide useful information to explore the data structure of the overall dataset, whereas the sorted data contribute to the success of the learning process.

Consider a supervised learning problem where we have a training set with N samples, $\{X, Y\} = \{x_i, y_i\}_{i=1}^{N}$. Here, $x_i \in R^{n_i}$, y_i is an n_o-dimensional binary vector with only one entry (corresponding to the class that x_i belongs to) equal to one for multi-classification tasks, where n_i and n_o are the dimensions of the input and output, respectively. Semi-supervised ELMs aim to learn a decision rule or an approximation function based on the training data. The semi-supervised ELM architecture follows two stages [21]: (1) random feature mapping; (2) output weights solving. The first stage is to construct the hidden layer using a fixed number of randomly generated mapping neurons, which can be any nonlinear piecewise continuous function, such as the Gaussian function given below:

$$g(x;\theta) = \exp(-b \parallel x - a \parallel), \tag{1}$$

where $\theta = \{\alpha, b\}$ are the parameters of the mapping function and $\| \|$ denotes the Euclidian norm. Generating feature mapping randomly enables semi-supervised ELMs for fast nonlinear feature learning and alleviates the problem of overfitting. Also, in this stage, a number of hidden neurons which map the data from the input space into an n_h-dimensional feature space (n_h is the number of hidden neurons) are randomly generated. We denote by $h(x_i) \in R^{n_{h \times n_0}}$ the output vector of the hidden layer with respect to x_i and $\beta \in R^{1 \times n_h}$ no the output weights that connect the hidden layer with the output layer. Then, the outputs of the network are given by [21]

$$f(x_i) = h(x_i)\beta, \quad i = 1, \ldots, N. \tag{2}$$

The second stage semi-supervised ELMs aim to solve the output weights by minimizing the sum of the squared losses of the prediction errors, which leads to the following formulation that is widely known as the ridge regression or regularized least squares problem [21]:

$$L_{\mathrm{ELM}} = \beta + \mathrm{CH}^T(Y - H\beta), \tag{3}$$

where $H = \left[h(x_i)^T, \ldots, (x_N)^T\right]^T \in R^{n_h \times n_h}$ and $h(x_i)\beta = y_i^T - e_i^T, \quad i = 1, \ldots, N.$

3.3 Izhikevich spiking neuron model

A typical spiking neuron model consists of dendrites, which simulate the input level of the network that collects signals from other neurons and transmits them to the next level, called soma. The soma is the process level at which when the input signal passes a specific threshold, an output signal is generated. The output signal is taken from the output level called the axon, which delivers the signal (short electrical pulses called action potentials or spike train) to be transferred to other neurons. A spike train is a sequence of stereo-typed events generated at regular or irregular intervals. Typically, the spikes have an amplitude of about 100 mV and a duration of 1–2 ms. Although the same elements exist in a linear perceptron, the main difference between a linear perceptron and a spiking model is the action potential generated during the stimulation time. Furthermore, the activation function used in spiking models is a differential equation that tries to model the dynamic properties of a biological neuron in terms of spikes. The form of the spike does not carry any information, and the number and the timing of spikes are important. The shortest distance between two spikes defines the absolute refractory period of the neuron that is followed by a phase of relative refractoriness where it is difficult to generate a spike.

Several spiking models have been proposed in the last years aiming to model different neurodynamic properties of neurons. Among these models, we could mention the well-known integrate-and-fire model, resonate-and-fire and Hodgkin-Huxley model. One of the simplest and versatile models is the one proposed by Izhikevich. This model has only nine dimensionless parameters, and it is described by the following equations [22]:

$$C\dot{\upsilon} = k(\upsilon - \upsilon_r)(\upsilon - \upsilon_t) - u + I, \tag{4}$$

$$\text{if} \quad \upsilon \geq \upsilon_{\mathrm{peak}} \quad \text{then} \begin{Bmatrix} \upsilon \leftarrow c \\ u \leftarrow u + d \end{Bmatrix}, \tag{5}$$

$$\dot{u} = \alpha\{b(\upsilon - \upsilon_r) - u\}. \tag{6}$$

Depending on the values of a and b, it can be integrator ($b < 0$) or resonator ($b > 0$). The parameters c and d do not affect the sub-threshold behavior (in a steady-state), whereas they affect the general model in the after-spike behavior. The parameter u is the membrane potential (membrane potential is the difference in electric potential between the interior and the exterior of a biological cell. With respect to the exterior of the cell, typical values of membrane potential range from -40 to -80 mV), u is the recovery current that represents a membrane recovery variable, which accounts for the activation of K+ ionic currents and inactivation of Na+ ionic currents, and it provides negative feedback to u. After the spike reaches its apex (+30 mV), the membrane voltage and the recovery variable are reset according to Eq. (5). C is the membrane capacitance of a neuron that influences synaptic efficacy and determines the speed with which electrical signals propagate along dendrites and axons, υ_r is the resting membrane potential in the model that is between 70 and 60 mV depending on the value of b, and υ_t is the instantaneous threshold potential which is the critical level to which the membrane potential must be depolarized to initiate an action potential. The parameter k occurs when the neuron's rheobase (*rheobase* is the minimal current

amplitude of infinite duration) and input resistance. The recovery time constant is α. The spike cutoff value is υ_{peak} and the voltage reset value is c. The parameter d describes the total amount of outward minus inward currents activated during the spike and affects the after-spike behavior [22]. Various selections of these parameters can lead to various native operating standards, depending on the objective and the problem it is required to solve.

4 Description of the proposed hybrid approach

The algorithmic process of the hybrid scheme proposed includes at first stage the use of the semi-supervised ELM classification approach to create classes, which contain the genetic material of a species family as Class 1—fish, Class 2—mammals, Class 3—algae and so on. To carry out this process, special samples with genetic material of each group (algae, Cnidaria, fish, mammals) are used in the training process of the semi-supervised ELM classification [21] as with the respective primers. These samples (primers) which are the fewest in the training set are the labeled data, which will be used for the training and they are denoted as:

$$\{X_l, Y_l\} = \{x_i, y_i\}_{i=1}^{l}. \tag{7}$$

The unlabeled data which are the biggest part of the training set are denoted as

$$\{X_u\} = \{x_i\}_{i=1}^{u}. \tag{8}$$

The l and u are the number of the labeled and unlabeled data, respectively. Then the following steps are used to calculate the mapping function of the SS-ELM: $R^{n_i} \rightarrow R^{n_o}$:

Step 1: construct the graph Laplacian L from both X_l and X_u.

Step 2: initiate an ELM network of n_{h} hidden neurons with random input weights and biases and calculate the output matrix of the hidden neuron $H \in R^{(l+u)xn_i}$.

Step 3: choose the trade-off parameter C_0 and λ.

Step 4: *if* $n_{\text{h}} \leq N$, compute the output weights β using function (9)

$$\beta = \left(I_{n_h} + H^T \text{CH} + \lambda H^T \text{LH}\right)^{-1} H^T C\tilde{Y}, \tag{9}$$

else compute the output weights β using function (10)

$$\beta = H^T \left(I_{l+u} + \text{CHH}^T + \lambda \text{LHH}^T\right)^{-1} C\tilde{Y}. \tag{10}$$

$$\text{Return the mapping function} f(x) = h(x)\beta \tag{11}$$

It should be noted that a change in the input current signal changes also the response of the Izhikevich neuron model, creating different firing rates. The firing rates are calculated as the number of spikes generated in an interval T. The neuron is excited for a time T ms when receiving an input signal and it fires when this spike or a train of spikes exceeds a particular threshold of membrane potential and then we have an action potential.

Having completed the first stage and the DNA grouped in classes, the second phase of the proposed algorithmic approach follows, in which the class fishes (obtained by the previous procedure) is taken as the initial dataset. Based on this class, the Izhikevich spiking model [22] tries to discover the DNA of the species *L. sceleratus* with the process of pattern recognition. This procedure is described as follows.

Following the hypothesis "patterns from the same class produce similar firing rates in the output of the spiking neuron and patterns from other classes produce firing rates different enough to discriminate among the classes," the Izhikevich model can be applied to solve the specified pattern recognition problem. Let $D = \{x^i, k\}_i^p = 1$ be a set of associations composed of input patterns, where $= 1,\ldots,$ is the class to which $x^i \in R^n$ belongs. The learning process adjusts the synaptic values of the model in such a way that the output generates a different firing rate for each class, reproducing the behavior described in the hypothesis. To use the Izhikevich neuron model to solve the *L. sceleratus* pattern classification problem, it is necessary to compute the input current that stimulates the model. In other words, the spiking neuron model is not directly stimulated with the input pattern $x^i \in R^n$ but with the input current . If we assume that each feature of the input pattern x^i corresponds to the presynaptic potential of different receptive fields, then we can calculate the input current that stimulates the spiking neuron as

$$I = x \cdot w, \tag{12}$$

where $w^i \in R^n$ is the set of synaptic weights of the neuron model. This input current is used in the methodology to stimulate the spiking model during ms.

Instead of using the spike train generated by the spiking model to perform the pattern classification tasks, we compute the firing rate of the neuron defined as

$$fr = \frac{N_{\text{sp}}}{T}, \tag{13}$$

where $_{\text{sp}}$ is the number of spikes that occur within the time window of length.

It is necessary to calculate the average firing rate AFR $\in R^K$ of each class, using the firing rates produced by each input pattern. In this sense, the learning process consists of finding the synaptic values of the spiking model in such a

way that it generates a different average firing rate for each class.

Suppose that the spiking neuron is already trained using a learning strategy. To determine the class to which an unknown input pattern x belongs, it is necessary to compute the firing rate generated by the trained spiking neuron. After that, the firing rate is compared against the average firing rate of each class. The minimum difference between the firing rate and the average firing rates determines the class of an unknown pattern. This is expressed with the following equation:

$$\mathrm{cl} = \mathrm{argmin}_{k=1}^{K}(|\mathrm{AFR}_k - fr|), \tag{14}$$

where fr is the firing rate generated by the neuron model stimulated with the input pattern $\tilde{x}$ [22].

To achieve the desired behavior at the output of the spiking neuron, it is necessary to adjust its synaptic weights. During the training phase, the synapses of the neuron model w are calculated using a powerful and efficient technique for optimizing non-linear and non-differentiable continuous space functions, which are called DEA [23]. This heuristic algorithm optimizes a problem by maintaining a population of candidate solutions and creating new candidate solutions by combining existing ones according to its simple formulae, and then keeping whichever candidate solution has the minimum score or error function on the optimization problem at hand. This approach has a lower tendency to converge to the local maxima; it evolves populations with a smaller number of individuals and has lower computation cost. To maximize the accuracy of the spiking neuron model during a pattern recognition task, the best set of synaptic weights must be found using this algorithm. The function that uses the classification error to find the set of synaptic weights is defined as follows:

$$f(w, D) = 1 - \mathrm{Performance}(w, D), \tag{15}$$

where w are the synapses of the model, D is the set of input patterns and performance (w, D) is a function which computes the classification accuracy in terms of (14), given by

$$\mathrm{Performance}(w, D) = \frac{P_{\mathrm{cc}}}{P_t}, \tag{16}$$

where $_{\mathrm{cc}}$ denotes the number of patterns correctly classified and $_{\mathrm{t}}$ denotes the number of tested patterns.

The general training methodology used to train the Izhikevich spiking model with DEA begins with the creation of a plurality of random populations of candidate solutions in the form of numerical vectors. The first of them are chosen as targets. Then the DEA creates a trial vector to perform the following four steps [22, 23]:

Step 1. Randomly select two vectors from the current generation.

Step 2. Use the selected to compute the difference vector.

Step 3. Multiply the difference vector by the weighting factor.

Step 4. Form the new trial vector by adding the weighted difference vector to a third one, randomly selected from the current population.

The trial vector replaces the target one in the next generation, if and only if the first produces a better solution than the current, after comparing the cost value obtained by the fitness function. The overall algorithmic approach that was proposed herein is described clearly and in detail in Fig. 1.

5 Training and testing datasets

Two datasets were created for the training of the hybrid system and the implementation of testing. These datasets resulted from the conversion of the initial genetic information recorded with the code of four letters A, T, C and G (abbreviations of the bases adenine, thymine, cytosine and guanine) in the DNA of the tested species, in numerical form.

The first eDNA sample dataset used in the training process of the semi-supervised ELM classification algorithm consists of three smaller subsets as below:

The Train Set includes 856 instances distributed to 580 independent variables and four classes (algae, Cnidaria, fishes and mammals). Each class came from the DNA of the respective species, namely: class algae from the DNA of 66 respective species, the class Cnidaria from the DNA of 58, the class fishes from the DNA of 81 and the class mammals from the DNA of 43 respective species [24].

This dataset is used as the primers (SSP), that is, specific samples with genetic material of each class (Algae, Cnidaria, Fishes and Mammals) which are used to train the SSC ELM. It is in fact the labeled data.

The Unlabeled Set is spread over 580 independent variables and includes 4382 instances which have no class (unlabeled). They came from the DNA of respective species like the Train Set. These are unlabeled data that provide useful information on the algorithm for the exploration of the data structure of the general test set.

The Test Set includes 184 instances distributed to 580 independent variables and 4 classes, which came from the DNA of similar species, like the Train Set. These are labeled data that are used by the algorithm to test its accuracy (after training).

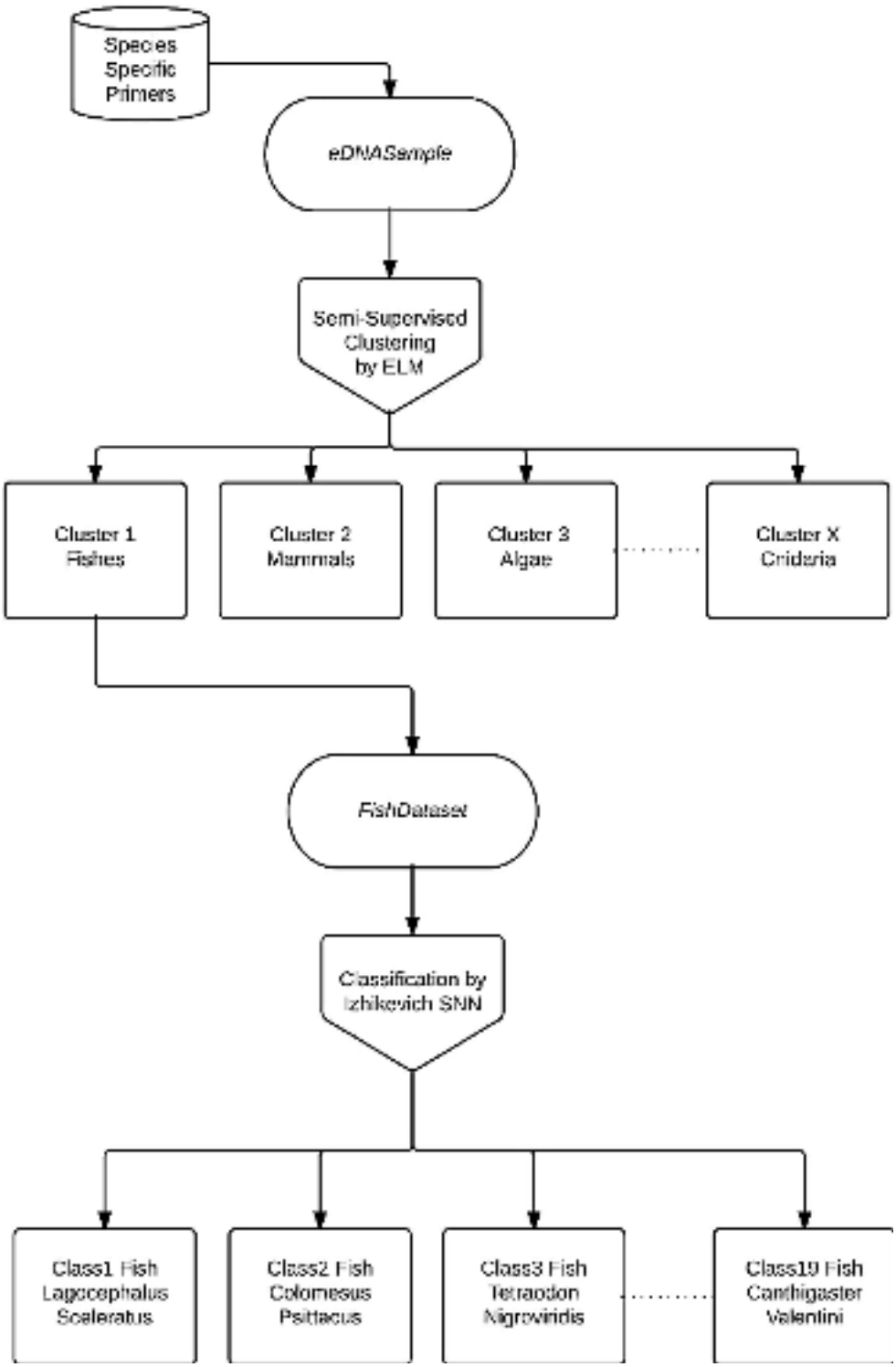

Fig. 1 Architecture of the proposed model

The second dataset in which the Izhikevich neuron model performs pattern recognition is the *Fish_Dataset*, which in essence is the class of fishes resulting from the process of the semi-supervised ELM classification. This is a highly sophisticated set of data, which resulted from the DNA of 81 fish species exhibiting high genetic similarities [24, 25]. Extensive comparisons were performed on the protein and DNA sequences (protein similarity search—PSS method) between the DNA of the fish species *L. sceleratus* and similar species. The PSS method provides sequence similarity searching against protein databases using the FASTA algorithm [25]. FASTA takes a given

nucleotide or amino acid sequence and searches a corresponding sequence database using local sequence alignment to find matches of similar database sequences. This algorithm follows a largely heuristic method which contributes to the high speed of its execution. It initially observes the pattern of word hits, word-to-word matches of a given length and marks potential matches before performing a more time-consuming optimized search. The FASTA is not a machine learning algorithm, but the most reliable method to select species with high genetic similarity and to create a particularly complex dataset.

The Fish_Dataset includes 1823 instances, distributed in 580 independent variables and 19 classes that represent fish of very high genetic similarity with the *L. sceleratus*.

6 Results and comparative analysis

Given that the datasets created a high genetic similarity among the species tested and the specificities resulting from the semi-supervised learning process, it is extremely impressive that the proposed hybrid system managed to solve a particularly complex, realistic genetic problem with high accuracy. It is characteristic that in the process of semi-supervised ELM classification, genetic characteristics of the species tested were placed correctly on the classes that symbolize the overall family of these species, accounting for about 91.3 %. To appreciate the actual importance of the above percentage, we have to consider that the said algorithm was trained with about 19 % of the total data (labeled data). The analytical values of the predictive power of semi-supervised ELM classification algorithm are presented in detail in Table 1, and the confusion matrix that shows the number of correct and incorrect predictions made by the classification model compared to the actual outcomes (target value) in the data is shown in Table 2.

The Izhikevich neuron model was tested on a very high complexity dataset, containing data derived from fish of the highest similarity index, which was determined after the comparisons made on protein and DNA sequences using the PSS method and the FASTA algorithm. The *Fish_Dataset* has a much higher degree of complexity and difficulty from the corresponding *lago_fasta_dataset* that includes the DNA barcode of 772 fish, corresponding to fish of high genetic similarity with *L. sceleratus*, which was proposed in our previous research effort [3]. The Izhikevich neuron model had a performance equal to 96.2 % with tenfold cross-validation. This percentage is higher than the result obtained by the ELM after resampling with bootstrap, applying the replacement method.

Due to the fact that the number of the used features was too high in the lago_fasta_dataset (558), several feature selection attempts were done for the reduction of the training time and for the enhancement of the generalization to avoid overfitting. The particle swarm optimization (PSO) was used in [3] to search for the optimal feature subset. The assessment of each subset was done by considering the value of each subset, which is based on the contribution and the degree of redundancy of each characteristic. The parameters considered for the final decision are related to the classification accuracy and to the correlation of the classification errors in comparison to the accuracy of the initial parameters set. After the feature selection, finally lago_fasta_dataset has 235 features (reduction by 57 %) and the accuracy of the ELM was 96.3 % [3] that is almost the same as the Izhikevich neuron model with a complete dataset with 558 features.

The corresponding comparative results are presented in Table 3.

This comparison generates very encouraging expectations for the wider use and exploitation of the hybrid developed model as a robust classification model for such a complex real-time problem. The proportion of the total number of predictions that were correct and the very high sensitivity rates, which represent the true identification cases of invasive species (true positive rate) are typical and indicative of the quality of the process. This is also shown by the size of the ROC curves that is a factor that played an important role toward the generalization capacity of the proposed system. Finally, lower values of RMSE indicate better prediction and how accurately the model predicts the response.

Table 1 Performance matrices of semi-supervised ELM classification method

Accuracy	91.3 %
Correctly classified instances	168
Incorrectly classified instances	16
Root mean-squared error (RMSE)	0.169
Mean absolute error (MAE)	0.0509
Avg. precision	0.909
Avg. recall	0.910
Avg. F-measure	0.907
Avg. ROC area	0.988

Table 2 Confusion matrix of semi-supervised ELM classification method

Algae	Cnidaria	Fishes	Mammals	
29	1	0	0	Algae
0	58	0	0	Cnidaria
2	09	32	0	Fishes
0	4	0	49	Mammals

Table 3 Accuracy (ACC) and performance matrices (PM) comparison between Izhikevich neuron model, ELM after resampling by bootstrap and ELM after feature selection by PSO

Izhikevich neuron model (*Fish_Dataset)*			ELM after resampling by bootstrap method (*lago_fasta_dataset*)			ELM after feature selection by the PSO search method (*lago_fasta_dataset*)		
ACC and PM			ACC and PM			ACC and PM		
ACC	RMSE	AROC	ACC	RMSE	AROC	ACC	RMSE	AROC
96.2 %	0.0693	0.995	96.0 %	0.0731	0.992	96.3 %	0.0684	0.995

7 Conclusions and further work

The advanced hybrid application of computational intelligence, described in conjunction with the extremely promising results obtained, offers a reliable innovative proposal in the formulation and design of biosecurity methods and protection of biodiversity. The simplification of the detection and identification of invasive species by the method of eDNA, allows the collection of data and therefore the recording of non-indigenous species that exist in some areas. It also creates the conditions for studying the behavior of different species and the seasonal fluctuation of their populations. Finally, it helps in accurate mapping of general intrusion and can contribute significantly to slowing the uncontrolled expansion of the problem of invasive species, avoiding the need for costly and long-term monitoring efforts. Clearly, the broad application of the proposed method which simplifies and reduces to a minimum the cost and the time of the genetic identification and the wide collection of these data is a prerequisite for the development of a risk management and prevention system, designed to protect the environment and public health.

The hybrid biologically inspired method proposed herein was tested successfully in controlling and automatic recognition of the invasive fish species *L. sceleratus* by digital machines.

It is important to note that the proposed method using eDNA is probably one of the best options for studies in intraspecific population levels and subspecies, especially for the characterization of hybrids which exhibit high levels of polymorphism through their environment, without necessarily being identified. Also, with this method, species can be identified by studying only their residues in any stage of their life cycle. Furthermore, the method may be used to describe the genetic diversity in populations, while considering the degree of genetic differentiation between them.

Another very important advantage presented by this classification method is that it can distinguish species which are very similar to each other using the SSP primers. In this way, it can reduce the uncertainty and the doubt among the classifications and it can also identify rare or even "extinct" species.

One of the future research directions that could be conducive to the proposed system is related to the choice of appropriate characteristics (feature selection) using an optimization method such as PSO, for calculating the value of the independent variables with the highest individual predictive ability. Also, another step forward would be the implementation of the proposed hybrid system combining this time three different learning methods (semi-supervised, unsupervised and reinforcement learning), to identify and exploit hidden knowledge among heterogeneous data combined by the analysis of eDNA. Also, an important innovation could be the use of corresponding advanced artificial intelligence technology such as Deep Learning, to solve the same problem. Finally, the application of the eDNA analysis under a different scale (big data) with bioinformatics approaches might be a challenge (metagenomic analysis).

References

1. Rahel F, Olden JD (2008) Assessing the effects of climate change on aquatic invasive species. Soc Conserv Biol 22(3):521–533
2. Miller W (2001) The structure of species, outcomes of speciation and the species problem: ideas for paleobiology. Palaeoclimatol Palaeoecol 176:1–10
3. Demertzis K, Iliadis L (2015) Intelligent bio-inspired detection of food borne pathogen by DNA barcodes: the case of invasive fish species *Lagocephalus sceleratus*. Eng Appl Neural Netw 517:89–99. doi:10.1007/978-3-319-23983-5_9
4. Kheifets J, Rozhavsky B, Solomonovich ZG, Rodman M, Sorooksky A (2012) Severe tetrodotoxin poisoning after consumption of *Lagocephalus sceleratus* (Pufferfish, Fugu) fished in Mediterranean sea, treated with cholinesterase inhibitor. Case Rep Crit Care. doi:10.1155/2012/782507
5. http://www.environmental-dna.nl/
6. Ficetola G, Miaud C, Pompanon F, Taberlet B (2008) Species detection using environmental DNA from water samples. Biol Lett 4:423–425
7. Yu-Li S, Naoki K, Cheng-Xu L, Yoshiko M, Haru K, Kunitomo W (2000) Rapid identification of 11 human intestinal *Lactobacillus* species by multiplex PCR assays using group- and species-specific primers derived from the 16S–23S rRNA intergenic spacer region and its flanking 23S rRNA. FEMS Microbiol Lett. doi:10.1111/j.1574-6968.2000.tb09155.x
8. Valentini A, Taberlet P, Miaud C, Civade R, Herder J, Thomsen P, Bellemain E, Besnard A, Coissac E, Boyer F, Gaboriaud C,

Jean P, Poulet N, Roset N, Copp H, Geniez P, Pont D, Argillier C, Baudoin M, Peroux T, Crivell J, Olivier A, Acqueberge M, Brun M, Møller R, Willerslev E, Dejean T (2015) Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. Mol Ecol. doi:10.1111/mec.13428

9. Herder E, Valentini A, Bellemain E, Dejean T, van Delft J, Thomsen P, Taberlet P (2014) Environmental DNA: a review of the possible applications for the detection of (invasive) species. Stichting RAVON, Nijmegen. Rapport 2013-104
10. Dejean T, Valentini A, Miquel C, Taberlet P, Bellemain E, Miaud C (2012) Improved detection of an alien invasive species through environmental DNA barcoding: the example of the American bullfrog *Lithobates catesbeianus*. Appl Ecol 49(4):953–959
11. Dejean T, Valentini A, Duparc A, Pellier-Cuit S, Pompanon F, Taberlet P, Miaud C (2011) Persistence of environmental DNA in freshwater ecosystems. PLoS One 6(8):e23398
12. Pan Y (2005) Protein structure prediction and understanding using machine learning methods. IEEE Granul Comput. doi:10.1109/GRC.2005.1547225
13. Ma X, Hu L (2013) Extracting sequence features to predict DNA-binding proteins using support vector machine. Comput Inf Sci. doi:10.1109/ICCIS.2013.48
14. Dong-Jun Y, Hu J, Li QM, Tang ZM, Yang JY, Shen HB (2015) Constructing query-driven dynamic machine learning model with application to protein-ligand binding sites prediction. IEEE Trans NanoBiosci 14(1):45–58. doi:10.1109/TNB.2015.2394328
15. Council Directive 2000/29/EC of 8 May 2000, Official J L 169, pp 0001–0112
16. New Zealand Biosecurity Surveillance Strategy 2020. http://www.hortnz.co.nz/assets/Uploads/SurveillanceStrategysubmissionNov08.pdf
17. Rapid DNA platform. http://integenx.com/wp-content/uploads/2016/02/RapidHIT-ID-Brochure-Desktop-DNA-is-here.pdf
18. FASTA Sequence Comparison. http://fasta.bioch.virginia.edu/fasta_www2/fasta_list2.html
19. Zhu X, Goldberg A (2009) Introduction to semi-supervised learning. Synth Lect Artif Intell Mach Learn 3(1):1–130. doi:10.2200/S00196ED1V01Y200906AIM006
20. Cambria E, Huang G-B (2013) Extreme learning machines. IEEE Intell Syst 28(6):30–31
21. Huang G, Song S, Gupta JN, Wu C (2014) Semi-supervised and unsupervised extreme learning machines. IEEE Trans Cybern. doi:10.1109/TCYB.2014.2307349
22. Vazquez R (2010) Izhikevich neuron model and its application in pattern recognition. Aust J Intell Inf Process Syst 11(1):35–40
23. Price K, Storn M, Lampinen A (2005) Differential evolution: a practical approach to global optimization. Springer. ISBN: 978-3-540-20950-8
24. Invasive Species Compendium. http://www.cabi.org/isc/
25. Protein Similarity Search. http://www.ebi.ac.uk/Tools/sss/fasta/