Math Geol (2007) 39: 657–671
DOI 10.1007/s11004-007-9120-x

Two Supervised Neural Networks for Classification of Sedimentary Organic Matter Images from Palynological Preparations

Andrew F. Weller · Anthony J. Harris · J. Andrew Ware

Received: 7 November 2005 / Accepted: 4 April 2007 / Published online: 5 October 2007

Abstract An improvement in the supervised artificial neural network classification of sedimentary organic matter images from palynological preparations is presented. Sedimentary organic matter encompasses the entire acid-resistant organic microparticles (typically with a diameter of 5–500 μm) recovered from a sediment or sedimentary rock. Supervised neural networks are trained to recognize patterns within databases for which the correct classifications are already known. Once trained, they are verified on pre-classified samples not seen by the network, and then used for classification of samples whose class is not known. Such networks have an input, hidden and output layer. Typically, these networks determine what the output class is by adjusting weights associated with the layer interconnects, and by modifying the signals that propagate through the hidden layer by a non-linear transfer function. In this example, the inputs in each network are the salient features selected from an available set of 194, while the outputs are the sedimentary organic matter classifications which were formerly developed with the rationalization of descriptive terms from previous classification schemes. The author's past work tested the supervised back propagation neural network for the classification of sedimentary organic matter images. This gave an overall correct classification rate of 87%. However, because the back propagation network underperformed on two of the four classes, the radial basis function neural network was tested on the same databases initially used in an attempt

A.F. Weller (✉)
Geological Institute, Department of Earth Sciences, ETH Zurich, 8092, Zurich, Switzerland
e-mail: weller@erdw.ethz.ch

A.J. Harris
School of Applied Sciences, University of Glamorgan, Pontypridd, CF37 1DL, UK
e-mail: ajharri1@glam.ac.uk

J.A. Ware
School of Computing, University of Glamorgan, Pontypridd, CF37 1DL, UK
e-mail: jaware@glam.ac.uk

to improve the recognition rate of these two classes. The difference between the back propagation and radial basis function networks lies in the non-linear transfer function applied in the hidden layer, which was modified by a Gaussian function in the latter. In the best-case scenario, this improved the recognition rate by 4% to just over 91%. This has also determined that a series of different supervised neural networks may be better for classification of sedimentary organic matter images. These results are encouraging enough to prompt further research that may result in a commercially viable system.

Keywords Back propagation · Radial basis function · Image analysis · Palynofacies

Introduction

The semi-automated capture, analysis and classification of sedimentary organic matter (OM) in palynological preparations were formerly described for palynofacies studies (Weller et al. 2005). Such studies encompass the entire acid-resistant organic micro-particles (typically with a diameter of 5–500 μm) recovered from a sediment or sedimentary rock (Fig. 1). Once these palynological residues have been extracted through a series of chemical digestion techniques and mounted on microscope slides for analysis and counting, the particles can be identified and classified for use in geochronological, biostratigraphical, paleoecological and/or paleoenvironmental analysis (Traverse 1988). Traditionally, this material is manually analyzed with a microscope in a time consuming manner. By automating this 'routine identification' component, more emphasis can be placed on distinguishing rarer ('unknown') particles and placing them in a descriptive context for assessment of geological change.

The classification scheme adapted in this work uses a rationalization of morphologically and texturally descriptive terms from three previous classification schemes (Boulter 1994; Tyson 1995; Batten 1996) with the removal of redundant descriptors (Table 1). Previously, a series of multi-layer back propagation (BPN) supervised artificial neural networks (ANNs) were used to classify the 1st order class and subsequently the 2nd order classes; this demonstrated an average correct recognition rate of 87% (Weller et al. 2005). The test data comprised 3266 manually assigned 1st order particles of sedimentary OM, which were split into four subsets: 1st order, 2nd order amorphous (501 particles), 2nd order palynodebris (1475 particles) and 2nd order palynomorphs (1290 particles), which corresponded to the proposed classification scheme (Table 1). By using BPN ANNs, the 1st order and 2nd order amorphous classes were found to have a relatively low classification rate (66–85%) compared to the 2nd order palynodebris and the 2nd order palynomorphs classes (>92%). To improve this recognition rate the radial basis function (RBF) supervised ANNs have been trained and tested on these original databases; the results of which are presented here. These studies indicate that different ANN paradigms can be used in a series for ultimate classification. This will be useful for characterization tasks that involve vast quantities of multivariate data and potentially use a series/hierarchy for classification.

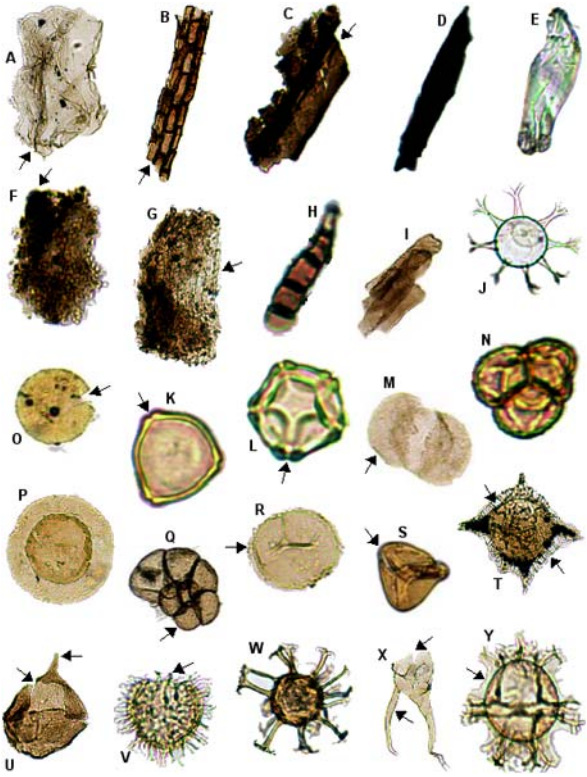

Fig. 1 Example sedimentary OM images (×200 magnification, not to scale); classified according to Table 1. Note dark particles on light background. **A** Membranous material; *arrow* indicates fibrous structural element. **B** and **C** Brown wood; *arrow* on (**B**) indicates (structural) cross-hatch pattern; *arrow* on (**C**) indicates fibrous structural element. **D** Opaque material with sharp, distinct outline. **E** Filament. **F** Amorphous OM of aquatic origin with a spongy, granular appearance and indistinct outline; *arrow* indicates possible pyritization. **G** Amorphous OM of terrestrial origin; *arrow* indicates remnants of structural elements. **H** Fungal material (elongated hyphae). **I** Compact colonial species of chlorophyte algae; note elongated cells with smooth surface textures. **J** Acritarch; note branching projections. **K** and **L** Angiosperm pollen grains; *arrow* on (**K**) indicates one of three pores present (triporate); *arrow* on (**L**) indicates one of five pores present (periporate). **M** Saccate gymnosperm pollen grain; *arrow* indicates one of two air sacs present (bisaccate). **N** Tetrad pollen grain. **O** and **P** Prasinophyte algae; *arrow* on (**O**) indicates suture through which the cell contents are released; note double wall layer in (**P**). **Q** Foraminiferal linings; *arrow* indicates single chamber. **R** and **S** Spores (note bilateral symmetry); (**R**) is a spheroidal spore, with *arrow* indicating a rectilinear germinal aperture (monolete spore); (**S**) is a tetrahedral spore, with *arrow* indicating a three branching germinal aperture (trilete spore). **T** Proximochorate dinocyst (i.e. projections between 10 and 30% of the diameter); *arrows* indicate different wall layers. **U** Proximate dinocyst (i.e. projections $<10\%$ of the diameter); *top arrow* indicates apical horn, *bottom arrow* indicates archeopyle (suture-breakage) where excystment has occurred. **V** Proximochorate dinocyst; *arrow* indicates archeopyle. **W** Chorate dinocyst (i.e. projections $>30\%$ of the diameter); note tip termination of projections which differs between dinocysts with projections. **X** Proximate dinocyst; *top arrow* indicates archeopyle, *bottom arrow* indicates one of pair of antapical horns. **Y** Chorate dinocyst; *arrow* indicates 'intergonal ridges' (i.e. features relating to the parasuture). For further examples and greater explanation, refer to Weller (2004)

Table 1 Adapted palynofacies classification scheme, including hierarchic descriptives (1st and 2nd order) with simple morphological and textural description

1st order descriptive	2nd order descriptive	Morphological and textural description
Amorphous (1)	Terrestrial origin (1)	Pale to brown; no opening in surface; diffusive outline (not sharp); irregular (no straight, curved or corners) edges; no coiling or projections; some internal (degraded) structure (not high proportion of lineations); speckled appearance; & not homogeneous
	Aquatic origin (2)	Pale to brown; no opening in surface; diffusive outline (not sharp); irregular (no straight, curved or corners) edges; no approx. parallel sides, coiling or projections; no internal structure; speckled appearance; & not homogeneous
Palynomorphs (2)	Spores and pollen (3)	Pale to brown; sharp distinct outline (not diffusive); some edge curvature; no approx. parallel sides, coiling; maybe some projections; some internal structure; no (black) inclusions; & not homogeneous
	Dinoflagellate cysts (dinocysts), acritarchs and other algae(4)	Pale to brown; sharp distinct outline (not diffusive); maybe projections; some internal structure; no (black) inclusions; maybe speckled texture; & not homogeneous
	Foraminiferal linings (5)	Pale to brown; no opening in surface; sharp distinct outline (not diffusive); some edge curvature (not straight or corners); no projections; maybe some (internal structure) lineations (not random, radiate from point or high proportion); no (black) inclusions; not speckled texture; & not homogeneous
	Fungal material (6)	Brown; not sheet-like; sharp distinct outline (not diffusive); either curved or straight edge (no corners or irregular); no coiling; maybe multiple lineations (not random); no (black) inclusions; not speckled texture; & not homogeneous
	Prasinophyte algae (7)	Pale to brown (not black); sharp distinct outline (not diffusive); curved edge; no corners or irregular edge; no elongation, approx. parallel sides, coiling or projections; some internal structure (no approx. parallel lineations); no (black) inclusions; & maybe speckled & homogeneous
Palynodebris (3)	Opaques (8)	Black; sharp distinct outline (not diffusive); no coiling or projections; no internal structure; maybe parallel pitting; & homogeneous
	Membranous material (9)	Pale; sheet-like; sharp distinct outline (not diffusive); no coiling or projections; maybe some internal structure; no speckled texture; & maybe homogeneous
	Brown wood (10)	Brown; maybe sheet-like; sharp distinct outline (not diffusive); no coiling or projections; maybe some internal structure; no speckled texture; & maybe homogeneous
	Tubes, filaments and hairs (11)	Pale to brown; no opening in surface; not sheet-like; sharp distinct outline (not diffusive); no irregular edge; elongated with approx. parallel sides; no coiling or projections; some internal structure; no random, radiating or high proportion of lineations; no (black) inclusions; no speckled texture; & maybe homogeneous

Data Preprocessing

Images of sedimentary OM are captured through a (transmitted-light) microscope-mounted camera at ×200 magnification, and each (darker) particle is segmented from the (lighter) background (Fig. 1) and filed along with a total of 194 image analysis (IA) measurements of its internal and external features. These include morphological, color, textural and Fourier descriptors, as well as geometric moments (for a full list, refer to Table 2). The IA software used (Halcon: MVTec GmbH 2004) is a platform-independent, "machine vision" software suite with a library of over 1150 operators. To facilitate the understanding of the individual contributions of each of these IA feature measurements, and to reduce the effects of the curse of dimensionality (Liu et al. 2001), classification tree models were developed with the Exhaustive search CHAID (chi-square automatic interaction detector) algorithm of SPSS AnswerTree (SPSS Inc 2004). The curse of dimensionality is a measure of the overall mean recognition probability as a function of input feature measurement complexity and database size. Given a pattern recognition environment with enough input complexity and a sizable database covering all classification scenarios, there is an underlying discrete probability structure found within (Hughes 1968). SPSS AnswerTree software highlights important segments of datasets where the best groups of salient features are efficiently determined by using scalable classification tree algorithms. SPSS AnswerTree is also able to give its own 'cross-validation' assessment (in percent) which is based on the predicted accuracy of the classification trees produced.

Twelve classification trees were constructed, three for each of the four classes (1st order, 2nd order amorphous, 2nd order palynodebris and 2nd order palynomorphs; Table 1) of varying accuracy based on SPSS AnswerTree cross-validation. Table 3 shows the number of salient features extracted from those available (194) at various accuracies. For a full description, including salient features identified for automated palynofacies studies, refer to Weller et al. (2006). The twelve sets of salient features were then extracted into databases that were used to train and test the ANN classifiers. For this, NeuralWorks (NeuralWare 2003) ANN software was utilized, a multi-paradigm ANN prototyping and development tool that can design, build, train, test and deploy ANNs to solve complex problems, such as classification.

Supervised Neural Networks

Supervised ANNs can be trained to recognize patterns within training data for which the correct classification is already known. The performance of the network is first evaluated by comparing the predicted classes of samples not seen by the network with their known class. The network can then be used to classify samples whose class is not known (Balfoort et al. 1992). As the identity of a particle is known in this system (Table 1), supervised ANNs are used for automated classification.

In general, such networks have an input and output layer with a number of simple processing units, or 'neurons', consisting of one or more hidden layers in between. Typically, layers receive signals, process them and feed them forward to units in the next layer. The way these signals are processed depends on the settings of the ANN

Table 2 Full list of 194 IA features measured. Manually assigned features are those that have a series of integrated steps within the IA software (Halcon) to derive a measurement based on the manual morphologically and texturally descriptive features (Table 1). Note the deletion of several Fourier shape descriptors; this is due to the repetition of measurements

Descriptor type(s)	Input	Specific measurement	Description
Morphological	1	Anisometry (external)	Elliptic radii (Ia/Ib) (calculated from geometric moments) (circle = 1)
	2	Bulkiness (external)	Relates elliptic axes to area (massiveness)
	3	Structure factor (external)	Elliptic shape parameter ((anisometry*bulkiness)-1)
	4	Elliptic Ra axis	Elliptic long (a) axis
	5	Elliptic Rb axis	Elliptic short (b) axis
	6	Convexity (external)	Measure of outward curves or bulges
	7	Compactness (external)	Area to contour length ratio
	8	Circularity (external)	Similarity to a circle (circle=1)
	9	Distance (external)	Mean distance of contour from area center
	10	Sigma (external)	Standard deviation of contour from area center
	11	Roundness (external)	Relation between distance and sigma (1-sigma/distance)
	12	Sides (external)	Number of polygon pieces if a regular polygon is concerned
Manually assigned	13	Small circle radius	Smallest surrounding circle
	14	Inner circle radius	Largest inner circle
	15	Circle difference	Small circle minus inner circle
Fourier	16	Real f0	Normalized Fourier shape descriptor—real (x) coefficient
	.	.	.
	54	Real f38	.
	55	Real f40	.
	56	Real f42	.
	.	.	.
	94	Real f80	.
	95	Imaginary f0	Normalized Fourier shape descriptor—imaginary (y) coefficient
	.	.	.
	133	Imaginary f38	.
	134	Imaginary f40	.
	135	Imaginary f43	.
	.	.	.
	172	Imaginary f80	.
Geometric moment	173	μ11	Normalized 2D boundary shape descriptor (product of inertia of axes through center parallel to coordinate axes)
	174	μ20	Normalized 2D boundary shape descriptor (2nd order line-dependent)

Table 2 (*Continued*)

Descriptor type(s)	Input	Specific measurement	Description
	175	μ02	Normalized 2D boundary shape descriptor (2nd order column-dependent)
Manually assigned	176	Straight lineations (external)	Amount of external straight lineations
	177	Curve lineations (external)	Amount of external curved lineations
Color	178	Mean red	Mean red color
	179	Mean green	Mean green color
	180	Mean blue	Mean blue color
Manually assigned	181	Straight lineations (internal)	Amount of internal straight lineations
	182	Curve lineations (internal)	Amount of internal curved lineations
	183	Radiate points	Amount of points that lineations radiate from
	184	Parallels (internal)	Amount of internal parallel lineations
	185	Number stretched lineations	Amount of lineations that stretch from side to side
Textural descriptors	186	Mean (internal)	Mean particle gray value
	187	Deviation (internal)	Gray value standard deviation over particle
	188	Entropy gray (internal)	Gray level disorder
	189	Anisotropy gray (internal)	Gray value symmetry
	190	Energy	Intensity of gray value distribution
	191	Correlation	Reciprocal relation of gray value distribution
	192	Homogeneity	Local similarity of gray values
	193	Contrast	Gray value differences over particle
Manually assigned	194	Number internal ROI	Amount of dark (near circular) inclusions

Table 3 Number of salient features (inputs) selected from 194 available (Table 2) for 80, 90 and 100% SPSS AnswerTree cross-validation accuracy. Gray cells indicate those databases used for RBF ANN classification

Database	Number of original inputs	Number of inputs for 80% accuracy	Number of inputs for 90% accuracy	Number of inputs for 100% accuracy
1st order	194	27	56	92
2nd order amorphous	194	1	9	25
2nd order palynodebris	194	8	13	23
2nd order palynomorphs	194	31	38	43

paradigm parameters. The neurons are connected in an organized way by interconnections of varying weights (Fig. 2). The training process involves repeatedly propagating different input vectors through the network from input to hidden to output layer until the network has learned the training samples. As vectors are passed forward through the network, they are multiplied by the appropriate weight associated with the interconnect. Initially weights are set to random values. At each neuron in the hidden layer, inputs are modified by a non-linear transfer function. In the output layer, the values associated with each hidden layer neuron are then summed, indi-

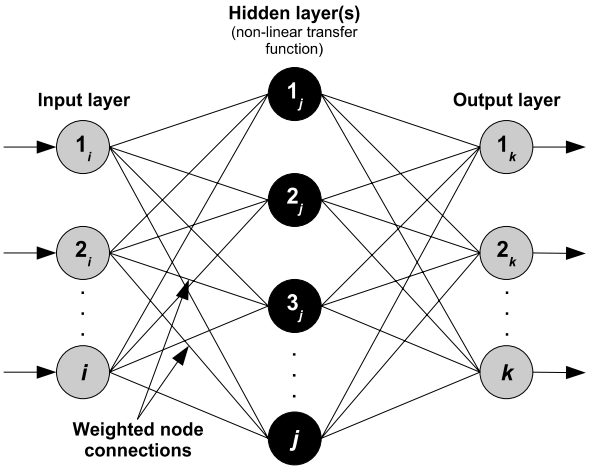

Fig. 2 Basic supervised ANN architecture displaying input layer (i.e. salient features: Table 3), hidden layer(s) and output layer (i.e. classification: Table 1)

cating the ANN's estimation of the corresponding output classifier (Malmgren and Nordlund 1997). The difference (or error) between the actual and expected output is then used as a basis for adjusting the weights and the parameters of the non-linear transfer function so as to reduce the difference. ANNs can also easily accommodate new classes of data by simply adding a new output node for each class and retraining the network. With multivariate statistical methods, this proves difficult as a new class would have to be introduced or programmed (Frankel et al. 1996).

Following the training process, the ANN is tested. Testing involves presenting the network with previously unseen examples for which the tester knows the correct classification. On the basis of a successful outcome (a high correspondence between the network's classifications and the actual classifications) the network can be used to classify data for which the classification is not known.

Radial Basis Function Neural Networks

The RBF ANN is a type of probabilistic neural network that was designed for classification tasks (Kartalopoulos 1996). They have been shown to have an efficiency of operation with the ability to produce reasonable decisions for, or reject patterns of, unknowns (Wilkins et al. 1996; Al-Haddad et al. 2000). They can also be easily retrained to incorporate unknowns into the network (Morris et al. 2001).

Previously, the RBF ANN was successfully applied to the classification of mineral deposits (Singer 2006). The results suggest that they can classify mineral deposits (when given comprehensive quantitative models) as well as experienced economic geologists. The difference between the BPN and RBF ANN lies in the neurons of the hidden layer. The RBF ANN replaces the BPN's sigmoidal non-linear transfer function with kernels (basis functions) to represent the data (Boddy et al. 2001). In this case, a Gaussian shaped kernel of the following form was chosen

$$Output = \exp\left(\frac{-x^2}{2\sigma^2}\right), \tag{1}$$

where σ controls the spread of the function, and x is the Euclidean distance between the kernel center and vector of interest. The kernels have a defined response to the input data that varies according to the distance of the data point from the kernel center. The kernel centers can be positioned with the aim of ensuring that their locations approximate the distribution of each class. For training, binary notation is required as an output. For example, if the classifier consists of three classes, three output columns are required, two of which could be 0, while the actual classification would be 1. During training, the positions of the kernel centers are determined along with the kernel widths and output layer vectors, respectively. Kernels are typically radially symmetric when the distance between the kernel center and vector of interest is Euclidean (as in this case), or non-radially symmetric (Wilkins et al. 1996). In testing, the output layer combines the nonzero response of the signals from the hidden layer whose magnitude is a function of the distance between the input and the kernel center and performs the classification (Boddy et al. 2001). The output node with the highest value (to a maximum of 1) is taken as the winner, and the network is said to have classified that particular input vector as belonging to the class represented by that node.

RBF ANNs were chosen because they have been found to produce consistent and accurate results (Culverhouse et al. 2002), display rapidity of training (Specht 1990), and to be at least as successful as other ANN paradigms or statistical methods (Wilkins et al. 1994; Culverhouse et al. 1996; Morgan et al. 1998). In addition, they do not make assumptions about distributions within the data (Singer 2006), and they are able to deal with incomplete datasets (Boddy et al. 1998).

Radial Basis Function Neural Network Results

For 1st order and 2nd order amorphous particle classification, the six original databases were used. Two classes each had 80, 90 and 100% SPSS AnswerTree cross-validation accuracy (Table 3). Each of the RBF ANN input vectors are made up of the salient features selected which were not normalized (values between 0 and 1). The output vectors indicate the class (Table 1). One layer of hidden neurons was used, their number determined heuristically by initiating with the same number of nodes as in the input layer and increasing by 10 until the best ANN configuration was found (the highest ANN recognition rate). If the number of nodes in the input layer were less than 10, then this number was used to initiate the number of neurons in the hidden layer. The ANNs were then trained until the lowest root mean square (RMS) error was found, which is an estimate of the standard deviation.

There were two test phases for the RBF ANNs. Initially, with both training and testing on the complete database (100% of data), a best classification rate of 89.41% was obtained for the 1st order class (with 80% SPSS AnswerTree cross-validation accuracy and 30 neurons in the hidden layer) and of 98.2% for the 2nd order amorphous class (with 90% SPSS AnswerTree cross-validation accuracy and 90 neurons in the hidden layer). This is an improvement over BPN ANNs of approximately 3 and 6%, respectively (Table 4). The best configuration was then used to test the RBF ANNs with a partitioned database (the best number of neurons in the hidden layer).

Table 4 RBF ANN results where 100% of data was used for both training and testing to determine best number of neurons in hidden layer

Database	SPSS AnswerTree cross-validation accuracy (%)	Number of neurons in hidden layer 1	RMS error	RBF ANN recognition rate (%)
1st order	80	10	0.43	85.73
1st order	80	20	0.41	87.84
1st order	80	27	0.52	78.08
1st order	80	30	0.39	89.41
1st order	80	40	0.42	86.56
1st order	80	50	0.46	83.96
1st order	90	10	0.48	81.78
1st order	90	20	0.47	82.82
1st order	90	30	0.45	83.99
1st order	90	40	0.45	85
1st order	90	50	0.45	84.48
1st order	90	56	0.5	79.98
1st order	100	10	0.5	80.16
1st order	100	20	0.47	82.42
1st order	100	30	0.47	82.55
1st order	100	40	0.47	82.52
1st order	100	92	0.5	79.36
1st; 2nd order amorphous	80	2	0.8	51.3
1st; 2nd order amorphous	80	10	0.69	71.66
1st; 2nd order amorphous	80	20	0.66	73.85
1st; 2nd order amorphous	80	30	0.65	73.85
1st; 2nd order amorphous	80	40	0.64	74.45
1st; 2nd order amorphous	80	50	0.64	75.05
1st; 2nd order amorphous	80	60	0.64	75.25
1st; 2nd order amorphous	80	70	0.64	75.25
1st; 2nd order amorphous	80	80	0.64	75.25
1st; 2nd order amorphous	80	90	0.64	75.05
1st; 2nd order amorphous	90	9	0.65	77.45
1st; 2nd order amorphous	90	10	0.45	90.02
1st; 2nd order amorphous	90	20	0.42	92.22
1st; 2nd order amorphous	90	30	0.4	93.21
1st; 2nd order amorphous	90	40	0.38	94.21
1st; 2nd order amorphous	90	50	0.37	94.21
1st; 2nd order amorphous	90	60	0.35	95.41
1st; 2nd order amorphous	90	70	0.33	96.81
1st; 2nd order amorphous	90	80	0.31	97.6
1st; 2nd order amorphous	90	90	0.3	98.2
1st; 2nd order amorphous	90	100	0.3	98

Table 4 (*Continued*)

Database	SPSS AnswerTree cross-validation accuracy (%)	Number of neurons in hidden layer 1	RMS error	RBF ANN recognition rate (%)
1st; 2nd order amorphous	100	10	0.43	90.62
1st; 2nd order amorphous	100	20	0.41	91.62
1st; 2nd order amorphous	100	25	0.54	83.83
1st; 2nd order amorphous	100	30	0.4	93.21
1st; 2nd order amorphous	100	40	0.39	94.01
1st; 2nd order amorphous	100	50	0.36	94.81
1st; 2nd order amorphous	100	60	0.33	96.01
1st; 2nd order amorphous	100	70	0.32	95.41

Table 5 Results of independently testing RBF ANNs on 10 and 20% of data

Database	SPSS AnswerTree cross-validation accuracy (%)	Test data size (%)	Number of neurons in hidden layer 1	RMS error	RBF ANN recognition rate (%)
1st order	80	10	30	0.43	85.93
1st order	80	20	30	0.46	84.1
1st order	90	10	40	0.44	85.02
1st order	90	20	40	0.46	83.03
1st order	100	10	30	0.53	77.37
1st order	100	20	30	0.48	81.35
1st; 2nd order amorphous	80	10	60	0.74	64
1st; 2nd order amorphous	80	20	60	0.78	62
1st; 2nd order amorphous	90	10	90	0.61	80
1st; 2nd order amorphous	90	20	90	0.55	85
1st; 2nd order amorphous	100	10	60	0.57	80
1st; 2nd order amorphous	100	20	60	0.56	83

This is done so that each ANN can be independently tested with previously unseen examples, and so that the recognition rate is not biased. Each of the six databases was split so that 80 and 90% of data was used for training and the remaining 20 and 10%, respectively, was used for testing the RBF ANN (Table 5).

Discussion

On initial examination, it can be seen that RBF ANNs have a difference of −0.79% mean, 2% mode and 1.28% median recognition rate when compared to BPN ANNs for both the 1st order and 2nd order amorphous classes (Table 6). With a standard deviation of 7.99% (RBF ANNs) and 5.06% (BPN ANNs), a difference of 4.82% is observed between the two paradigms. This anomaly is due to poor training of the

Table 6 Comparison of best BPN and RBF ANNs configurations with 10 and 20% of test data

Database	SPSS AnswerTree cross-validation accuracy (%)	Test data size (%)	BPN ANN recognition rate (%)	RBF ANN recognition rate (%)	Difference (%)
1st order	80	10	84.4	85.93	1.53
1st order	80	20	81.8	84.1	2.3
1st order	90	10	84.71	85.02	0.31
1st order	90	20	82.11	83.03	0.92
1st order	100	10	78.59	77.37	−1.22
1st order	100	20	84.71	81.35	−3.36
1st; 2nd order amorphous	80	10	78	64	−14
1st; 2nd order amorphous	80	20	66	62	−4
1st; 2nd order amorphous	90	10	78	80	2
1st; 2nd order amorphous	90	20	80	85	5
1st; 2nd order amorphous	100	10	80	80	0
1st; 2nd order amorphous	100	20	82	83	1
Average (mean)			80.03	79.23	−0.79
Average (mode)			78	80	2
Average (median)			80.9	82.18	1.28
Standard deviation			5.06	7.99	4.82

RBF ANN on the 2nd order amorphous class with 80% SPSS AnswerTree cross-validation accuracy. Having just one salient feature in this class (Table 3), there is a difference of −14 and −4 for 10 and 20% of the partitioned test data, respectively. This has proved difficult for an RBF ANN to train on.

If these two anomalous results are eliminated, the recognition rate difference between the RBF and BPN ANNs is improved by 0.85% mean, 0% mode and 1.12% median for both the 1st order and 2nd order amorphous classes. This is justified because this database has also been shown empirically to be too small and limited to facilitate the construction of accurate classifiers (Weller et al. 2005). This was assessed using the Gamma M-Test which is an indicator of the minimum quantity of data required to construct an effective model by knowing the variance of the statistical noise within it (Corcoran et al. 2003), which is inevitable in natural data. The standard deviation difference has also dropped to 0.27%. Separately, the RBF ANN demonstrated a classification improvement by approximately 0.5% for the 1st order class and approximately 1.5% for the 2nd order amorphous class. This is in accordance to previous RBF ANN classification studies. Table 7 gives the best ANN paradigm to use for each of the twelve databases after independent verification.

Across the spectrum of sedimentary OM classification, the initial BPN ANN series recognition rate of 87% has been improved to a best-case scenario of just over 4% to >91%. This is an improvement on previous automated sedimentary OM classification studies. Dabros and Mudie (1986) developed a system that identified and counted particulate OM which reduced labor-intensive user analysis (including identification and counting) by at least 60%, and France et al. (2000) developed a system

Table 7 Number of images per database, best ANN paradigm, test data size and ANN recognition rate

Database	SPSS AnswerTree cross-validation accuracy (%)	Number of available images	Best ANN paradigm	Test data size (%)	ANN recognition rate (%)
1st order	80	3266	RBF	10	85.93
1st order	90	3266	RBF	10	85.02
1st order	100	3266	BPN	20	84.71
1st; 2nd order amorphous	80	501	BPN	10	78
1st; 2nd order amorphous	90	501	RBF	20	85
1st; 2nd order amorphous	100	501	BPN	20	82
1st; 2nd order palynodebris	80	1475	BPN	20	98.65
1st; 2nd order palynodebris	90	1475	BPN	10	97.3
1st; 2nd order palynodebris	100	1475	BPN	10	98.65
1st; 2nd order palynomorphs	80	1290	BPN	20	94.57
1st; 2nd order palynomorphs	90	1290	BPN	10/20	94.57
1st; 2nd order palynomorphs	100	1290	BPN	10	92.25

to differentiate between three taxa of pollen grains which correctly classified to an average of 82%. This is also marginally better than previous RBF ANN-automated particulate classification studies. Culverhouse et al. (1996) identified 23 species of toxic and noxious dinoflagellates from European coastal waters to an accuracy of 85%. Jonker et al. (2000) identified 20 species of marine phytoplankton to an accuracy of 88.9% and Culverhouse et al. (2002) categorized 23 species of dinoflagellates to consistent and accurate results of between 72 and 90% recognition.

Summary and Conclusions

In a previous study (Weller et al. 2005), the authors tested a BPN ANN on twelve databases comprising 1st order, 2nd order amorphous, 2nd order palynodebris and 2nd order palynomorph sedimentary OM classes of different size, which gave an overall correct classification rate of 87%. Since the BPN ANN performed poorly for 1st order and 2nd order amorphous classes, the authors tested a RBF ANN on the same databases in an attempt to improve the recognition rate. This demonstrated a classification improvement by approximately 0.5% for the 1st order class and approximately 1.5% the 2nd order amorphous class, making their incorporation in an ANN series appealing for sedimentary OM classification. The initial BPN ANN series recognition rate of 87% has been improved to a best-case scenario of just over 4% to >91%. For better recognition performance, a classifier consisting of both BPN and RBF ANN paradigms may be utilized for the classification of sedimentary OM. For a productive system, a RBF ANN may be used to classify the 1st order and 2nd order amorphous classes, while a BPN ANN should classify the 2nd order palynodebris and palynomorph classes (Fig. 3).

This system was not designed to replace or threaten human experts. Instead, it was designed to facilitate their work by reducing the cost (and time) involved in analysis,

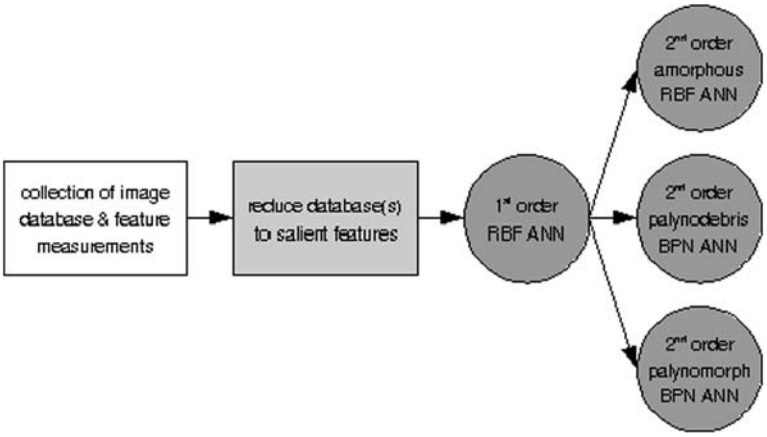

Fig. 3 The system initially captures an image of a particle which is segmented from background and filed in database along with 194 feature measurements (Table 2). Each database is then reduced in size to associated salient features (Table 3). These are finally passed through RBF and BPN ANN series to provide classification

and to accommodate precise, consistent and continuous operation. It frees the worker from routine microscopy tasks with automatic slide scanning and increases objectivity in classification. The system acts as a new tool for the quantification of particle numbers and characteristics by increasing the sampling volume (and thus the accuracy). It provides standardized and automated measurement procedures. Ultimately, an electronic morphotype (species?) database may be developed through this system.

Acknowledgements We would like to thank the anonymous reviewers for helpful comments on the manuscript. This work was made possible by support from the EPSRC (GR/N28405/01), to whom the Authors are grateful.

References

Al-Haddad L, Morris CW, Boddy L (2000) Training radial basis function neural networks: effects of training set size and imbalanced training sets. J Microbiol Methods 43:33–44

Balfoort H, Snoek J, Smits J, Breedveld L, Hofstraat J, Ringelberg J (1992) Automatic identification of algae—neural network analysis of flow cytometric data. J Plankton Res 14(4):575–589

Batten DJ (1996) Palynofacies and palaeoenvironmental interpretation. In: Jansonius J, McGregor D (eds) Palynology: principles and applications. American Association of Stratigraphic Palynologists Foundation, Salt Lake City, pp 1011–1065

Boddy L, Wilkins MF, Morris CW (1998) Effects of missing data on neural network identification of biological taxa: RBF network discrimination of phytoplankton from flow cytometry data. Intell Eng Syst Artif Neural Netw 8:655–660

Boddy L, Wilkins MF, Morris CW (2001) Pattern recognition in flow cytometry. Cytometry 44:195–209

Boulter MC (1994) An approach to a standard terminology for palynodebris. In: Traverse A (ed) Sedimentation of organic particles. Cambridge University Press, Cambridge, pp 199–217

Corcoran J, Wilson ID, Ware JA (2003) Predicting the geo-temporal variations of crime and disorder. Int J Forecast 19(4):623–634. Special Issue on Crime Forecasting

Culverhouse PF, Simpson RG, Ellis R, Lindley JA, Williams R, Parisini T, Reguera B, Bravo I, Zoppoli R, Earnshaw G, McCall H, Smith G (1996) Automatic classification of field-collected dinoflagellates by artificial neural network. Mar Ecol Prog Ser 139:281–287

Culverhouse PF, Herry V, Ellis R, Williams R, Reguera B, González-Gil S, Umani SF, Cabrini M, Parisini T (2002) Dinoflagellate categorisation by artificial neural network. Sea Technol 43(12):39–46

Dabros MJ, Mudie PJ (1986) An automated microscope system for image analysis in palynology and micropaleontology: current research, part A. Geol Surv Can 86-1A:107–112

France I, Duller A, Duller G, Lamb H (2000) A new approach to automated pollen analysis. Quat Sci Rev 19(6):537–546

Frankel D, Frankel S, Binder B, Vogt R (1996) Application of neural networks to flow cytometry data analysis and real-time cell classification. Cytometry 23(4):290–302

Hughes GF (1968) On the mean accuracy of statistical pattern recognizers. IEEE Trans Inf Theory IT-14(1):55–63

Jonker R, Groben R, Tarran G, Medlin L, Wilkins M, Garcia L, Zabala L, Boddy L (2000) Automated identification and characterisation of microbial populations using flow cytometry, the AIMS project. Sci Marina 64(2):225–234

Kartalopoulos SV (1996) Understanding neural networks and fuzzy logic: basic concepts and applications. IEEE Press, Piscataway, 205 p

Liu J, Dazzo F, Glagoleva O, Yu B, Jain A (2001) CMEIAS: a computer-aided system for the image analysis of bacterial morphotypes in microbial communities. Microb Ecol 41(3):173–194

Malmgren BA, Nordlund U (1997) Application of artificial neural networks to paleoceanographic data. Palaeogeogr Palaeoclimatol Palaeoecol 136:359–373

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(8):975–984

Morris CW, Autret A, Boddy L (2001) Support vector machines for identifying organisms—a comparison with strongly partitioned radial basis function networks. Ecol Model 146:57–67

MVTec GmbH (2004) Halcon. Available from: http://www.mvtec.com/halcon/ [Accessed 2 Nov 2005]

NeuralWare (2003) NeuralWorks professional II/PLUS. Available from: www.neuralware.com/products_pro2.jsp [Accessed 20 Apr 2004]

Singer DA (2006) Typing mineral deposits using their associated rocks and grades and tonnages in a probabilistic neural network. Math Geol 38(4):465–475

Specht D (1990) Probabilistic neural networks. Neur Netw 3:109–118

SPSS Inc (2004) AnswerTree. Available from: http://www.spss.com/answertree/ [Accessed 26 Apr 2005]

Traverse A (1988) Paleopalynology. Allen and Unwin Ltd, London, 600 p

Tyson RV (1995) Palynological kerogen classification. In: Tyson RV (ed) Sedimentary organic matter, organic facies and palynofacies. Chapman and Hall, London, pp 341–367

Weller AF (2004). The semi-automated classification of sedimentary organic matter and dinoflagellate cysts in palynological preparations. Doctoral dissertation, University of Glamorgan, UK, 170 p

Weller AF, Corcoran J, Harris AJ, Ware JA (2005) The semi-automated classification of sedimentary organic matter in palynological preparations. Comput Geosci 31(10):1213–1223

Weller AF, Harris AJ, Ware JA, Jarvis PS (2006) Determining the saliency of feature measurements obtained from images of sedimentary organic matter for use in its classification. Comput Geosci 32(9):1357–1367

Wilkins MF, Morris CW, Boddy L (1994) A comparison of radial basis function and backpropagation neural networks for identification of marine phytoplankton from multivariate flow cytometry data. Comput Appl Biosci 10(3):285–294

Wilkins MF, Boddy L, Morris CW, Jonker R (1996) A comparison of some neural and non-neural methods for identification of phytoplankton from flow cytometry data. Comput Appl Biosci 12(1):9–18