FPGA Implementation of Artificial Neural Networks: An application on Medical Expert Systems

G. - P. K. Economou, E. P. Mariatos, N. M. Economopoulos,
D. Lymberopoulos, and C. E. Goutis

Department of Electrical Engineering
University of Patras, GR 261 10, Patras, Greece

Abstract

In this paper, the FPGA implementation of an Artificial Neural Networks (ANNs) composition for a Medical Expert System (MES) focused on Pulmonary Diseases (PDs) is discussed. Using a specially designed Neuron based on pipelined bit-serial arithmetic and a successful approximation of its determinant sigmoid function, a computation module has been structured that can accommodate eight (8) Neurons in one FPGA. The use of memory elements allows for up to 256K synapses to be mapped with high speed and great accuracy performances. Also, due to the FPGA reconfigurability, new structures and training patterns can be used to update this Medical Expert System, in order to fit in more Pulmonary or other Diseases, with minimal effort.

Keywords

Medical Expert Systems, Artificial Neural Networks, Pulmonary Diseases, FPGA.

1 Introduction

Artificial Neural Networks (ANNs) are recently playing a significant role towards improved MES, due to their special qualities [1], [2], [3], [4], [5]. A new ANNs' composition for the Diagnosis of PDs is presented in this paper. It has shown very good performance while being taught and tested with real-world data. It is based on a cascade of three layers of ANNs that process PDs' symptoms and other input data, and provide as outputs the possible PDs with an accuracy of 92%. Furthermore, the third layer of the structure, that is currently being designed, will suggest possible medical treatment and medicines according to the results of previous layers.

The proposed implementation is based on Field Programmable Gate Arrays (FPGAs), in order to be portable and easily reconfigurable. The latter quality is an important factor, since it allows for new structures (upgrading the current MES, supporting other diseases, mapping into other fields of human expertise, etc.) to be implemented with no changes in the given hardware.

The key issue when designing an FPGA for ANNs, is to reduce the size of each Neuron in order to fit as many as possible in one chip. A special Neuron circuit based on bit-serial pipelined arithmetic and on a linear but effective approximation of its sigmoid function has been designed. This way, eight (8) Neurons can fit in one FPGA chip. The use of large external memory allows the computation of up to 256K synapses, with speed that overcomes that of fast conventional processors.

Section 2 of this paper refers to the introduced ANNs' composition; its development and training is discussed and its structure is presented. Section 3 briefly explains the need for an FPGA implementation, and is followed by a presentation of the designed Neuron element in section 4. The issue of mapping the whole ANNs' formation on an FPGA-based system is discussed in section 5 and the paper ends with a reference to the procedure of updating the posed MES (section 6).

2 The Medical Expert System

Medical patients' data constitute this MES's input patterns. Yet, due to the single weight and the rendering that different Medical Doctors (MDs) assign, it is often a complex task to transmute them in usable information. Also, in order for MDs to properly use an MES, it must follow step by step the Clinical Differential Diagnosis Methodology (CDDM), whereas intermediate results have to be made accessible in each step of induction.

0-8186-6710-9/94 $4.00 © 1994 IEEE

Thus, the formation of an MES based on an ANNs composition, has been forwarded to provide for the categorization and generalization of the medical data-input patterns into new patients cases' symptoms. A mapping of their symptoms' exhibition to the classes of and to possible PDs, is therefore achieved.

The boundaries of the system were established by medical experts in PDs. A definite number of inputs were set, i.e., the questions that MDs ask when inspecting a patient. They contain related findings of each one of PDs' symptoms, i.e. Cough, Sputum, Haemoptysis, Fever, Dyspnea, Wheezing and Chest Pain and historical as well as data obtained from physical examinations. Moreover, those data were fed to a large number of ANNs [7], [9], [13], [15] and related to both a sum of thirty-five (35) PDs and to twelve (12) major PDs' classes. Data were fed by introducing their existence or non-existence in possible PDs' symptoms. Major influences, as the gravity of findings to determine certain PDs, multiple PDs' interference in a diagnosis and resulted PDs' ordering on a higher-fitness basis, were left to the ANNs to learn. Still, lethal PDs a patient could suffer, were made certain to be excluded or confirmed by this MES, by using suitable input patterns, by a percentage of accuracy that approximates the 92%.

This MES formation, in order to follow the CDDM, is composed of three layers. Each layer is structured by a number of three-levelled ANNs [6], [7], [8], [9] of different number of input and hidden, but of the same number of output Neurons per layer. Data connections from the previous to the next ANNs are provided, both internally and externally to these layers, thus processing knowledge from more general to more specific. A large number of experiments lead to this particular newly proposed MES's formation, based on these ANNs' architecture, which are pictured in Figure 1.

In the first MES's layer, real-world patient's data are treated in order to define possible PDs' general classes. ANNs covering the PDs' most important symptoms, i.e., Cough, Sputum, Haemoptysis, Fever, Dyspnea, Wheezing and Chest Pain, as well Historical and Physical Examinations' data, and their related findings, are fed with a patient's clinical data which are then processed parallel in time. Hence, the outputs of these ANNs, form pairs along to the outputs of Physical Examinations' ANN so as for their outputs to be fed in the next ANNs of the chain. These ANNs too output PDs' possible general classes, granting more reliability to all input data, a patient's answers and MDs' inspection results. Consequently, the final ANN, that suggests possible Clinical Examinations (CEs) to be performed, concludes the first layer of this MES.

The same number of quasi-identical to the first layer's ANNs, form the second one, plus a number of inputs to all of them: the outputs of the possible PDs' general classes from the previous layer. In addition, the outputs of these ANNs are possible PDs. This way, a strong positive feedback is exercised to each second layer's ANNs and intermediate results are made clear.

The final layer consists of two (2) similar structured to the second one ANNs, that process PDs' old and new symptoms' findings. Additional inputs, however, are provided from CEs and the previous layers' results and are correlated all together. The outputs of this layer are planned to be the recommended medical treatment, appropriate nutrition, possible medicine(s)' dosage(s), the proper way that these should be administrated (Per Oral - po, Intra Venus - iv, Intra muscular - im) and the proper time-schedule those should be taken in a given period of time. Figure 2 depicts the ANNs' composition for the first layer of the proposed MES. More detailed information can be found in [10], [11].

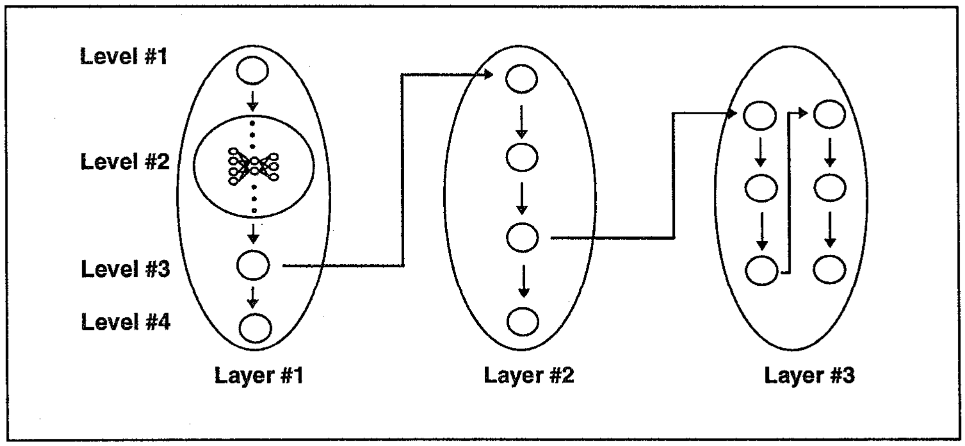

Figure 1: MES's Layers and ANNs' Composition.

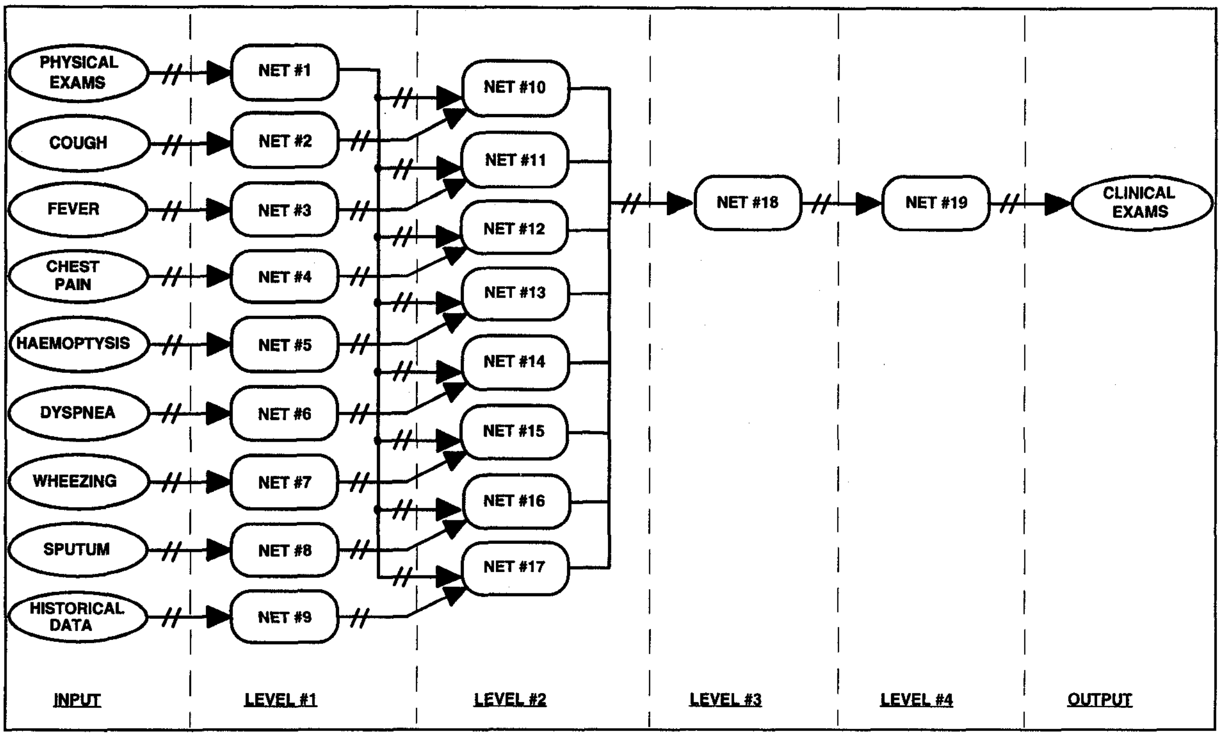

Figure 2: Structure of Layer #1.

3 The need for an FPGA-based implementation

This proposed MES has been implemented in software, as well as taught and tested with real-world data. Double precision arithmetic was used and a sigmoid function for each of the Neurons of the form:

$$f(x)=\frac{1-e^{-\alpha x}}{1+e^{-\alpha x}},\ \alpha\text{: slope of the sigmoid} \qquad (1)$$

Yet, a disadvantage of the software solution, is low speed. Still, this is not a major problem, at least if a powerful computer is used to run the software. A few mseconds delay can not be considered important for a full medical diagnosis. This could be an actual drawback of the software approach, only if even more complex systems, that could probably cover many different infected parts of the human body, are to be constructed. Such an MES, to cover blood diseases, is already set.

The main reason for using a hardware-based implementation, is the need for a portable diagnoses MES. Medicine is not always practised behind an office's desk, in fact a great part of it is practised at patients' homes, at long distances from medical centres.

The main disadvantages of a hardware, ASIC-based, implementation, is the high development cost and the low reconfigurability it allows for. The FPGA solution ensures that new structures for the proposed ANNs can be mapped on the system without having to make costly changes in the hardware. In this paper an FPGA-based system is presented; it can be used to map the previously described ANNs, easily assuring this attribute.

Two issues appear in the process of translating a software program to a digital hardware structure. Firstly, the size of the hardware has to be dealt with, since the proposed structure require a large number of Neurons and synapses, which corresponds to a large number of adders and multipliers. Secondly, the accuracy of the arithmetic used affects the quality of the results. The data width must be selected to be as small as possible but without affecting the ANNs' performance. Another issue related to accuracy, is the approximation that will by used for the sigmoid function. As it can be seen from equation 1, it cannot be implemented in hardware without taking too much area. An approximation is proposed that has worked well in many simulations, and requires very little space, all the proposed structures and circuits, have been emulated by proper programs written in the C programming language, by a, 386@40 platform. They have not been actually placed on an FPGA. Though, estimations of their performance can be given.

Thus, the results that are presented in the following sections are obtained by information taken from the FPGA's data sheets and the team's experience with FPGAs. Next section deals with the size and accuracy issues, and section 5 presents the mapping of the ANNs.

4 Design of the Neuron

The processing element of an ANN is the Neuron. A Neuron can be viewed as processing data in three steps; the weighting of its input values, the summation of them all and their filtering by a sigmoid function. The Neuron can be expressed by an equation of the form [12]:

$$y_j = f\left(\sum_i w_{ij} x_i - \theta_j\right) \tag{2}$$

The summation can be calculated, either by a parallel input Wallace tree, or by a serial accumulation. For the weighted inputs to be calculated in parallel using conventional design techniques, a large number of multiplier units would be required. To avoid this, a Multiplier/Accumulator architecture has been selected. It takes the inputs serially, multiplies them with the corresponding weight and accumulates their sum in a register. Figure 3 shows the proposed Neuron design.

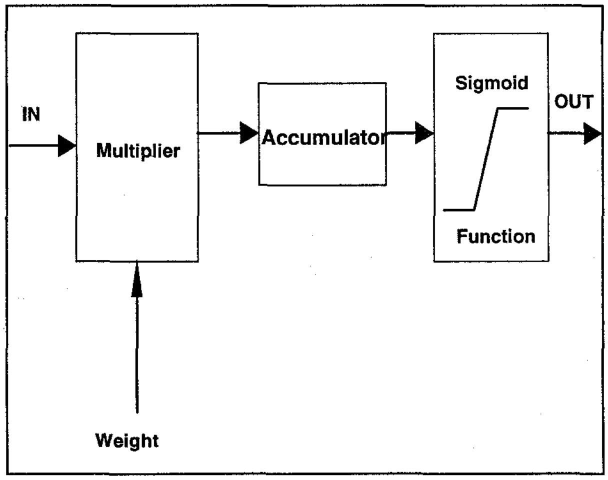

Figure 3: The proposed Neuron's architecture.

The Accumulator's unit is composed of a bit-serial adder and an eight (8) bit register. The multiplier is based on a simple architecture that consists of eight (8) shift elements, some simple logic gates and one (1) Wallace tree of adders. It works with bit-serial data, in order to save size, and it is presented in figure 4.

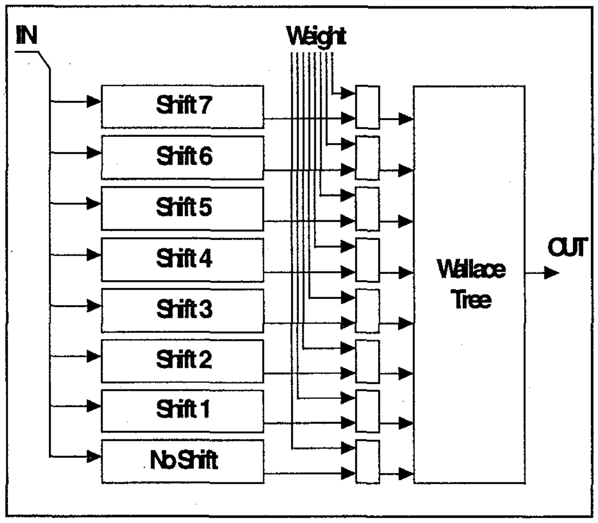

Figure 4: Multiplier's Circuit.

Pipelining is used to avoid speed reduction that might be caused by bit-serial arithmetic. Since the Most Significant bits of the shifted values are always constant "0" or "1" and equal to the MS bit of the original number, they cannot be used. Thus, shifting of the next input value can start before multiplication of the previous one has finished. This multiplier is estimated to occupy about twenty-five (25) Configurable Logic Blocks (CLBs) of the XILINX XC3090 FPGA [13].

The width of data has been set to eight (8) bits, a value obtained from simulations with software. A seven (7) bit accuracy might also be adequate for the specific application but the use of eight (8) bits has been decided in order to be compatible with possible future variations of the ANNs. The sigmoid of equation 1 has been approached with a function that is linear for some values and reaches zero or one for all the others. Figure 5, shows a circuit that implements this sigmoid and the comparison of the two functions. The factor 0,0625 has been selected though extensive simulations with a lot of different values, and since it is a negative power of 2, can be implemented by a simple right shift.

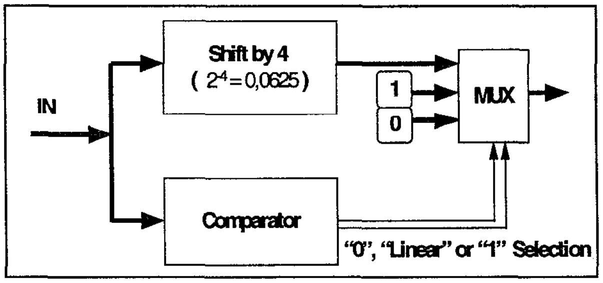

Figure 5: The Sigmoid's Approximator.

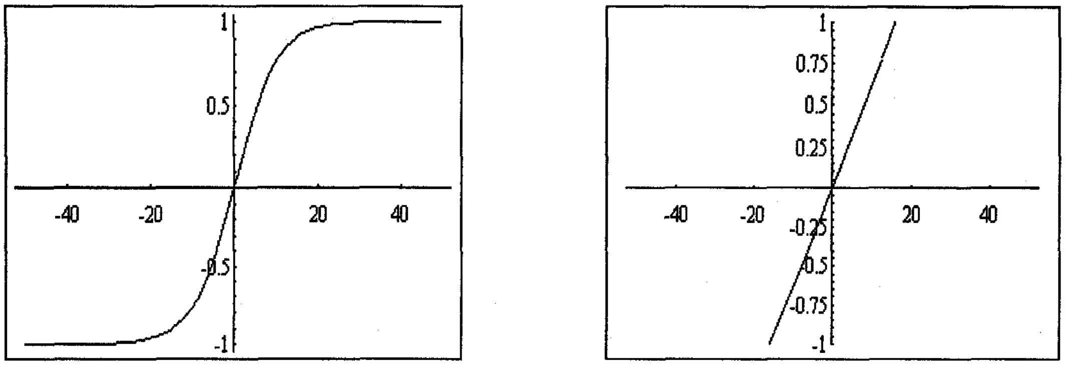

Figure 5 (cont.): The Sigmoid's Approximator.

5 Mapping of the ANNs' in an FPGA

The mapping of the proposed structure on an FPGA module, has to be based on a trade-off between speed and chip size. Since the total number of Neurons is approximately 8,000 connected with 220,000 synapses, a scheme for serial calculation of each ANN has to be used. Each ANN, consists of three slabs (Neural layers). This architecture deals with the implementation of one slab; also, it is based on an eight (8) Neuron processing module and on large external memories.

One FPGA (XILINX XC3090) chip will contain eight (8) Neurons. A parallel-in serial-out register is used to carry the inputs from an external memory to the synapses, whereas a multiplexer (MUX) module addresses the outputs and stores them in the same memory to be used in the next ANN [14]. Figure 6 shows the proposed implementation.

A control module that monitors the execution of the entire structure (not shown), can also be integrated in the FPGA. Yet, to allow for possible increase in the complexity (and thus in the size) of the control module, the use of a second FPGA chip could be considered. Still, this approach bears no further complexity.

The MES will be composed by the next items: the FPGA, the RAM for intermediate store of ANNs' outputs and the memory that contains the weights. The proposed system structure is shown in figure 7. Two more elements are there, the input MUX, that is used to select inputs for the FPGA module either from previous results or from the user, and the configuration's RAM that holds the configuration bit stream for the FPGA.

The weights and the configuration's RAM are loaded externally. This allows for the easy update of the system as will be discussed in the next section. The sizes of the memories, depend on the complexity of the MES and the total and maximum number of the ANNs' synapses.

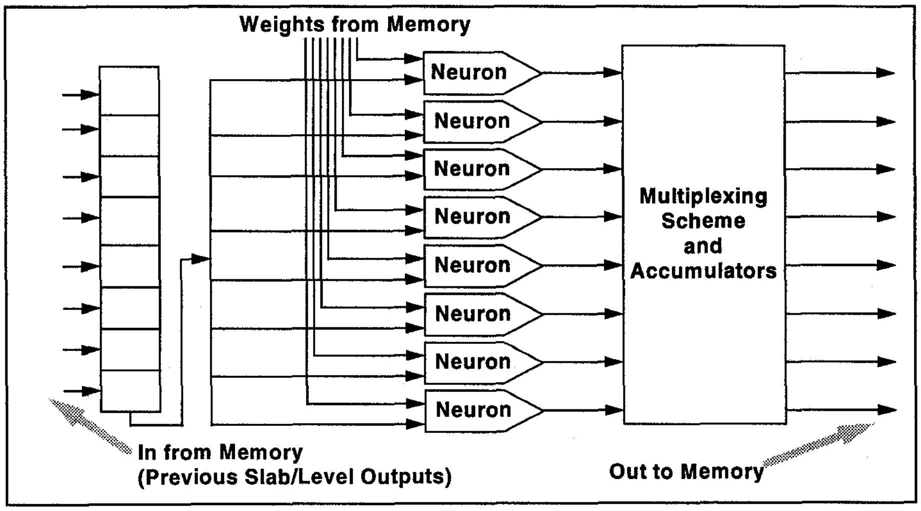

Figure 6: FPGA's Contents.

More specifically, the Weights Memory contains one byte for each synapse; the use of a 256 Kbyte RAM to hold all the 220,000 weights of this structure, is thus proposed. The size of the Intermediate RAM is proportional to the maximum possible number of inputs for one slab per ANN. This size is 290 inputs for this application. Hence, this RAM's size results in 2 Kbytes.

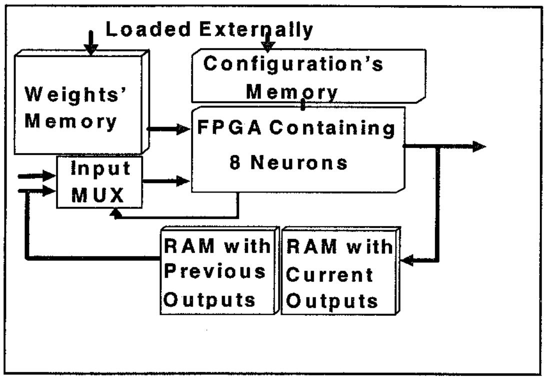

Figure 7: System Modules' Connections.

As already mentioned, speed performance may not be the key factor, but is still important. For each slab with N Neurons and M inputs (or M synapses/Neuron), the time required is $N \times M/8$ operation cycles. On the other hand, with adequate pipelining, each operation cycle will be performed at the time of 2 clock cycles. The total time for one ANN slab will then be $N \times M/4$ clock cycles.

Using a clock at 20MHz, the full MES will require about 3mseconds. A powerful PC (386@40) has already been used and implements about 0.1 seconds.

6 Training and Reconfiguring the MES

Updating the system can be considered mandatory, mainly for two reasons. Firstly, should any new data be fed into the training algorithm causing new weights to show a better performance, and secondly, should a new structure be developed. A powerful computer can be used at a medical centre, that will constantly train the network (MES) with new real-world data. The outcome of this process can be stored in a Weights File.

The development of a new structure that will either enhance the existing MES or add new capabilities to it (consider more symptoms, etc.), should pass through a phase of software simulation. When the results show that it can be used in practical situations, a designer will have to make an FPGA prototype (using the Tools of XILINX) and prepare a Configuration Bit stream File.

Both Files can be then loaded on the system via magnetic media or even through systems that communicate by using eMAIL communication protocols. Figure 8 shows the full presented updating procedure. Note the simplicity, high speed and portability of the approach, that leaves MDs out of having to deal with the technical aspects of this MES.

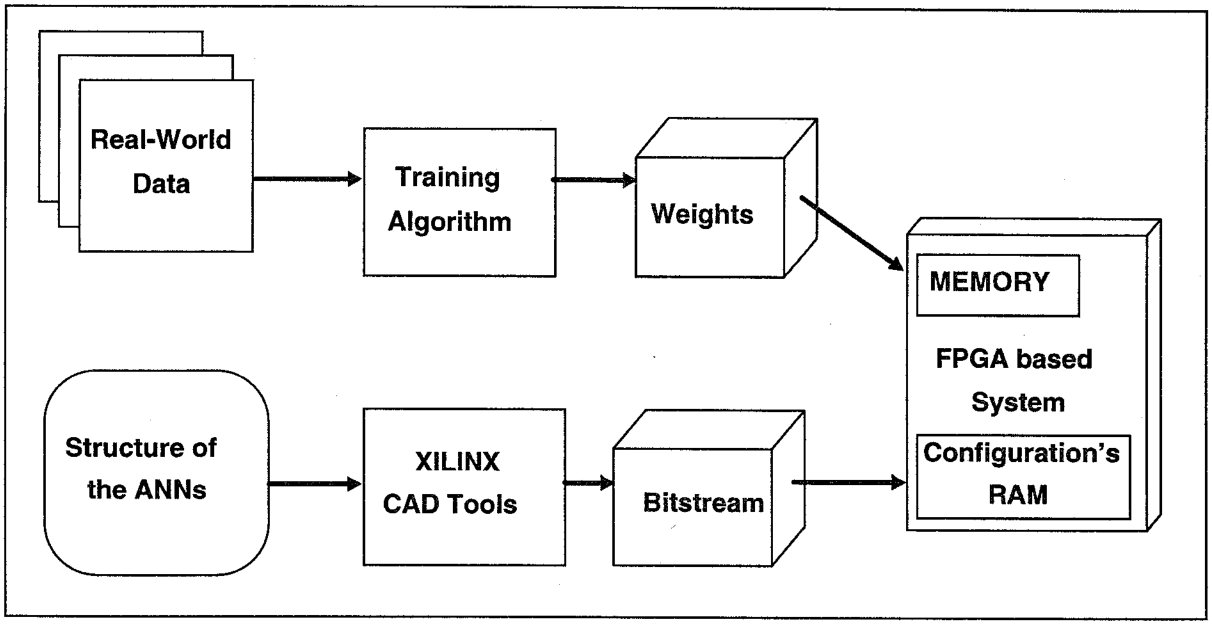

Figure 8: MES's Updating Procedure.

8 Conclusions

The proposed mapping of an ANNs' composition in an FPGA-based system, was proven to provide for very prominent results. This composition was used for the structuring of an efficient portable and reconfigurable MES to assist MDs in PDs' diagnoses. Software and hardware implementations were studied and by simulation results size and speed factors proved to be overcome. On the other hand, the enhancement of this MES through the augmentation of its patterned input data, will have to follow. The integration of medical theoretic knowledge and the interference with other pulmonary teams' knowledge as well as to the learning by new algorithms, would be the next target, easily transferred on the proposed FPGA-structured basis. The final realisation of a general-purpose MES to be the basis of inductively diagnosing other medical diseases, is the outmost scope of this research team.

References

[1] D. G. Bounds, P. J. Lloyd, B. Matthew, and G. Waddell, "A Multi Layer Perceptron Network for the Diagnosis of Low Back Pain", *Proc. Int. Conf. on Neural Networks*, San Diego, Vol. 2, pp. 481-489, 1988.

[2] A. Durg, W. V. Stoecker, J. P. Cookson, S. E. Umbaugh, and R. H. Moss, "Identification of Variegating Coloring in Skin Tumors: Neural Network vs. Rule-Based Induction Methods", *IEEE Eng. in Med. and Biol.*, Vol. 12, pp. 71-74 & 98, 1993.

[3] B. H. Mulsant, "A Neural Network as an Approach to Clinical Diagnosis", *M. D. Computing*, Vol. 7, pp. 25-36, 1990.

[4] T. J. O' Leary, U. V. Mikel, and R. L. Becker, "Computer-Assisted Image Interpretation: Use of a Neural Network to Differentiate Tubular Carcinoma from Sclerosing Adenosis", *Modern Pathology*, Vol. 5, pp. 402-405, 1992.

[5] R. Poli, S. Cagnoni, R. Livi, G. Coppini, and G. Valli, "An NN Expert System for Diagnosing and Treating Hypertension", *IEEE Comp.*, Vol. 24, pp. 64-71, 1991.

[6] D. R. Hush, and B. G. Horne, "Progress in Supervised Neural Networks", *IEEE Sig. Proc. Mag.*, Vol. 10, pp. 8-39, 1993.

[7] R. P. Lippmann, "An Introduction to Computing with Neural Nets", *IEEE ASSP Mag.*, pp. 4-22, 1987.

[8] R. S. Scalero, and N. Tepedelenlioglu, "A Fast New Algorithm for Training Feedforward NN", *IEEE Trans. on Sig. Proc.*, Vol. 40, pp. 202-210, 1992.

[9] B. Widrow, and M. A. Lehr, "30 years of Adaptive Neural Networks: Perceptron, Madaline, and Backpropagation", *Proc. of the IEEE*, Vol. 78, pp. 1415-1442, 1990.

[10] G. - P. K. Economou, C. Spiropoulos, N. M. Economopoulos, N. Charokopos, D. Lymberopoulos, M. Spiliopoulou, E. Haralambopulu, and C. E. Goutis, "Medical Decision Making Systems in Pulmonology: A Creative Environment based on Artificial Neural Networks", *1994 IEEE Int. Conf. on Systems, Man and Cybernetics.*

[11] G. - P. K. Economou, C. Spiropoulos, N. M. Economopoulos, N. Charokopos, D. Lymberopoulos, M. Spiliopoulou, E. Haralambopulu, and C. E. Goutis, "Medical Diagnosis and Artificial Neural Networks: A Medical Expert System applied to Pulmonary Diseases", *1994 IEEE Workshop on Neural Networks for Sig. Proc.*.

[12] F. N. Sibai, "A Fault Tolerant Digital Artificial Neuron", *IEEE Des. & Test of Comp.*, pp. 76-82, 1993.

[13] XILINX Inc., *The Programmable Gate Array Data Book*, 1991.

[14] F. Distante, M. Sami, R. Stefanelli, and G. Storti-Gajani, "Mapping Neural Nets onto a Massively Parallel Architecture: A Defect-Tolerance Solution", *Proc. of the IEEE*, Vol. 79, pp. 444-460, 1991.