PHENOMENON OF TOLERANCE TO DAMAGE IN ARTIFICIAL NEURAL NETWORKS

Ryszard Tadeusiewicz*, Izabela Figura

AGH University of Science and Technology
*Corresponding author: rtad@agh.edu.pl

Abstract

Neural networks are computer tools with many advantages. They are powerful because of possibility of complex nonlinear systems modeling and they are user-friendly because of learning abilities. But neural networks have additional advantages. One of them is increased tolerance to damages. Technological systems are mainly not resistive for damages. Microprocessor if is even a little damaged – just doesn't work. In contrary to this biological systems, especially brain are almost insensitive for damages – many cell can be dead or can have many malfunctions – but the brain as a whole can functioning properly. The research presented in the paper is dedicated for discovery, if neural networks, as a models of natural neural system elements, can show tolerance to damages. This assumption was experimentally proven and results of simulations showing such resistance, are presented in the paper.

Key words: artificial neural networks, tolerance to damage

1. INTRODUCTION - BIOLOGICAL HERITAGE OF NEURAL NETWORKS

We have been accustomed to the fact that neural networks are used as computational tools with a variety of beneficial characteristics (Tadeusiewicz, 2010a). Theoretically, we know that these beneficial characteristics stem from the fact that neural networks are models of real nervous system fragments and this, for example, results in their ability to learn. However, on an everyday basis we do not focus on this issue since it is not needed during the practical use of networks - for example, for the purpose of creating models of various processes and systems. Meanwhile, neural networks **are** really models of brain fragments, which is demonstrated in figure 1 showing how the knowledge about the brain acquired by biologists has been converted by bio-cyberneticians into the form of abstract mathematical models that have become a basis for the creation of neural networks or, in other words, functional information technology tools (Tadeusiewicz, 2010b).
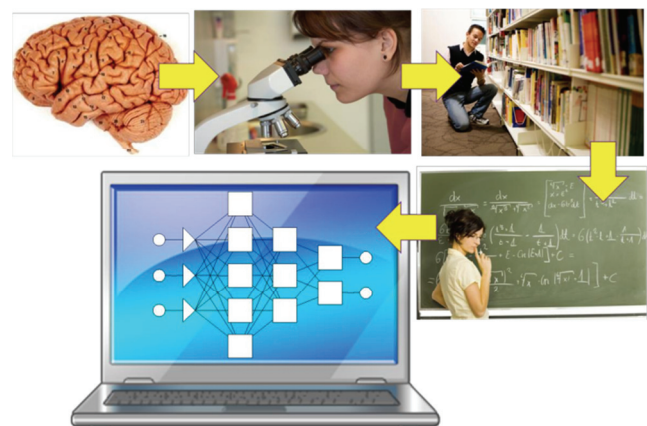
Fig. 1. *A simplified model showing the transition from the knowledge about the brain to artificial neural networks (own model created with the use of ClipArt elements from MS Office).*

The fact that neural networks are modelled after the biological brain has various, often surprising, consequences. This article describes one of such consequences.

approximately the same as the proportion between the mass of a mouse and the total mass of ten elephants (figure 2).
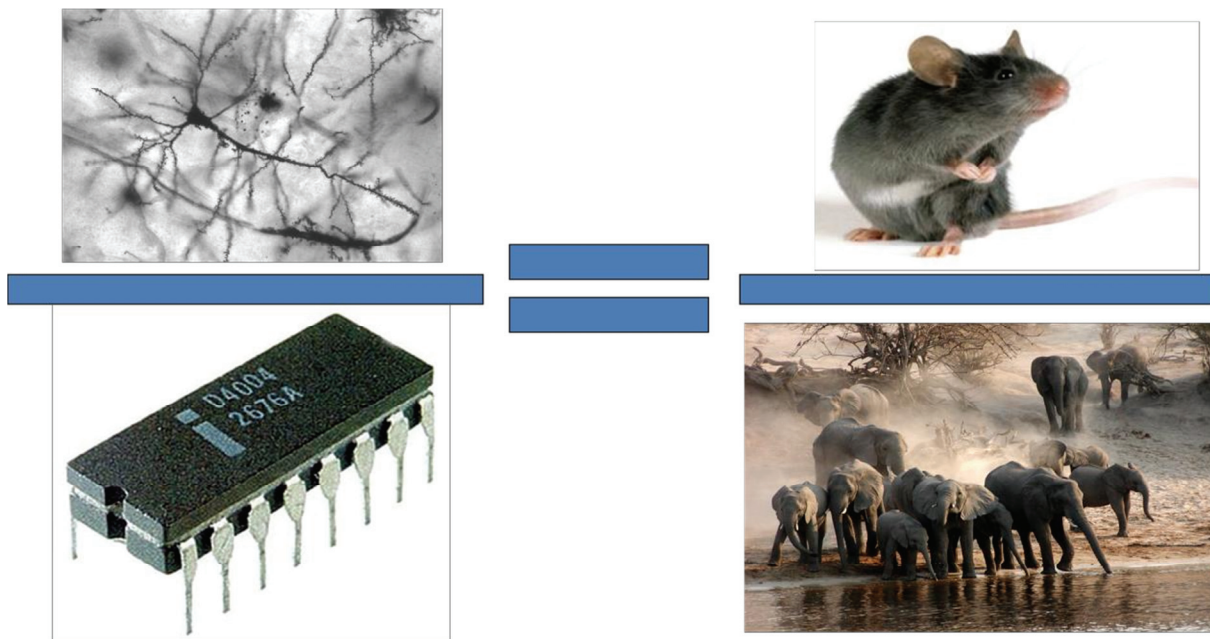

Fig. 2. Proportion graphically showing how much a neuron is slower than a microprocessor.

We shall begin with a statement known among neurobiologists and clinical neurologists (Longstaff, 2002), but usually unknown to technicians and scientists. Our brain that we are so proud of and that we value so much as a source of our intelligence is made of a very poor material. One could point at numerous signs of this poor quality of the neurobiological material. For example, the operational speed of "biological processors", i.e. nerve cells, is at least six orders of magnitude (million times!) lower than the speed of typical electronic processors. It suffices to compare gigahertz frequencies of clocks in modern computers with physiological data showing that brain cells use signals with frequencies usually not exceeding several hundred hertz, and one kilohertz is the ultimate upper limit for them. The speed of communication in the brain is also much smaller in a similar proportion (million times!) than the speeds easily achieved in modern telecommunications. The electronic signal in a fibre-optic cable or a copper cable reaches the speed of 300 thousand kilometres per second, which is obviously the highest speed to be achieved in Nature. In comparison, a nerve impulse in fibres that connect neurons usually achieves the speed of several dozen centimetres per second. The proportion between the speed of a biological neuron and the speed of a microprocessor is then
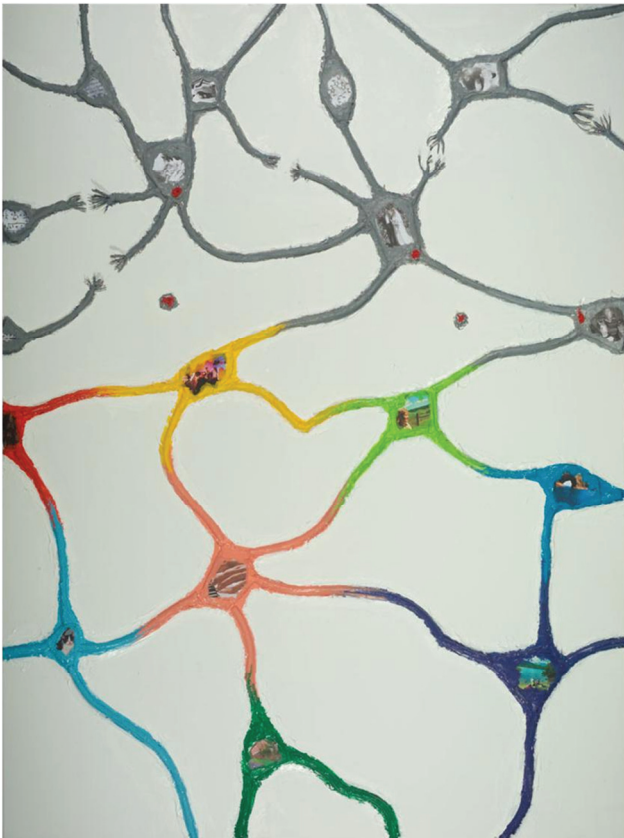
2. NEURODEGENERATION

Unfortunately, there is yet another affliction that besets both technical information processing systems and biological neural structures. It is susceptibility to damages. Both systems break down but the biological ones break down much more often. In fact, they break down constantly! Beginning with the age of 12 (in women) and 16 (in men), neurons in the brain start dying. Large amounts of neurons – from several hundred to several dozen thousand **per day.** Their death is irrevocable as the natural regeneration of nerve cells does not occur in the human brain. The truth is that as far as the amount of the nervous tissue is concerned, humans are born with a certain amount of it that they constantly lose throughout all their life. This process called neurodegeneration is presented in figure 3 in an artistically simplified form.

In this figure, neurodegeneration is marked as the filling of cell interior with grey colour and as the loss of connections between cells. In this very symbolic figure, one may see that some information resources (presented as pictures inside cells) are lost and that some routes of signal transfer cease to function. This results in the impoverishment of the

nervous system functions and the system becomes less efficient.

Fig. 3. The process of neurodegeneration - visible in the upper part of the network - shown as an artistic metaphor (Source: http://www.endocytosis.org/ImaginingTheBrain/NeuroArt2009/ SummerWork2009.html last access August 2011).

As it has been mentioned above, figure 3 is an artistic metaphor, a visual representation of an idea, rather than a biological reality, but the same is shown in figure 4, which is a real microscope image of the cerebellum tissue with the developing process of neurodegradation (visible in the upper part so as to emphasise the analogy with figure 3). One may notice that the bodies of information processing neurons (with visible black dots in the middle resulting from the method of colouring) disappear in the upper part of the preparation and mainly glial cells that do not process information are left there.

The forms of nerve cell damages, which are the symptoms of neurodegeneration, vary (figure 5), but the result is always the same: some nerve cells stop functioning or malfunction.

Why do we pay so much attention to the discussion of this regrettable susceptibility to damages of the natural nervous tissue?
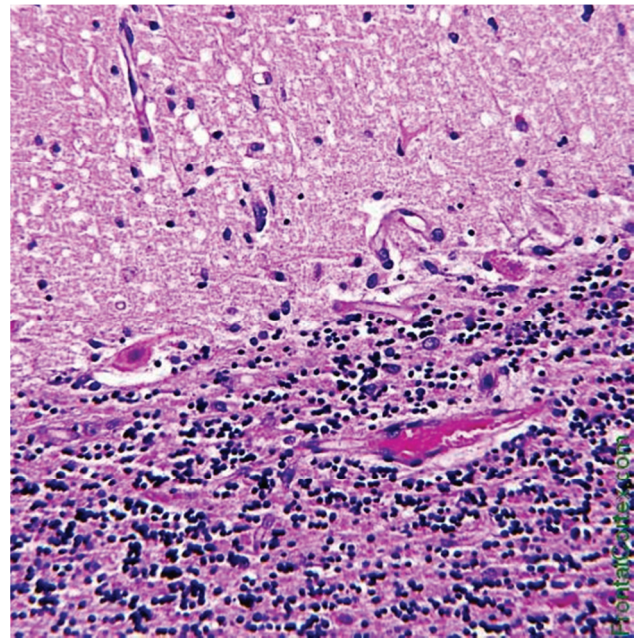
Fig. 4. A microscope preparation of the cerebellum with a visible process of neurodegeneration in the upper part (Source: http://frontalcortex.com/gallery/pics/gliageek_EosinophilicNeur onalDegeneration CerebellumX200.jpg last access August 2011).

This is due to the fact that the process of dramatic degradation of nervous structures constantly taking place in every human being for a long period of time practically does not influence the realisation of complex biological functions by the neural system. We say that the biological nervous system shows **an increased tolerance to damage.** Although its elements become worn out and damaged to a surprisingly large extent and scope, the structure as a whole fulfils its functions efficiently and effectively despite the loss of these elements. At least we perceive it in this manner while subjectively observing our own mental processes that accompany various activities and while objectively analysing the behaviours of other people. Obviously, when damage caused in the human brain by neurodegeneration becomes too extensive, we observe the deterioration in the quality of the brain functioning and the loss of some of its functions. However, it occurs only in the case of very serious neurodegenerations, such as in Alzheimer disease. This disease causes such extensive damages in human brain (figure 6) that even compensatory mechanisms functioning in the case of smaller damages are unable to change this.
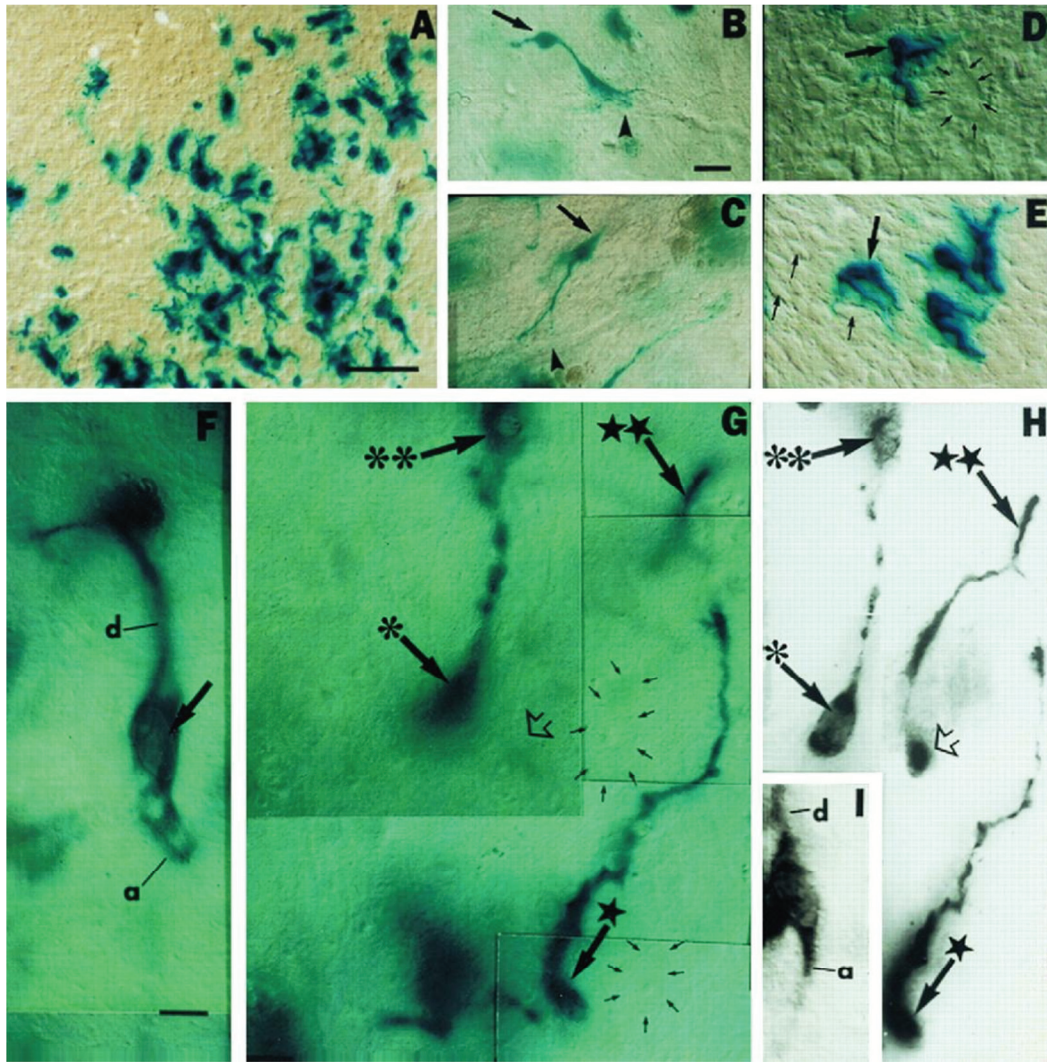
Fig. 5. *Various forms of nerve cell damages seen in microscope images*
(Source: http://www.pnas.org/content/94/21/11663/F2.large.jpg last access August 2011).
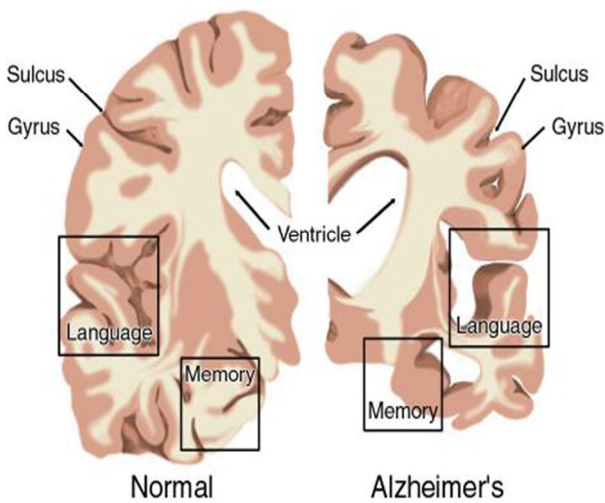
Fig. 6. *Section of a normal and pathological brain shows extensive damage caused by Alzheimer disease.*
(Source: http://www.b12patch.com/blog/wp-content/uploads/ 2009/07/alzheimer-brain.bmp, last access August 2011).

3. STUDY OF SUSCEPTIBILITY TO NEURAL NETWORK DAMAGE

The considerations presented above are aimed at pointing to the basic difference between the **results** of damage we deal with in the biological nervous system and in computer tools designed by modern electronics. The products of modern technology, especially computers, are characterised by the fact that the break-down of one component usually causes the failure of the whole system. Meanwhile, components of the human brain constantly break down and the brain as a whole functions incessantly without a noticeable deterioration of its functioning.

The question arises whether neural networks being technical products but reflecting the brain structure will be able to show this increased tolerance to damages.

This issue keeps appearing in numerous studies concerning neural networks Graupe, 2007);

however, the majority of authors merely mention the fact that neural networks probably possess the feature of increased tolerance to damages and the authors of this study have not managed to find a single study in which this phenomenon would be studied or analysed in more detail. Bearing this in mind, an experiment has been designed and conducted; its aim has been to collect particular empirical data on the issue whether in artificial neural networks there is the phenomenon of an increased tolerance to damage and on how the reliability of the whole network's functioning depends on the extent of the introduction of purposefully damaged elements.

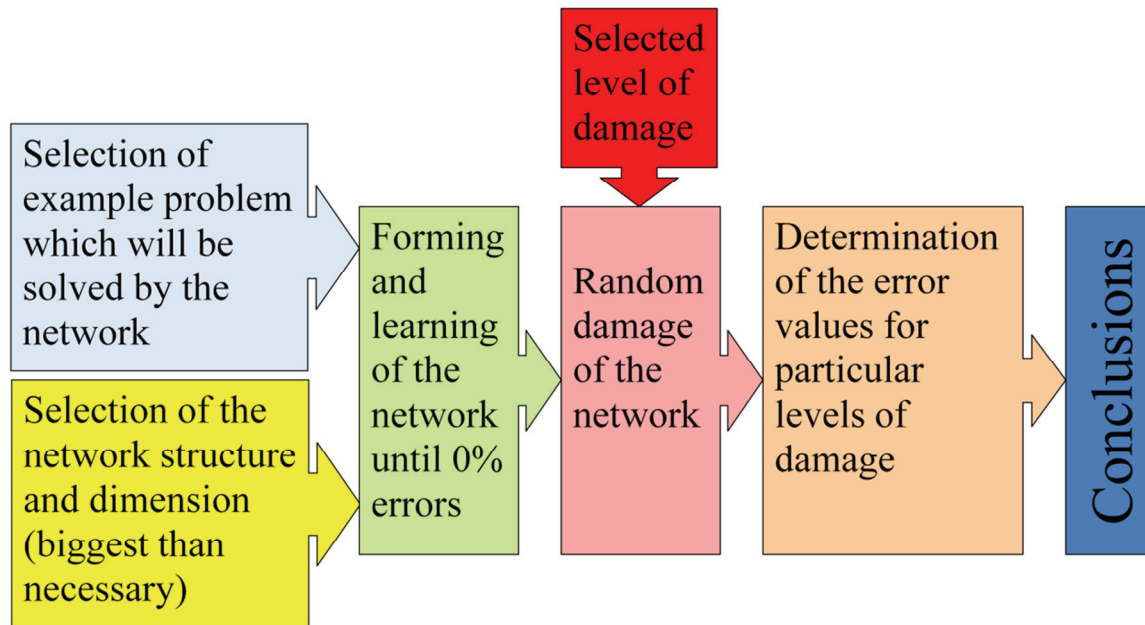An outline of the designed and conducted experiment is shown in figure 7.

characterised by the fact that its properties in the context of applying various machine learning systems are very well-known, so it will be possible to compare the results obtained with the data from the literature and interpret them quite easily. The task selected is the problem of the recognition of iris species on the basis of the structure of their petals. This example has been chosen because the expected deterioration in the quality of the functioning of a (deliberately) damaged neural network will be easier to detect while requiring the classification (recognition) of objects from the network than for example while observing changes in the value of the output signal in the network accomplishing a regressive model of a given process. In the case of a classifying network, the situation is always clear and easy

Fig. 7. Outline of the designed and conducted experiment.

## 4. CHOICE OF A HYPOTHETICAL PROBLEM SOLVED BY THE NETWORK

The first element of the outline shown in figure 7 is the choice of a hypothetical problem that the tested network will solve – first as a complete and fully functioning structure and then as a structure in which some artificial damage has been deliberately done. It is obvious that the type of a task solved by the network influences the behaviour of the network when damages occur. A detailed analysis of the relation between the type of a task and the extent of the network's susceptibility to damages to its structural elements will require more thorough and in-depth studies. For the purpose of this study, only one task has been selected; the task is

to assess: the network either classifies correctly or makes mistakes. In the network producing an output signal in the form of numerical values, divergences between what the network does and what it should do require more subtle assessments. The second argument related to the choice of the task of iris recognition stems from a commonly known feature of this task consisting in the fact that it comprises both a simple and a difficult classification problem. One may observe how a damaged neural network loses its abilities – both easily acquirable abilities (relating to obvious differences between iris species) and those whose learning requires a longer period of time (using more subtle and complex differences between the species).

In order to present the characterisation of the task in full, we should add that the studies have been conducted with the use of a popular and easily accessible database (*Fisher's Iris data set*), which is a multidimensional set introduced for general use by an English scientist Ronald Fisher in 1936. The database contains 50 examples for each species out of the three known iris species (*setosa, virginica,* and *versicolor*). Four features have been determined for every object in the database: length and width of sepals and petals, as shown in figure 8.
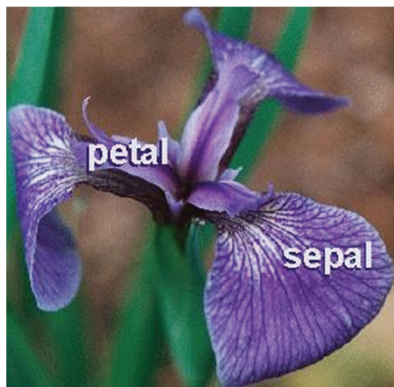
Fig. 8. *Petals and sepals of an iris.*
(Source: http://www.mathworks.com/help/toolbox/stats/bqzdnrv-1.html, last access August 2011).

The vector of input data to the network consists of four elements $X = <x_1, x_2, x_3, x_4>$ and contains the following components:

x_1 – Sepal length
x_2 – Sepal width
x_3 – Petal length
x_4 – Petal width

While representing the input data, the rule has been assumed according to which one iris species is coded on the following basis "*one-out-of-N,*" which means that at the output of the network there is a three-element vector with the following structure (values that should occur in the case of an ideal recognition of particular iris species are given):

Y = <1, 0, 0>   for *setosa* iris
Y = <0, 1, 0>   for *virginica* iris
Y = <0, 0, 1>   for *versicolor* iris

The full repertoire of the data concerning irises (downloaded from the database being an attachment to the package *Neural Network Toolbox* of the

Matlab programme) has been divided into L, V, and T subsets, according to the proportion:

L – 75% of input data treated as training data,
V – 15% of input data randomly selected as validation data,
T – 15% of input data used as test data.

As it has been mentioned above, the database used in the study is known to contain a simple task and a difficult task. This stems from the distribution of points corresponding to irises of particular species in space determined by elements of the *X* vector. It is impossible to draw a picture in a four-dimensional space but by selecting any three of these four coordinates we can see an image such as the one shown (for example) in figure 9.
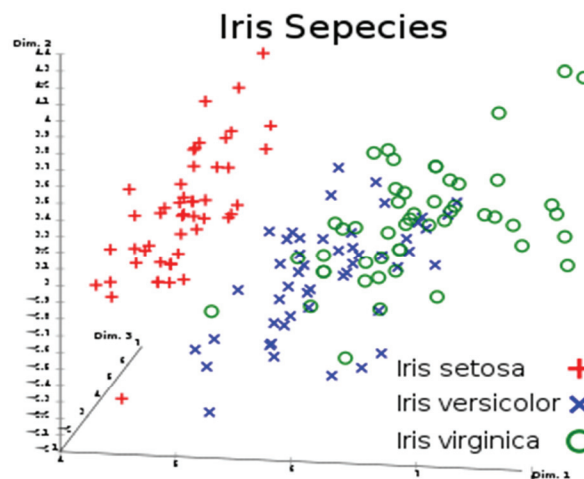
Fig. 9. *Distribution of points representing particular iris species in a three-dimensional space.*
(Source: http://en.wikipedia.org/wiki/File:Iris_Flowers_Clustering_kMeans.svg, last access August 2011).

One may easily see that the distinction between the *setosa* iris and the other two species should not give rise to any difficulties. However, the distinction between the species *versicolor* and *virginica* will be difficult and will require the use of all four input variables.

5. STRUCTURE OF A RECOGNISING NEURAL NETWORK AND ITS LEARNING

In order to recognise iris species, a small neural network would be sufficient, as although this task is partially quite difficult, it may be solved with the use of simple means. However, in the studies described in this article we want to have a network that

would possess numerous elements because later on we intend to damage some of them. This is why a highly surplus network with four inputs (corresponding to four elements of the vector *X*) and three outputs (corresponding respectively to the three species of recognised irises) has been selected. The internal network structure contains two hidden layers: the first one with 40 neurons and the second one with 30 neurons. The connections between the layers are round robin connections, which means that the network has in total 1450 connections (4 x 40 between the input and the first hidden layer, 40 x 30 between the two hidden layers and 30 x 3 between the second hidden and the output layer).

During the process of learning, in every learning layer (in both hidden layers and in the output layer) the values of weight matrix *W* and of the vector of threshold values (*bias*) *b* are set.

The structure of the network used is shown synthetically in figure 10. A simplified network presentation without drawing every individual neuron and without drawing all the connections has been applied, as with the dimensions mentioned above such a drawing would be totally illegible.

The network has been taught with the *back-propagation* method using the training set L defined above, with the validation set V in order to prevent over-learning, and the behaviour of the
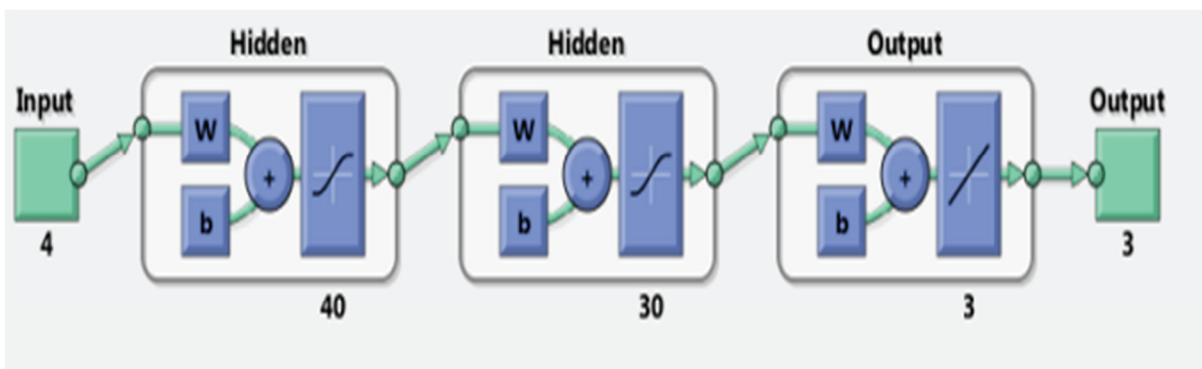

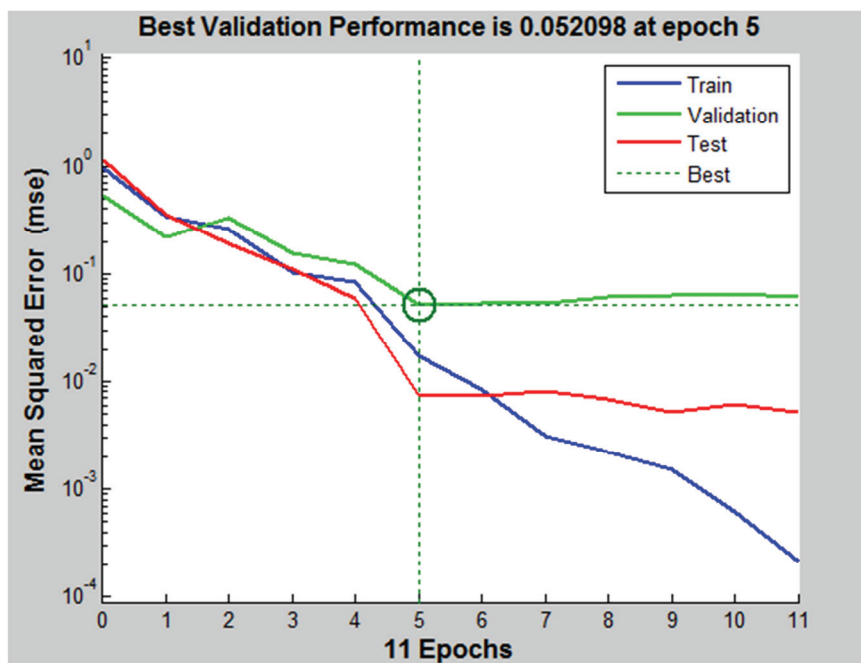**Fig. 10.** *Structure of the neural network used in the studies.*

Fig. 11. *Course of errors during network learning.*

Neurons of both hidden layers are non-linear and have bipolar sigmoidal transfer function according to the hyperbolic tangent mathematical formula, whereas neurons of the output layer are linear ones.

network has been (independently) tested with the test set T. The course of errors during the process of learning is shown in figure 11.

Learning has been continued up until 11 epochs and continuous improvement of error in the training set is obtained; however, beginning with the 5th epoch the improvement of results in the validation set has ceased being noticeable (and even a slight deterioration of results in this set could be observed), which would suggest the occurrence of the over-learning effect. The process of learning has then been stopped at five generations, assuming the obtained result of learning as a basis for further studies. It is worth noticing that the result of learning is not ideal (the level of 0% errors has not been achieved), which after verifying with the real data has been manifested in the fact that whereas irises of *setosa* and *virginica* species have been recognised correctly in 100%, 10% of *versicolor* irises (namely – 5 pieces) have been recognised incorrectly (as *virginica*). A glimpse on figure 9 allows us to notice the reason for this fact.

The analysis of the errors made by the network (figure 12) makes it possible to state that the errors are on the whole insignificant and undergo normal distribution, which suggests their random (unsystematic) character.
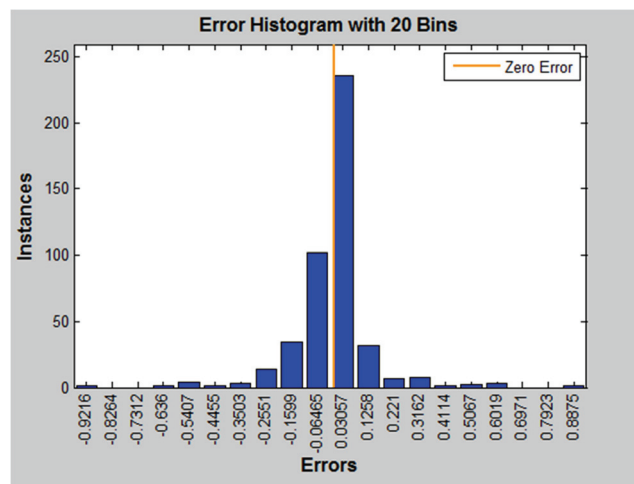

Fig. 12. *Histogram of network errors.*

Figure 13 shows responses of output neurons of the trained network in the form of the height of pyramids. These are the values produced at a given output described with the species name that represents this output with *"one out of N"* coding. The signal presented on the graph shows an average value for the whole set of data from a selected class. The data is grouped (from the left to the right) in
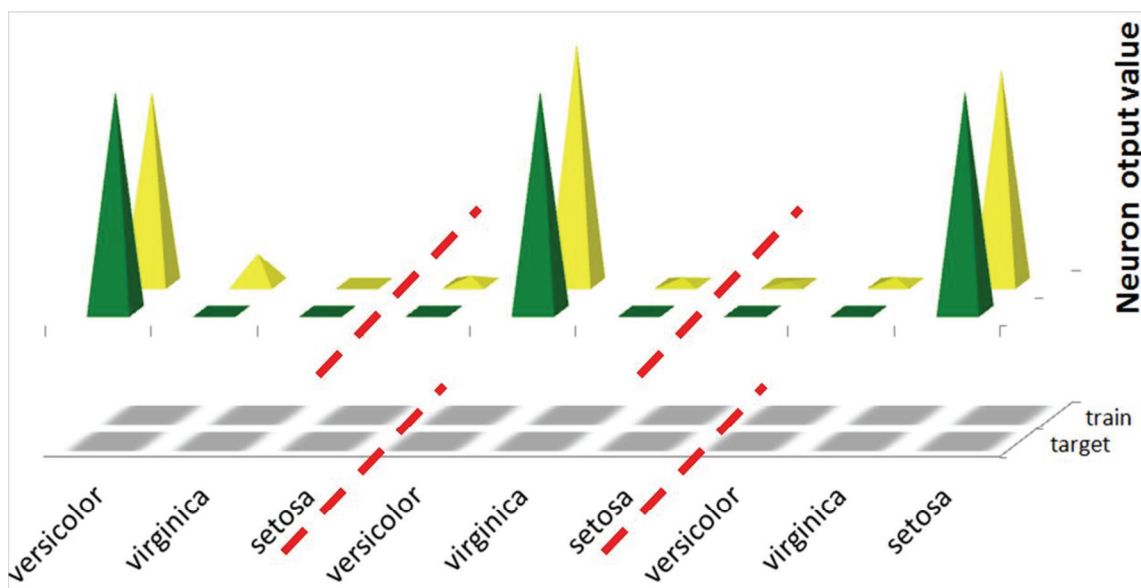

Fig. 13. *Illustration of the behaviour of a trained undamaged network.*

Figure 13 shows the result of functioning of a trained network in a form worth analysing in detail since later on a similar form will be used to present the functioning of a network subject to damages. Since next figures will be much more complex, the analysis and comprehension of the suggested method of result presentation with this first simple example will be of much help while interpreting all the subsequent results.

such a way that the first three pyramids represent responses of the network while presenting its data concerning irises of the *versicolor* species, the second group of three output signals shows values of averaged network responses when its input has been presented with the data on irises of the *virginica* species, and the third one respectively belongs to *setosa* irises. In figure 13, these groups of data are additionally separated with broken lines, which

will be missing in subsequent figures due to the legibility of these figures. In front of the pyramids representing the behaviour of the trained network there are pyramids (marked with a dark colour) showing model behaviour (target) of an ideal classifier. It is worth noticing that the pyramids in the figure are lifted above the level of the ground because when the same type of the chart will be used to present the behaviour of a network with damages, then some averaged values of network responses will be **negative**, which will require the possibility of drawing such signals - additionally in these subsequent figures presented as a pyramid with the vertex directed downwards. A precise localisation of the pyramid in relation to the chart axis is shown by its "shadow" projected onto the basis of the chart.

Figure 13 confirms what has basically been noticed; namely, that while recognising irises of the *versicolor* species errors sometimes occur (quite a big pyramid in the column corresponding to the output signalling the species *virginica*). But a detailed analysis of the figure shows that while recognising other species the network does not function perfectly, because although it has made correct decisions, non-zero values of signals at "false" outputs have occurred in the case of the recognition of all species. However, they have been small and have not led to the neural classifier taking false decisions.

6. STUDY OF NETWORK DAMAGE RESULTS

The neural network formed in the way described in the previous chapter was subject to studies concerning the influence of damages on its functioning. It was assumed that the damages studied would have the form of **disruptions** of randomly selected connections inside the network. In practice, it was done in such a way that randomly selected weight coefficients were simply cleared because a signal did not pass through a connection whose weight equalled zero. This process of random clearing of weights was continued until an assumed percentage of damaged connections was achieved in the whole network.

In order to observe better the changes in the network behaviour during the process of connection damaging as their number increased, networks with respectively 1%, 2%, 5%, 10%, 15%, 20%, 30%, 48%, 60%, 70%, 80% and 90% of damaged connections were simulated.

Obviously, the random selection of damage localisation caused that a single study of a damaged network behaviour did not give reliable results. It could happen that randomly selected damage locations would hit a crucial region of the network structure or – conversely – only insignificant connections were damaged. This was why every damage level was simulated in the network ten times in order
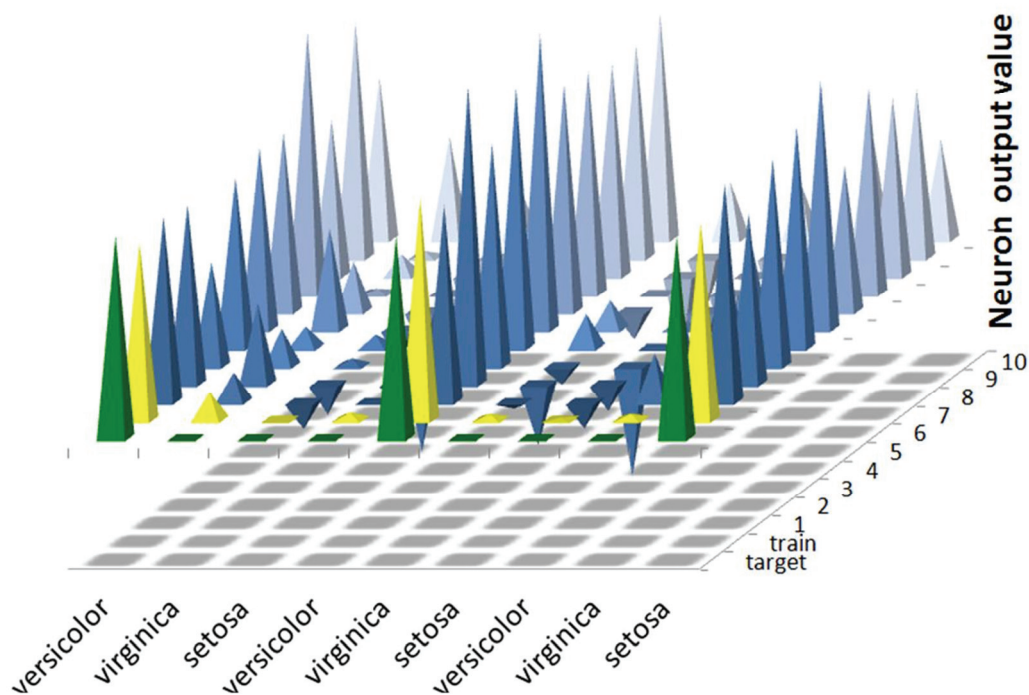

Fig. 14. *Behaviour of 10 networks with damages at the level of 2% of their elements. Discussion in the text.*

to obtain an appropriate amount of data for the statistical analysis. Every time the process of damaging began with a full network without any damage, so each network out of the ten networks studied with the same proportion of damaged elements had in fact a different set of damaged connections. A sample result of a set of such simulations for network damages at the level of 2% is shown in figure 14.

Just as figure 13, this figure shows the responses of output neurons in the form of pyramid heights, first in an undamaged network (in the front) and then in subsequent rows in particular damaged networks numbered from 1 to 10. One may notice that for every experiment with the damage of 2% of network elements, the quality of network functioning has slightly decreased but it has not led in any of the analysed examples to **an additional**
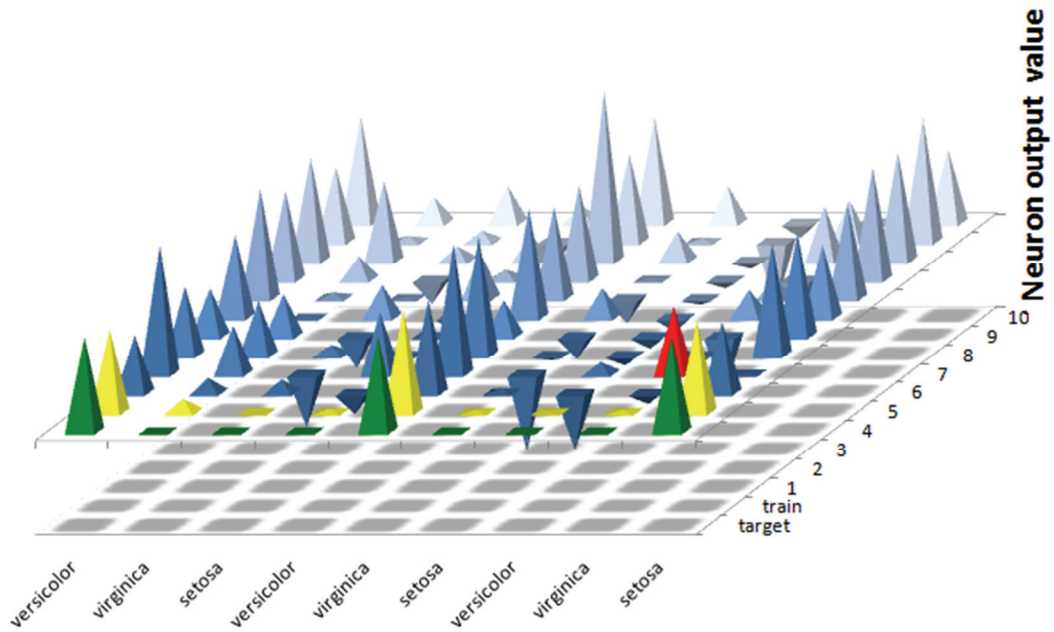

Fig. 15. *Behaviour of 10 networks with damages at the level of 5% of their elements.*
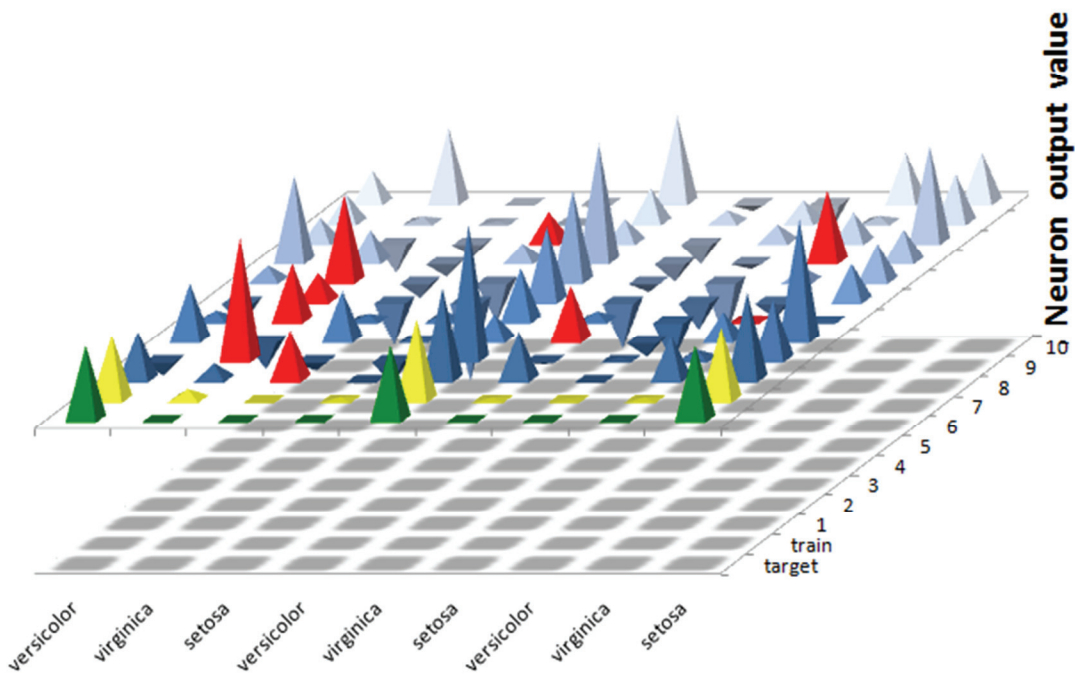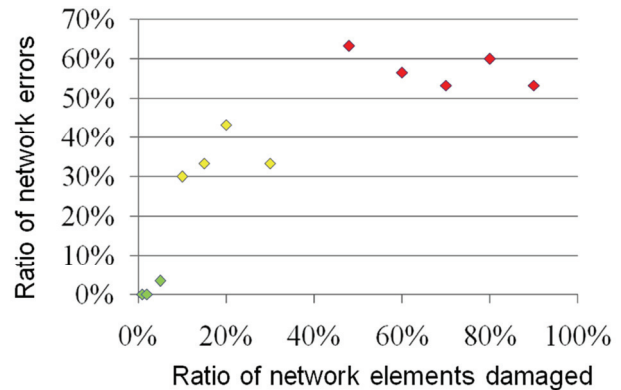

Fig. 16. *Behaviour of 10 networks with damages at the level of 10% of their elements.*

erroneous recognition (let us remind you that an undamaged network has misrecognised 5 examples of *versicolor* irises as *virginica*, which has been left but cannot be put down to damages in the network). What has happened as a result of the analysed level of damage is expressed merely by lower certainty of recognition; as a consequence of network damage, the output signal of a neuron signalling a correct solution has slightly decreased, whereas output signals of neurons that may cause erroneous recognitions have slightly increased. However, despite the impairment of recognition certainty, its correctness has not been affected.

A similar network behaviour was obviously observed at a smaller level of elements damaged (1% of broken connections) and at the level of a slightly higher percentage of damages (5%), although in that case one erroneous recognition occurred – and, what is interesting, within a generally simple task of distinguishing between *setosa* and *virginica* (figure 15).

another. What is characteristic is the fact that errors first occur where the task performed by the network is more difficult (distinguishing between *versicolor* and *virginica* irises), though also other recognitions begin to be erroneous.

Fig. 18. *Correlation between the percentage of damaged elements in the network and the percentage of errors made by the network.*
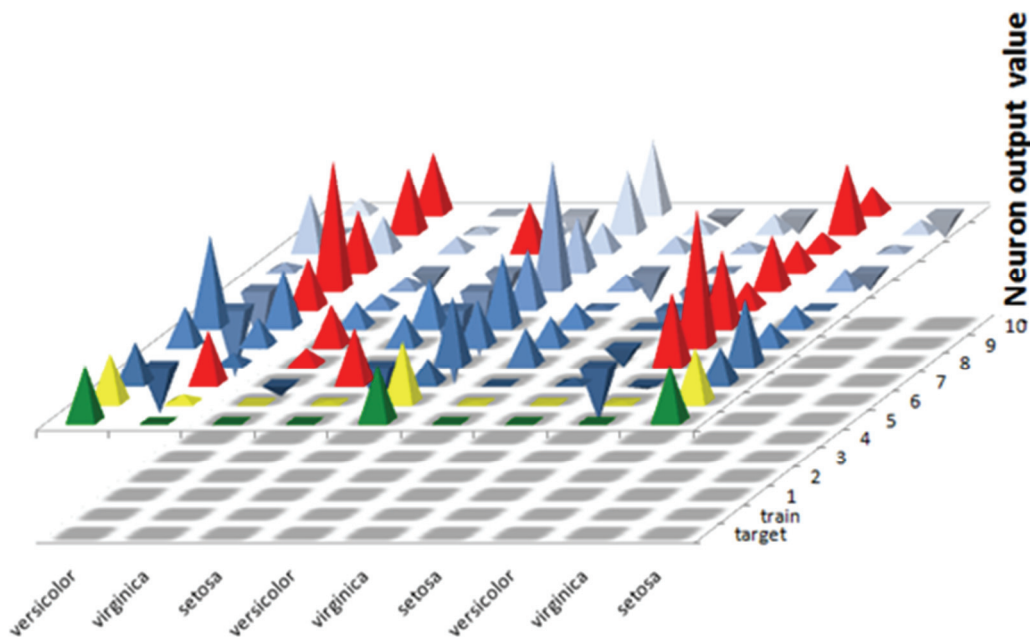
Fig. 17. *Behaviour of 10 networks with damages at the level of 48% of their elements.*

As it occurred only once per 10 experiments conducted, it may be assumed that in this case we deal with "an accident at work". Apparently one process of accidental damaging of network elements h a s led to the elimination of such connections that are really very crucial and needed.

A definitely new quality appeared in case of 10% of damages.

In this case, the network begins to err quite often, sometimes erroneous solutions (marked in figure 16 with a different colour) occur one after

The percentage of errors generated by the network increases as the number of damages increases, until finally (at 48% of damaged network elements) the network produces erroneous solutions more often than correct recognitions (figure 17).

The usefulness of a network as a recognising device is problematic from this point on but it is worth noticing that it occurs when almost 50% of network elements have been damaged, so it is not surprising that the network errs.

Out of curiosity, the experiment was continued for an even higher percentage of damaged connections in the network. It turned out that further change of the percentage of erroneous recognitions was not a constant increase and even when 90% of network elements were damaged, the network still tried to perform its task, achieving quite good results, as for such a damaged structure. Figure 18 presents the overall experiment results.

7. CONCLUSIONS

Studies described in this article are piloting studies. They are based on a single example of a neural network and one task solved by this network, so over-generalisation of conclusions that may be drawn from these studies would be unjustified. However, it seems that several observations are worth noticing and they should become the subject of further, more detailed studies.

First of all, it was proven that neural networks might show an increased tolerance to damages. In the conducted studies, the trained neural network retained the ability to perform the learnt function properly (recognition of selected elements) despite damages that were artificially introduced into its structure.

Secondly, the hypothesis according to which the size of network damage grows while its ability to perform the learnt task decreases has been confirmed. However, this correlation seems to be non-linear. In the case of a small extent of damages (several percent of the total number of elements being a part of the network), the impairment of function is slight. One may say that in this region the network defends itself against the results of damages and shows almost unchanged functioning despite visible damages in its structure.

After exceeding the first threshold of the number of damages tolerated, the quality of network functioning rapidly decreases, but then on a given segment (in the studies discussed it was the range from 10% to 40% of damaged elements) the network performs its tasks with an error that is much bigger than in the case of an undamaged network or a slightly damaged network (ca. 30%). Yet still it is within the tolerance limit for practical uses. Let us point to the fact that in the analysed task one of the three iris species had to be recognised, so the level of errors expected for the system that would be totally unable to recognise and would make only random guesses would be expressed as a per-

centage of 66% of erroneous recognitions. A significantly damaged network (40% of connections damaged) reached the rate of errors at the level of 33%, that is twice smaller.

Upon exceeding the second threshold (ca. 50% of damaged elements), the network again, almost abruptly, significantly deteriorates its behaviour, reaching the 60% level of errors, whereby this error practically stops increasing with further damage to the network and remains at the given level even when up to 90% of elements are damaged. It is worth noticing that this level still remains slightly lower than for a classifier functioning randomly, which may prove that even this extremely damaged neural network still tries to solve problems in a reasonable way.

Finally, it was noticed that the process of deteriorating the quality of the network functioning caused by damages of its elements was more visible in relation to tasks performed by the network than in relation to easier tasks. There might be exceptions from this rule stemming from the fact that random damages of the network structure may selectively destroy some of its properties and abilities, leaving the other network "competences" intact.

The phenomenon of tolerance to damages observed in artificial neural networks was signalled in this article but it was not fully explained and studied. However, with the awareness that this phenomenon really exists, it would be worth undertaking further studies to make our knowledge on this subject more extensive, detailed and better documented.

REFERENCES

Graupe, D., 2007, *Principles of Artificial Neural Networks*, World Scientific Publishing Co., Hackensack.
Longstaff, A., 2002, Neurobiology. Short courses [In Polish: *Krótkie wykłady, Neurobiologia*], Warszawa: Wydawnictwo Naukowe PWN, Warszawa (In Polish).
Tadeusiewicz, R., 2010a, New Trends in Neurocybernetics, *Computer Methods in Materials Science*, 10/1, 1-7.
Tadeusiewicz, 2010b, R., Cybernetic Modeling of Neural System – From Neuron-Like Structures, through Neural Networks till Neurocybernetics [In Polish: Modelowanie cybernetyczne systemu nerwowego – od struktur neuropodobnych, przez sieci neuronowe aż do neurocybernetyki], *Bio-Algorythms and Med-Systems*, Suplement No. 1, 6/12, 15-17.

ZJAWISKO ODPORNOŚCI NA USZKODZENIA SZTUCZNYCH SIECI NEURONOWYCH

Streszczenie

Sieci neuronowe są narzędziami komputerowymi posiadającymi szereg zalet. Mają szerokie możliwości w zakresie modelowania nieliniowych systemów oraz są przyjazne dla użytkownika ze względu na zdolność do uczenia się. Oprócz wymienionych, sieci neuronowe mają jeszcze wiele dodatkowych zalet. Jedną z nich jest zwiększona odporność na uszkodzenia. Systemy technologiczne z reguły nie są odporne na uszkodzenia. Mikroprocesor jest niezdolny do pracy nawet po niewielkim uszkodzeniu. W przeciwieństwie do tych systemów, systemy biologiczne są w dużej mierze niewrażliwe na uszkodzenia. Specjalnym przykładem jest mózg – wiele komórek może być martwych lub źle funkcjonować, podczas gdy cały mózg funkcjonuje prawidłowo. Badania opisane w niniejszej pracy są dedykowane do oceny czy sieć neuronowa, która jest modelem naturalnego systemu nerwowego, może również wykazywać odporność na uszkodzenia. To założenie zostało udowodnione doświadczalnie, a wyniki symulacji potwierdzających taką odporność są zamieszczone w artykule.

COMPUTER METHODS IN MATERIALS SCIENCE