Volume 2(1): 2024, Page20-28, 10.59785/tjhest.v2i1.42

Application of approximation algorithms to the detection and categorization of diseases

Heyam Al-Aaraj[1*], Ahmad Al-Sayeh[2] and Aida Mohammad[3,]

[1] Al-Hussein Bin Talal University, Princess Aisha Bint Al Hussein College for Nursing and Health Sciences, Jordan; **hiyam_aaraj@yahoo.com**
[2] Al-Hussein Bin Talal University, Princess Aisha Bint Al Hussein College for Nursing and Health Sciences, Jordan**; a.sayeh@ahu.edu.jo**
[3] King Faisal University, College of Education, kingdom of Saudi Arabia. College of Education**;amohammad@kfu.edu.sa**
***** Correspondence: **hiyam_aaraj@yahoo.com**

Abstract: The diagnosis of diseases can be formulated as a classification problem, making it an NP-hard problem. This is the case for the two problems that this work aims to solve: the classification of tumor samples from patients suspected of having breast cancer as benign or malignant and the classification of samples from patients suspected of having type II diabetes as negative or positive. In order to make accurate diagnoses (classification) of these disorders, our idea is to construct approximate algorithms based on multilayer perceptrons, genetic algorithms, and algorithms that hybridize these alternatives. Numerical experiments enable assessing and contrasting the effectiveness of various approaches using actual data sets. The results demonstrate that, in addition to outperforming algorithms suggested in the literature in terms of performance, our ideas produce outcomes with classification errors that are nearly zero.

Keywords: Classification, medical diagnosis, multilayer perceptron, genetic algorithms, hybrid algorithms.

1. Introduction

The medical diagnosis is one of the most complex tasks to carry out since it requires the analysis of many factors such as Anamnesis (this is the information provided by the patient during the clinical interview only useful to analyse your clinical situation), symptoms, signs, physical examination [1]. The analysis of these factors and the doctor's experience results in a diagnosis that can be presumptive or definitive and, in certain cases must be confirmed by medical tests to reach a diagnosis.

In the last decade, the trend of using Machine Learning Algorithms has become popular, offering many possibilities for autonomous decision-making. The areas of health have a remarkable evolution despite the fact that they are very special areas due to the complexity of dealing with the integrity of a human being, emphasizing that any error can be fatal for the health of the individual, being able to until it cost him his life.[2]

Currently, it is possible to model complex problems by means of approximate algorithms, modeling the behavior of complex systems, taking into account the particular characteristics and random factors that define them. Such is the case of problems belonging to the field of Biomedicine [3], in particular, the diagnosis of diseases that can be formulated as a classification problem, given that it is necessary to examine the medical data (clinical

tests, physical examination, symptoms, resonance, tomography, etc.), or attributes, to determine disease, or class. This makes medical diagnosis an NP-hard problem, as is the case with the two problems that are attempted to be solved in this work: classification, as benign or malignant, of tumor samples from patients suspected of suffering from breast cancer; and classification, negative or positive, of samples from patients suspected of having type II diabetes. [4]

Artificial Neural Networks (ANNs), Metaheuristics and hybrid algorithms have become efficient tools to solve this class of problems [5, 6]. Due to the massively parallel and distributed structure, ANNs have a high capacity for learning, generalization and nonlinear approximation and are therefore suitable for classification [7, 8]. On the other hand, Metaheuristics represent a family of approximate optimization techniques that maintain the balance between intensification and diversification during the search. This allows complex and difficult problems to be solved efficiently, as is the case with classification problems [8]. Also, hybridizing these techniques has provided high-quality solutions to these problems [10]. This work intends to conduct a study on ANNs, Metaheuristics and hybrid algorithms, analysing the characteristics, advantages and disadvantages of their use; to then use and adapt these techniques to the classification and diagnosis of potential cases of breast cancer and type II diabetes. ANNs based on the multilayer perceptron, called $BackP_{base}$ and $BackP_{mod}$, two metaheuristic algorithms based on genetic algorithms, GA_{1x} and GA_{2x}, and a hybrid algorithm that combines these techniques, GA_{BackP}, is proposed to provide reliable diagnoses.[11]

The article is organized as follows. In Section 2, the algorithms that deal with the problems in question in the literature are introduced, and the ones proposed here are also explained. Whereas section 3 specifies how the experimentation was carried out, the data's origin and the results obtained are analysed. The last section summarizes our conclusions and future work is projected.

2. Approximate algorithms used in classification problems

This section describes approximate algorithms that solve classification problems (eg, disease diagnosis), such as multilayer perceptron, genetic and hybrid algorithms. In addition, each of these descriptions will be accompanied by an explanation of our algorithmic proposal.

2.1. Multi-Layer Perceptron (MLP)

An MLP utilizes a feed-forward network with supervised training to learn from examples [12]. A perceptron's main learning method is backpropagation (reverse error propagation), a generalisation of the Delta rule [13]. While not perfect, this strategy has advanced the discussion of MLPs' potential in a wider range of contexts

The effects of an input pattern are propagated through the network's hidden layers after it is provided as a stimulus to the network's first layer of neurons (forward step). A linear combination of the inputs and their corresponding synaptic connections gives a weighted total, which is subsequently subjected to an activation function for the purpose of said propagation (or weights). Then, a cost function is utilised to compute an overall score based on how much of an error was made while comparing the neurons' output to the output pattern. The errors are then sent back to the intermediate layer neurons, who are directly responsible for contributing to the output. Repeat this process until each neuron in the network has been assigned an error that represents their share of the overall

network error. The weights of connections between neurons are adjusted using this knowledge to produce more precise outcomes when the same pattern is presented in the future. The learning process is accelerated by adding the momentum component to momentum [13], which accelerates the convergence of the algorithm. Another technique for improving learning is adaptive η (Learning rate) [13], which comprises altering the value of η during training to prevent the error from becoming stuck because local optima are present.

We propose a pair of MLPs, which we've named $BackP_{base}$ and $BackP_{mod}$, as part of the solution. We present $BackP_{base}$, a backpropagation learning platform. However, $BackP_{mod}$ enhances $BackP_{base}$ by using momentum and adaptive ' η ' techniques to regulate the learning rate and, by extension, the algorithm's convergence. The algorithms were developed using the guidelines provided by Hertz et al. [14]. In Figure 1 we see the structures of BackPbase and BackPmod, where n is the number of input neurons and m is the number of hidden layer neurons. The synapses are represented by the matrices.: W_{ij}^{1} (with i = 1, …, m and j = 1, …, n) and W_{1j}^{2} (with j = 1, …, m), as shown in Figure 2.

Because there are various numbers of qualities being studied in the two cases, the number of input neurons differs between them (8 for breast cancer and 9 for diabetes). Section 4 provides a description of the configuration procedure that determined the quantity of neurons in the hidden layer. One neuron in the output layer represents the network's ultimate classification choice for a particular instance.

For all MLPs, W_{ij}^{1} and W_{1j}^{2} are created using uniformly distributed random values in the [-0.5, 0.5] range. By changing the values of the parameters β, η, you may customise the learning error and the number of iterations. A constant named β is used to calculate the hyperbolic tangent activation function; it is the learning rate used to adjust the synaptic weights during the reverse iteration. The deployment of hyperbolic tangent, exponential, and linear activation functions are all being considered. We consider both the mean square error and the entropic measure when assessing the pricing [14].

The input-output patterns utilised to train the network represent 70% of the data in each case study. The evaluation stage makes use of all remaining resources. In order to fairly compare the results of the exponential and hyperbolic tangent activation functions, the patterns also need to be normalised. Both approaches also employ an incremental update [14] to lessen the possibility of the network forgetting its training material.

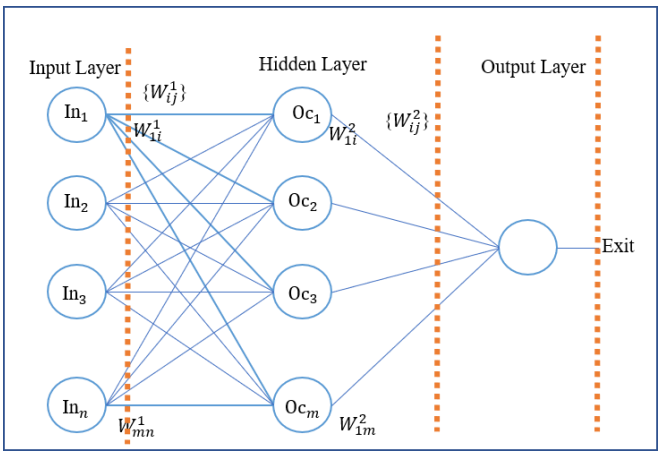

Figure 1. General architecture of the MLPs.

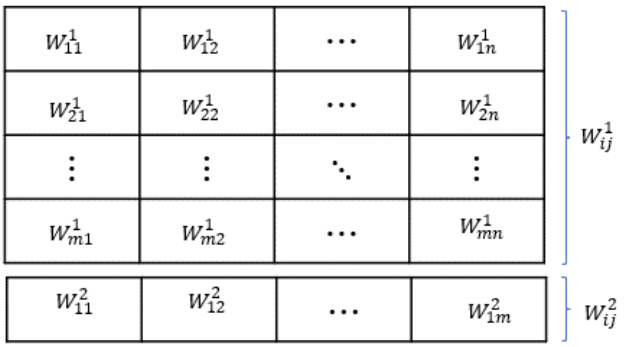

Figure 2. Synaptic weight matrices.

In particular, for $BackP_{mod}$, the constants *α, a, b* and *steps* are added. Both *α, a* and *b* are used to improve η in the *momentum* and adaptive η techniques; while steps are the number of iterations after which the value of η is verified.

2.2. Genetic Algorithms (GAs)

Since evolutionary algorithms are frequently employed to describe the progressive improvement of a population through Darwinian natural selection, GAs [15] are particularly well-liked as a subset of Metaheuristics. By maintaining a pool of candidates (individuals) and promoting both the creation of new knowledge (via mutation) and the distribution of current knowledge, they are technically engaging in a two-way search (by crossover).

Which individuals will be permitted to reproduce (through crossover and mutation) is determined by probabilistic selection operators [15]. (roulette, ranking, binary tournament etc.,).

Newly born people increasingly replace the existing population over time. This type of replacement aims to preserve the best individuals while eliminating everyone else (elitism). Depending on whether the existing population is taken into account, there are two different replacement strategies that can be used: (μ; □) if the replacement is carried out by selecting individuals from the union of the existing population and the population of children, and (μ + □) if it is carried out by using only the population of children [16].

Both of the proposed GAs, GA_{1X} and GA_{2X}, represent people with actual numbers, and they select their parents in a binary tournament. The elitist replacement procedure (μ + □) also uses this selection mechanism. Altering a randomly chosen gene (one component of the solution) is another method they employ to cause mutation in a person with a predetermined chance of success (pm).

The crossover approach used sets GA_{1X} and GA_{2X} apart from one another. The use of a crossover at a single point in time, as demonstrated by GA1X, is one technique to put a standard GA to the test. Similar to this, GA2X employs a two-point crossover to increase the genetic variety of each offspring. Both operators are employed with some assurance (p_c).

An individual's genetic makeup is based on the synaptic weights of an MLP, as shown in Figure 3. Individuals are represented by a vector in this case as opposed to MLPs where they are organised in a matrix. We produce random values in the range [-0.5, 0.5] for each connection weight, which represents the individual's gene, using a uniform distribution.

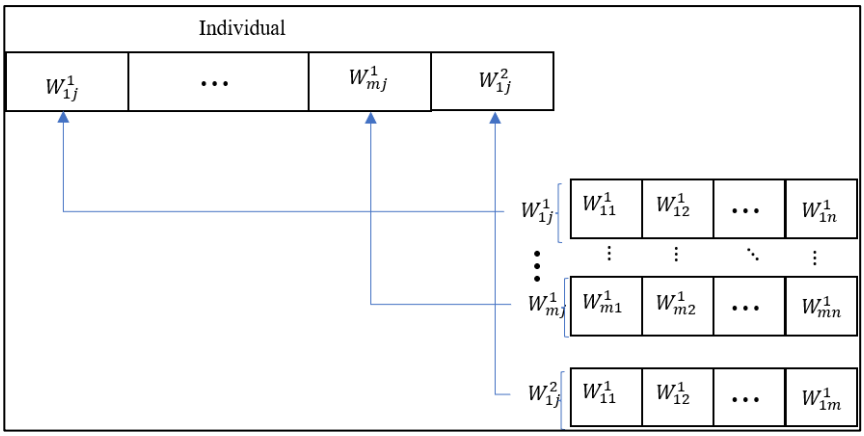

Figure 3. The usual composition of a GA population.

One measure of *fitness* is the typical number of incorrect diagnoses made when trying to classify instances of diabetes or breast cancer. The better a person's *fitness* level, the less mistakes they make. In other words, calculating the mean error by only running the forward step of the MLP throughout an epoch is akin to evaluating someone's *fitness*.

2.3. Hybrid Metaheuristics

Combining algorithms such as metaheuristics, mathematical programming, constraint programming, and machine learning techniques provide powerful search algorithms for solving NP-hard problems [17]. Our hybrid proposal, GA_{BackP}, combines metaheuristics with machine learning techniques by integrating the ease of adaptation of GAs with the machine learning present in MLPs. The GA provides the containing structure, and the MLP is used as the mutation operator.

Unlike those implemented by GAs, this new mutation operator does not apply a "blind" mutation but rather a "smart" mutation. This is because the MLP mutates the individual, trying to minimize its classification error. With this, it is hoped that the individual will increase its probability of survival during evolution.

The figure 4 shows the GA_{BackP} mutation process, which is achieved by entering the individual into the network and training it to refine the fitness of said individual. First, the individual is decomposed into two vectors: w_{ij}^1 and w_{ij}^2 . The vectors are restructured in the matrices that make up the synaptic connections of the network, then training of the MLP is carried out as explained in their design. Finally, the inverse work to the decomposition of the individual is carried out, converting the matrices into vectors and obtaining a mutated individual.

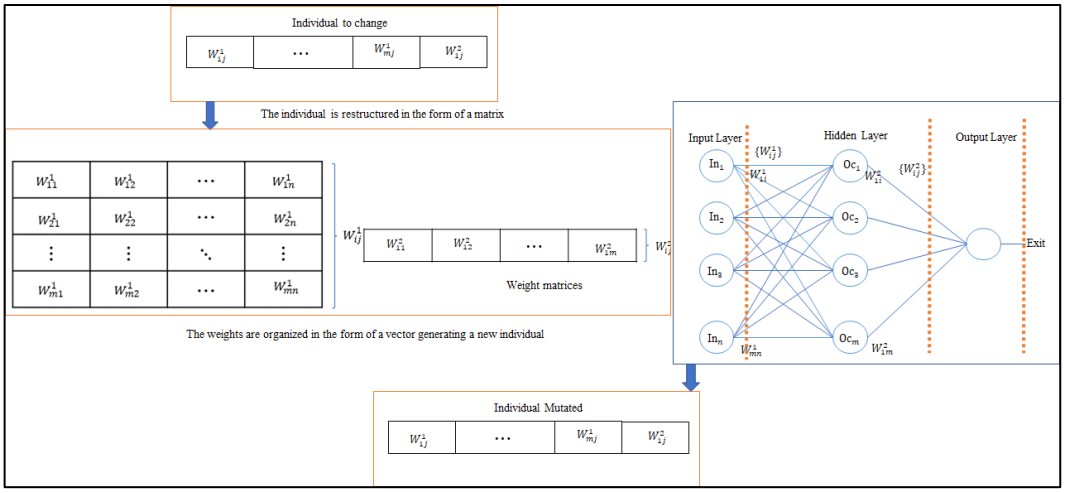

Figure 4. Mutation of an individual by means of RNA.

3. Experimentation and Analysis

This section explains the methodology used to carry out the experiments, and the results obtained are analysed and compared with those published by Alba and Chicano in [9]. The experiments use data from real cases stored in public repositories and are carried out on identical PCs equipped with: Intel Atom CPU N2600 processor, 1.6GHz, 2GB DDR3 RAM and Microsoft Windows 10 Ultimate x86.

In the case of breast cancer diagnosis, the data used belongs to the University of Wisconsin1 and consists of 699 instances composed of 8 attributes each. 65.5% (458) of these cases correspond to instances classified as benign and the rest as malignant. Meanwhile, the data used for the diagnosis of diabetes belong to the National Institute of Diabetes and Digestive and Kidney Diseases and correspond to patients belonging to the Pima aboriginal group. They consist of 500 instances with 9 attributes each, of which 325 (65.1%) are classified as negative, that is, non-diabetic patients and 175 (34.9%) are positive; that is, they correspond to diabetic patients.

Table 1. Parametric configuration for each algorithm and case study.

	$BackP_{base}$	$BackP_{mode}$	GA_{1X}	GA_{2X}	GA_{BackP}	$BackP_{base}$	$BackP_{mode}$	GA_{1X}	GA_{2X}	GA_{BackP}
Case study	Breast cancer					Type II Diabetes				
Act function.	TanH					TanH				
cost function	Mean Square Error					Mean Square Error				
β	1					1				
η	1					0.02				
n & m	7 and 2					8 and 49				
training times	500	500			15	500	500			20
Steps		15					15			
α		0.01					0.001			
A		0.001					0.05			
B		0.05					0.01			
Population Size			64					64		
p_c			0.98					0.98		
p_m			0.1					0.1		
Total generations			250	250	250			250	250	125

The parametric configuration used in the experiments (see table 1) arises from comparing the results obtained by giving each parameter different values and performing all

possible combinations of them3. For each algorithm and case study, the parameter configuration that offers the best results, in terms of mean training errors, is chosen. With the chosen configurations, each algorithm is executed 30 times to obtain a reliable sample and to perform a statistical analysis of the results to determine if there are significant differences between them and which ones differ. For the latter, the ANOVA test is applied if the data follow a normal distribution or the non-parametric Kruskal Wallis test otherwise; both with a confidence level of 95%.

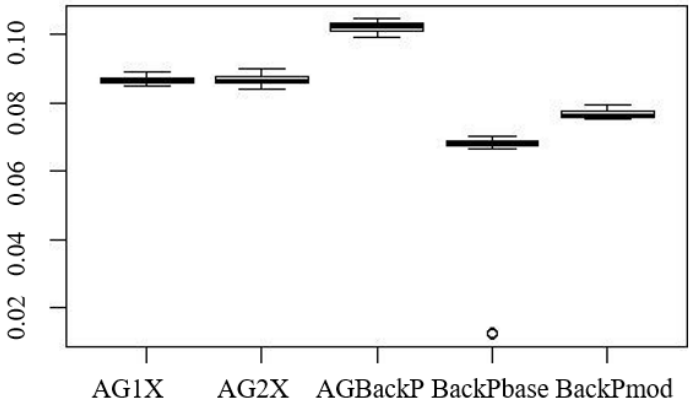

Figure 5. Box-plot diagram of each algorithm's mean square error obtained for the case study: Breast cancer.

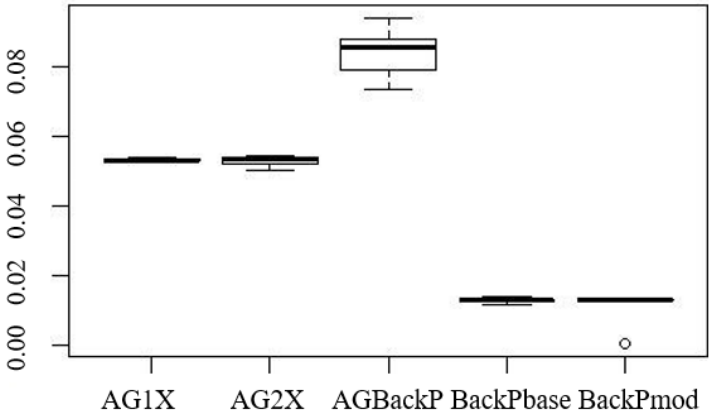

Figure 6. Box-plot diagram of the mean square error obtained by each algorithm for the case study: Type II diabetes.

The analysis of the results obtained in the experimentation can be deduced from the box-plot diagrams shown in figures 5 and 6. These graphs allow visualizing of the error made by each algorithm in all the executions and reflect the results of the analyses statistics. That is, whether or not there are significant differences in their behavior. On the one hand, we analyse the results related to the case study 'Breast cancer' shown in Figure 5. First, it is observed that the algorithms that classify with errors close to zero are $BackP_{base}$ and $BackP_{mod}$. In addition, there are no statistical differences in their behaviors, being $BackP_{base}$ the one that obtains the minimum error. Meanwhile, GA_{1X} and GA_{2X} classify with relatively low errors (approximately 5%) and present statistically similar behaviors, with GA_{2X} making the least error. Finally, and with errors ranging between 5 and 7%, we find GA_{BackP}, whose behavior is significantly different from that of the other 4 algorithms.

On the other hand, figure 6 allows us to analyse the results obtained for the study case 'Type II Diabetes. The five algorithms behave in a similar way as for the case of breast cancer, but this time the minimum error is greater than 6%.

In [9], the authors propose a GA, an MLP called BP and two algorithms that hybridize these techniques under the name GABP and GALM to solve the two case studies discussed here. The metric used for the comparison is known as Classification Error Percentage (CEP), which indicates the percentage misclassified by the algorithm in question. The average of these values for BackPbase, GA_{2X}, GA_{BackP}, BP[9], GA[9], GABP[9] and GALM[9], in each case study, is shown in Table 2.

$$CEP = \text{Mean Error} \times 100 \ldots \text{Eq. (1)}$$

Analyzing Table 2, it is observed that, in both study cases, BackPbase and GA_{2X} incur fewer classification errors than those made by BP and GA, respectively. This decrease in PEC values varies between 38 and 77%. However, in the two works, it is noted that the algorithms based on MLP improve the results obtained by those based on genetics between 20 and 94%.

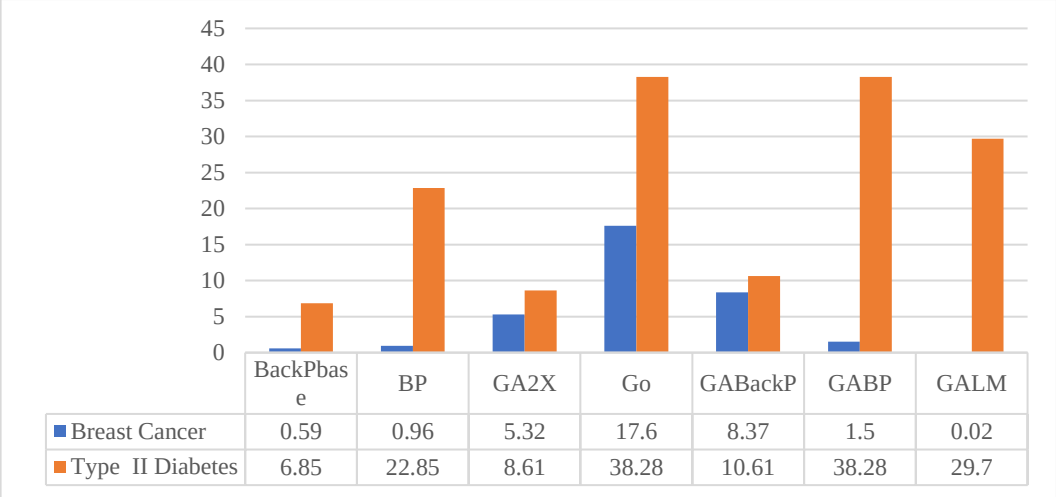

Table 2. Average of the PEC values for each algorithm and case study.

As for the hybrid algorithms, GA_{BackP}, GABP and GALM, significant differences in behavior are observed depending on the case study analyzed. For breast cancer, the PEC values corresponding to the two hybrid versions proposed in [9] are between 5 and 300 times lower than those of GA_{BackP}. Instead, for the second case, it is $GA_{\text{-}BackP}$ that incurs an error 2 and 3 times smaller. However, these hybrids do not outperform the MLPs.

4. Conclusions

In this article, two multilayer perceptrons, two genetic algorithms, and an algorithm that hybridizes the previous ones were developed to solve classification problems in medicine: diagnosis of cases of breast cancer and type II diabetes. These algorithms were evaluated using real test cases published by institutions recognized by the scientific community, and their behavior was compared with that of other algorithms published in the literature.

Once the experiments were carried out, it was observed that the algorithms based on MLPs are the ones that offer the best results, followed by the GAs, leaving the hybrid algorithm in last place. From the comparison with other algorithms in the literature, it

emerged that the algorithms proposed here obtained better results in five of the six comparisons made.

In future lines of research, we will first try to improve the performance of the GAs by incorporating heuristics in the genetic operators. Finally, and in order for the hybrid algorithm to achieve the expected results, improvements will be analyzed in the implementation of the mutation operator to accelerate convergence to the minimum error.

Author Contributions: Conceptu-alization, H.Al and A.Al; methodology, H.Al, A.Al,and A.M; validation, H.Al and A.M ; formal analysis, A.Al and A.M; data curation, H.Al, A.Al,and A.M; writing—original draft preparation, A.Al and A.M; writing—review and editing, H.Al and A.Al; visualization, H.Al and A.M; supervision, H.Al.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Conflicts of Interest: No potential conflict of interest was reported by the authors.

References

1. Davis JL, Murray JF. History and Physical Examination. Murray and Nadel's Textbook of Respiratory Medicine. 2016:263–277.e2. doi: 10.1016/B978-1-4557-3383-5.00016-6.
2. Raffaele Pugliese, Stefano Regondi, Riccardo Marini, Machine learning-based approach: global trends, research directions, and regulatory standpoints, Data Science and Management, Volume 4, 2021, Pages 19-29,
3. Watson, D.S. Conceptual challenges for interpretable machine learning. Synthese 200, 65 (2022). https://doi.org/10.1007/s11229-022-03485-5
4. Yongjun Xu, Xin Liu, Xin Cao, Changping Huang, Enke Liu, Sen Qian, Xingchen Liu, Yanjun Wu, Fengliang Dong, Cheng-Wei Qiu, Junjun Qiu, Keqin Hua, Wentao Su, Jian Wu, Huiyu Xu, Yong Han, Chenguang Fu, Zhigang et al, Artificial intelligence: A powerful paradigm for scientific research, The Innovation, Volume 2, Issue 4, 2021, 100179,
5. Jairo Rojas-Delgado, Rafael Trujillo-Rasúa, Rafael Bello, A continuation approach for training Artificial Neural Networks with metaheuristics, Pattern Recognition Letters, Volume 125, 2019, Pages 373-380,
6. Tran, Hoa & Khatir, Samir & Ho-Khac, H. & De Roeck, Guido & Thanh, Bui & Abdel Wahab, Magd. (2020). Efficient Artificial Neural Networks based on a hybrid metaheuristic optimization algorithm for damage detection in laminated composite structures. Composite Structures. 10.1016/j.compstruct.2020.113339.
7. Long, Lyle & Gupta, Ankur. (2008). Scalable Massively Parallel Artificial Neural Networks. Journal of Aerospace Computing, Information, and Communication. 5. 10.2514/1.31026.
8. Nicoletti, Guy. (2000). An Analysis of Neural Networks as Simulators and Emulators. Cybernetics and Systems. 31. 253-282. 10.1080/019697200124810.
9. Gu, Rong & Shen, Furao & Huang, Yihua. (2013). A parallel computing platform for training large scale neural networks. Proceedings - 2013 IEEE International Conference on Big Data, Big Data 2013. 376-384. 10.1109/BigData.2013.6691598.
10. Grossi, Enzo & Buscema, Massimo. (2008). Introduction to artificial neural networks. European journal of gastroenterology & hepatology. 19. 1046-54. 10.1097/MEG.0b013e3282f198a0.
11. Mehrabi, Mohammad & Moayedi, Hossein. (2021). Landslide susceptibility mapping using artificial neural network tuned by metaheuristic algorithms. Environmental Earth Sciences. 80. 10.1007/s12665-021-10098-7.
12. Bansal P, Lamba R, Jain V, Jain T, Shokeen S, Kumar S, Singh PK, Khan B. GGA-MLP: A Greedy Genetic Algorithm to Optimize Weights and Biases in Multilayer Perceptron. Contrast Media Mol Imaging. 2022 Feb 24; 2022:4036035. doi: 10.1155/2022/4036035.
13. Alsmadi, Mutasem & Omar, Khairuddin & Mohd Noah, Shahrul Azman & Almarashdah, Ibrahim. (2009). Performance Comparison of Multi-layer Perceptron (Back Propagation, Delta Rule and Perceptron) algorithms in Neural Networks. 2009 IEEE International Advance Computing Conference, IACC 2009. 296 - 299. 10.1109/IADCC.2009.4809024.
14. Hertz, J. Palmer, R.G., Krogh.A.S., "Introduction to the Theory of Neural Computation", Perseus Publishing, 1991.
15. Maaranen, Heikki & Miettinen, Kaisa & Mäkelä, Marko. (2003). Training Multi Layer Perceptron Network Using a Genetic Algorithm as a Global Optimizer. 10.1007/978-1-4757-4137-7_20.
16. Ghaheri, ali & Shoar, Saeed & Naderan, Mohammad & Hoseini, Sayed shahabuddin. (2015). The Applications of Genetic Algorithms in Medicine. Oman Medical Journal. 30. 406–416. 10.5001/omj.2015.82.
17. Talbi, El-Ghazali. (2013). Combining metaheuristics with mathematical programming, constraint programming and machine learning. 4OR. 11. 10.1007/s10288-013-0242-3.