





Article

# A New Algorithm for Multivariate Genome Wide Association Studies Based on Differential Evolution and Extreme Learning Machines

David Álvarez Gutiérrez<sup>1</sup>, Fernando Sánchez Lasheras<sup>2,3,\*</sup> , Vicente Martín Sánchez<sup>4</sup> , Sergio Luis Suárez Gómez<sup>2,3</sup>, Víctor Moreno<sup>5,6,7,8</sup> , Ferrán Moratalla-Navarro<sup>5,6,7,8</sup> and Antonio José Molina de la Torre<sup>9</sup> 

- <sup>1</sup> SERGAS, UAP CS, 27720 A Pontenova, Spain; davidalvarezgutierrez@gmail.com
  - <sup>2</sup> Department of Mathematics, University of Oviedo, 33007 Oviedo, Spain; suarezsergio@uniovi.es
  - <sup>3</sup> Instituto Universitario de Ciencias y Tecnologías Espaciales de Asturias (ICTEA), University of Oviedo, 33004 Oviedo, Spain
  - <sup>4</sup> CIBERESP, University of Leon, Vegazana Campus, 24400 Leon, Spain; vicente.martin@unileon.es
  - <sup>5</sup> Oncology Data Analytics Programme, Catalan Institute of Oncology (ICO), Hospitalet de Llobregat, 08907 Barcelona, Spain; v.moreno@iconcologia.net (V.M.); fmoratalla@iconcologia.net (F.M.-N.)
  - <sup>6</sup> Colorectal Cancer Research Group, ONCOBELL Programme, Institut d'Investigació Biomèdica de Bellvitge (IDIBELL), Hospitalet de Llobregat, 08907 Barcelona, Spain
  - <sup>7</sup> Consortium for Biomedical Research in Epidemiology and Public Health (CIBERESP), 28029 Madrid, Spain
  - <sup>8</sup> Department of Clinical Sciences, Faculty of Medicine and Health Sciences, University of Barcelona, 08907 Barcelona, Spain
  - <sup>9</sup> IBIOMED, University of Leon, Vegazana Campus, 24400 Leon, Spain; ajmolt@unileon.es
- \* Correspondence: sanchezfernando@uniovi.es; Tel.: +34-985-103-338



**Citation:** Álvarez Gutiérrez, D.; Sánchez Lasheras, F.; Martín Sánchez, V.; Suárez Gómez, S.L.; Moreno, V.; Moratalla-Navarro, F.; Molina de la Torre, A.J. A New Algorithm for Multivariate Genome Wide Association Studies Based on Differential Evolution and Extreme Learning Machines. *Mathematics* **2022**, *10*, 1024. <https://doi.org/10.3390/math10071024>

Academic Editor: Ioannis G. Tsoulos

Received: 24 February 2022

Accepted: 21 March 2022

Published: 23 March 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Abstract:** Genome-wide association studies (GWAS) are observational studies of a large set of genetic variants, whose aim is to find those that are linked to a certain trait or illness. Due to the multivariate nature of these kinds of studies, machine learning methodologies have been already applied in them, showing good performance. This work presents a new methodology for GWAS that makes use of extreme learning machines and differential evolution. The proposed methodology was tested with the help of the genetic information (370,750 single-nucleotide polymorphisms) of 2049 individuals, 1076 of whom suffer from colorectal cancer. The possible relationship of 10 different pathways with this illness was tested. The results achieved showed that the proposed methodology is suitable for detecting relevant pathways for the trait under analysis with a lower computational cost than other machine learning methodologies previously proposed.

**Keywords:** machine learning; differential evolution; extreme learning machines; genome-wide association studies; single nucleotide polymorphism; pathways analysis

**MSC:** 68T20; 92B20; 68T05; 68W50

## 1. Introduction

Since the completion of the Human Genome Project [1] and the International HapMap Project [2], it has been well known by the scientific community that there is a clear link between some genes and certain traits and illness. Genome-wide association studies (GWAS) [3] are a powerful methodology that has proved its importance in the analysis of complex genomic problems.

Although GWAS studies present certain limitations, mainly concerning the need to make use of a large sample size to perform the study [4–6] and to have a population stratification [7] as well as the high probability of false positives [8,9] that must be managed with the help of the statistical techniques, the use of GWAS has allowed researchers to produce some outstanding scientific results.

The first GWAS date back to the years 2005 and 2006 [10,11]. Those first studies were of great interest at the time, as they found common variants associated to age-related macular degeneration. The fact that this type of study can look for genetic relationships in a multivariate way is one of the principal points in its favor. This means that a GWAS is able to deal with a large number of SNPs simultaneously and look for any relationship between them and a particular trait being analyzed [12]. It is also remarkable that GWAS does not require any a priori hypotheses, as they just take into account the relationships of SNPs that can be placed in any gene with the trait without needing to study a previous list of loci [13].

While in the first GWAS analysis performed, the link between SNPs and phenotypes was considered in a univariate way [14], the studies performed in recent years made use of multivariate methodologies [15] that, in some cases, include machine learning techniques [12]. Previous research has employed random forest, and this is just one instance of machine learning methodologies being used in GWAS in recent times [16–18]. In one particular case, [16], a method was used that was based on random forest designed with the express purpose of interpreting imbalanced genomic data. Another work [17] developed an algorithm called Reg-SNPs-intron with the help of random forest classifiers, while in the research performed by Roshan et al. [18], random forest was employed to study the number of casual variants and associated regions identified by top SNPs, and the results obtained were compared with other methodologies.

Furthermore, the gradient boosting methodology was already applied in the framework of GWAS [19,20]. Previous work [19] used gradient boosting models with the aim of differentiating inflammatory bowel disease (IBD) genes from non-IBD genes through the use of information from expression data and gene annotations. In other instances of the use of gradient boosting [20], its goal is to analyze genetic loci in order to make the task of interpreting large-scale genetic studies simpler, whilst maintaining their inherently unbiased character.

Support vector machines (SVM) were also employed to classify genes that cause inflammatory bowel diseases and those that do not [19].

Another article [21] applied SVM classification models to assess the potential of using GWAS data to predict duloxetine, and further research [12] employed them in a hybrid algorithm that combined SVM with genetic algorithms to determine which pathways would have an influence on colorectal cancer. Additionally, deep neural networks were employed in deep learning to predict the effect of genomic variants on tissue-specific expression [22] and genetic algorithm in GWAS analysis that makes use of hybrid algorithms combined with SVM [12].

This study was performed to propose and validate a new methodology based on machine learning techniques that can be applied in GWAS studies and could improve those currently available in a certain way. More specifically, the proposed methodology can find the most relevant SNPs in each pathway that lead to determining if an individual has a certain trait or is suffering from a particular illness. More specifically, the method presented in this paper is based on differential evolution and extreme learning machines. The performance of the new proposed methodology was checked with the help of a GWAS database, which helped to monitor the performance of this new proposed methodology as did a number of well-known pathways, all of which made it possible to make a comparison of the results with those obtained in some of our previous work [12].

## 2. Materials and Methods

### 2.1. Differential Evolution

Differential evolution (DE) is a metaheuristic algorithm that was proposed by Storn and Price [23], whereby the use of operators effects a random initialization of the population which then evolves to produce trial offspring. These operators are also employed in methods, such as crossover and mutation. In addition, DE uses a selection operator that

admits this offspring into the population or otherwise discards it, depending on the values of their objective function.

As was mentioned in the previous paragraph, DE uses a population of randomly chosen vectors as starting points. That population is the first one employed by the algorithm, while the following ones are created with the help of functions that modify the initial points. DE perturbs the vectors that make up a certain population by performing the scaled difference of two randomly selected vectors of the said population. To produce the new vector  $v_0$ , the result is added to another vector member of the same population. The procedure is repeated until all members of a population have competed against the newly generated vector  $v_0$ .

The issue of optimization is considered one of minimization in which the vector chosen for the next population is the one whose objective function value is the lowest. After all the trial vectors are tested, the survivors are the members of the next generation and will be responsible for creating new offspring.

The DE algorithm requires a population structure. The set of individuals of the  $x$ -th generation is represented by  $P_x$ , while each vector of that generation can be represented by  $v_{ix}$ , where  $i$  means that it is the  $i$ -th vector of the  $x$ -th generation.

Before the random population initialization, the upper and lower boundaries for each variable must be fixed in order to make sure that all components (variables) of all the members (vectors) of the population are inside both limits. Once the population of the DE algorithm is initialized, the algorithm mutates and recombines all the vectors that are part of the population. As was stated before, this process is based on re-scaling, subtraction and addition according to the following equation [24] that is employed in order to obtain each of the new vectors:

$$V_n = v_{r0x} + F(v_{r1x} - v_{r2x}), \quad (1)$$

where

$V_n$  is the new vector that is created, making use of two members of the  $x$ -th generation.

$F$  is a scale factor larger than 1 with no upper limit.

$v_{r0x}$  is the  $r0$ -th vector of the  $x$ -th generation.

$v_{r1x}$  is the  $r1$ -th vector of the  $x$ -th generation.

$v_{r2x}$  is the  $r2$ -th vector of the  $x$ -th generation.

DE also makes use of the uniform crossover. It involves taking two vectors from the same population and copying a certain proportion of components of one to another with certain probability  $p$ . In the DE algorithm selection, the operator is included, too. It works as follows: if the trial vector  $V_n$  has a lower value than the target vector  $V_{ix}$  (i.e., the  $i$ -th vector of the  $x$ -th generation), it replaces the target vector in the next generation; otherwise, the target retains its place in the population.

Once the new population is created, the process is repeated until the optimum is located, or a prespecified termination criterion is achieved. Appendix A contains Algorithms A1 that describes the pseudocode of the DE algorithm.

## 2.2. Extreme Learning Machines

When dealing with problems of regression and classification, extreme learning machine (ELM) has shown itself to be a highly practical learning algorithm [25]. A major point in its favor is that it can learn from data much more quickly than other machine learning methodologies [25]. It is not necessary to tune the parameters of the hidden layer iteratively, and it is possible to calculate the output weights by employing the least square optimization methodology [26,27].

In the case of the present research, ELM is used for classification. More specifically, regularized extreme learning machine (RELM) [28] will be employed. One of the main advantages of RELM is that not only is it able to obtain a better generalization that reduces the training error, but it also maximizes the edge distance.

RELM can consider empirical and structural risks at the same time [29]. The RELM mathematical model can be expressed as the following optimization problem [29]:

$$\min \left( \frac{1}{2} \|\beta\|^2 + \frac{\gamma}{2} \|\varepsilon\|^2 \right)^2, \quad (2)$$

Subject to

$$\sum_{i=1}^N \beta_{ig} (a_i x_j + b_i) - t_j = \varepsilon_j \quad (3)$$

where

$\beta$  is a parameter employed to smooth the cost function.

$\gamma$  is the proportion of the two kinds of risk parameters.

$\varepsilon$  represents the differences between the reference feature vectors and the feature vectors generated by the hidden layer of the RELM.

The equations presented above can be converted to an unconditional extremum problem with the help of a Lagrange function [30].

### 2.3. The Proposed Algorithm

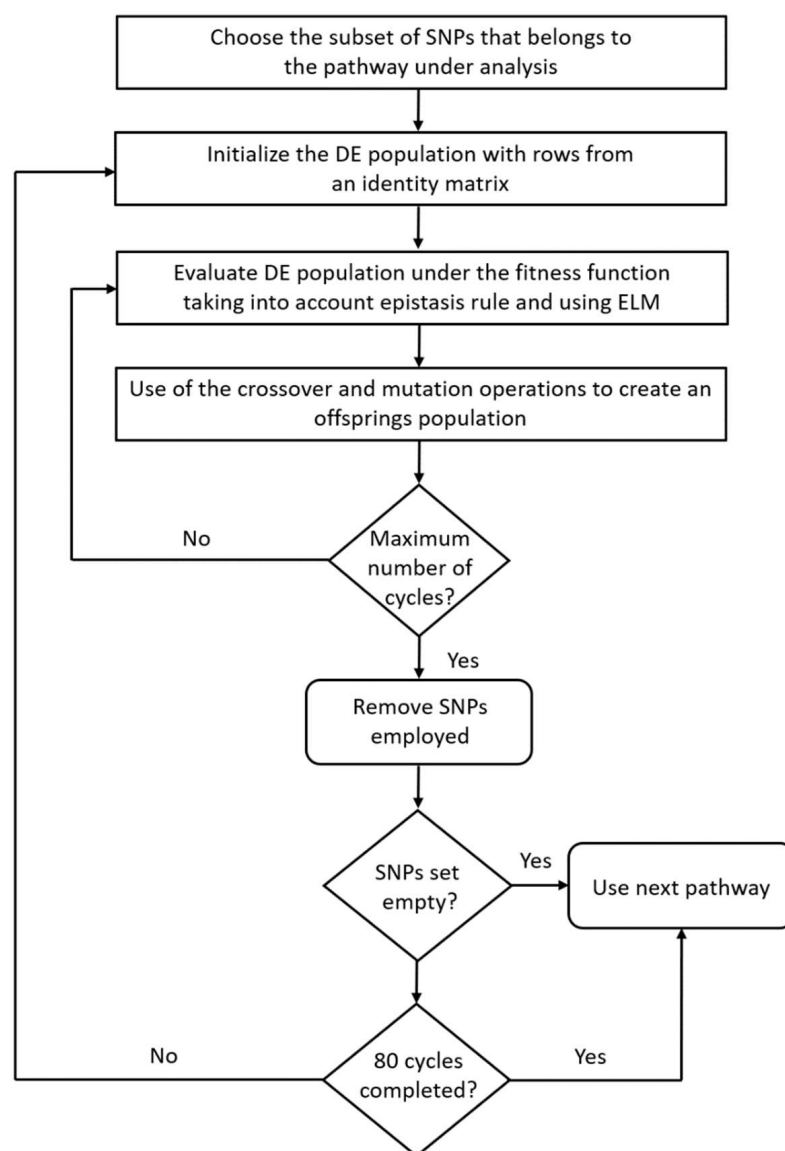
The algorithm presented in this paper employs DE and ELM. The aim of this new development is to find a certain subset of SNPs belonging to a previously defined pathway that would be able to perform a classification of individuals in cases and controls with certain accuracy. The flowchart of this new methodology is presented in Figure 1. Although the machine learning tools employed are different, the overall foundations of the algorithm are similar to those proposed in previous research [12] that made use of genetic algorithms and support vector machines.

The algorithm first takes all the SNPs pertaining to the pathway being considered. It should be pointed out that every time the algorithm is run, it only uses those SNPs corresponding to that pathway. As a consequence, only those SNPs belonging to this subset will end up being members of the search space used by the DE algorithm. All of the candidate solutions are considered an argument, taking the form of a vector of real numbers and produces a value indicating the fitness of the candidate solution under analysis as output. Please note that in the case of the present problem, the vector of real numbers is a string of "1s" and "0s" that indicates which SNPs of those that are in the subset will take part in the ELM model employed as a fitness function of the ELM. In our case, "1" indicates that the SNP will participate in the ELM model, while "0" means that it will not. It should also be mentioned that for each member of the population, a different ELM model will be trained.

In the first generation of the population, there will be only one active SNP in each of its members. This means that the initial population is built as if random rows from a square identity matrix of rank equal to the number of SNPs of the pathway under study were chosen as elements of the initial population. In the following generations, these population members will evolve through making use of the DE rules. The evolution is performed in such a way that only those members of the population with a higher value of the fitness function will be kept and employed to create the next generation of individuals. What this means is that in this algorithm, those variables that are active in the DE are employed as input information for the ELM model. The next step consists of assessing the performance of the ELM model.

Please note that the fitness function employed in this algorithm calculates the area under the ROC curve that is obtained by the individuals classified according to the results of the ELM model that makes use of the active variables. As in previous research [12] performed by the authors, the use of more than one SNP from the same gene is not allowed so as to prevent the epistasis phenomenon [31], which in this algorithm implementation means that the ROC value obtained by population members with epistasis is replaced by 0.





**Figure 1.** Flowchart of the proposed algorithm (DE—differential evolution, ELM—extreme learning machine).

The algorithm is applied as shown in Figure 1. The first step involves choosing the subset of SNPs that belongs to the pathway under analysis. The DE population is then initialized with a set of individuals in which one SNP is active. This is why in the flowchart it says that the DE population is initialized by making use of the rows of an identity matrix. The following step consists of evaluating the DE population under the fitness function, taking into account the epistasis rule and using ELM as the classification algorithm whose performance is assessed by means of the area under the ROC curve. Taking into account the performance obtained, a new DE population is created, using the current population as a starting point. Should the maximum number of cycles be reached, the SNPs already employed in the element of the population which has performed the best are taken out. This process is then carried out again, and only halted when the SNPs set is empty or when 80 cycles in total have reached completion.

The proposed algorithm is also applied through permutating the labels of cases and controls. Taking into account the advice in the existing literature, the permutation process was repeated 10,000 times [32]. Please note that in our previous research [12], due to computational limitations, only 1000 permutations were employed. This option is also

valid, and some examples can be found in the literature [33–35] but in this case, it was possible to perform 10,000 permutations.

#### 2.4. The Database

In this work, the data set used forms part of the Colorectal Cancer Transdisciplinary Study (CORECT), which was an observational multicentric multicase control study carried out between September 2008 and December 2013, while the subset of information employed came from Leon University Hospital and Bellvitge Hospital [12].

This information set had a total of 2019 people, of whom 1076 had colorectal cancer. For the present research, 370,570 distinctive SNPs of each person were employed. The cases considered in this research were histologically affirmed, and their ages were within the range from 20 to 85 years old. The choice of controls was made at random from the population records assigned to family doctors within the same area of the hospitals that participated in this research with the same age and sex, all having lived in the area under study for at least 6 months.

The subjects that took part in this study were all volunteers. The safety and privacy of the research was guaranteed through the protocol and ethical document approved by the Ethics Committees of the Study. As no personal information was required, it was removed from the database to ensure confidentiality. All the files that include information about the subject complied with Spain's Organic Law 15/1999. The files employed in the study were registered with the Spanish Data Protection Agency with the record number 2102672171. Please also note that not only was approval obtained of the Ethics Committee of the study, but also a double informed consent was requested to all the patients. Genetic samples were stored in regional genetic banks of the autonomous communities of Castilla y León (Spain) and Catalonia (Spain).

It is of interest to remark that comparison of the results of the present research with those obtained in previous work [12] is possible, as the same pathways were chosen. These are 10 pathways that belong to the KEGG database [36–38]. These pathways can be divided into three different kinds: pathways whose relationship with colorectal cancer was already proved, pathways whose relationship with colorectal cancer is inconclusive according to the current medical literature and finally, pathways that according to current medical literature are unlikely to have a relationship with colorectal cancer.

### 3. Results

In the present research, the DE population size was fixed at 5000, and the maximum number of iterations allowed was 6000. The crossover probability was fixed at 50%, and a differential weighting factor (F) of 0.8 was employed. The proposed algorithm was applied in the same way to all the pathways described in the database section.

This algorithm first creates a subset made up exclusively of SNPs pertaining to the pathway being analyzed. In Table 1, the number of SNPs belonging to each pathway may be seen. As this table shows, the longest of these is the Huntington's disease pathway with 1980 SNPs, while the shortest is the mitochondrial biogenesis pathway with 679.

In total, 5500 individuals make up the initial population. As mentioned before in the Materials and Methods section, only one active SNP was found in all these members of the initial population. It is to be noted that fewer than 2600 SNPs are involved in any of the pathways being analyzed, and so there are some elements in the initial population that are the same as one another. Taking one such population as a starting point, new generations are created in which two or more SNPs may simultaneously be active. These elements evolve, generation after generation, in search of the AUC maximization. This process is repeated until generation 6000 is reached.

The AUC value obtained after the application of the algorithm to the adipocytokine-signaling pathway was 0.542125, the AUC value obtained for the AMPK-signaling pathway was 0.569859, etc. The values obtained for each of the pathways under analysis can be consulted in the column called AUC of Table 1.

**Table 1.** Pathways under analysis. Total number of SNPs per pathway (Tot. SNPs), SNPs employed in the 80 iterations by the non-permuted phenotypes (SNPs employed), average AUC obtained in the 80 iterations by the non-permuted phenotypes (AUC), AUC obtained by the permuted phenotypes (AUC perm.), percentage of non-permuted AUC values that are higher than the maximum permuted AUC value (win subsets).

Pathway Name	Tot. SNPs	SNPs Employed	AUC	AUC Perm.	Win Subsets
Adipocytokine signaling pathway	752	558	0.542125	0.542302	17.30%
AMPK signaling pathway	1812	508	0.569859	0.555751	89.45%
Apelin signaling pathway	2525	626	0.578882	0.542724	100%
Colorectal cancer pathway	813	399	0.583862	0.569623	100%
Glucagon signaling pathway	1707	439	0.563172	0.550494	81.95%
Huntington's disease	1980	493	0.552506	0.546617	85.15%
Insulin resistance	1574	489	0.554821	0.558040	30.05%
Insulin signaling pathway	1215	487	0.557398	0.551704	95.90%
Longevity regulating pathway	1481	443	0.535514	0.533965	48.85%
Mitochondrial biogenesis	679	362	0.578295	0.550624	100%

In this research, an algorithm combining DE and ELM was tested, and on average, each of the 6000 iterations used took 0.73 s. When the iterations were completed, those SNPs that were employed for training the ELM model with the best performance were removed, so that the algorithm would repeat the process in search of those SNPs that are able to provide the best classification but without being able to make use of those that have already shown the best classification performance. It would be possible to repeat this process for as long as the pathway has SNPs available, but instead of repeating the process until no more SNPs were available, the algorithm was programmed to stop the process after 80 cycles.

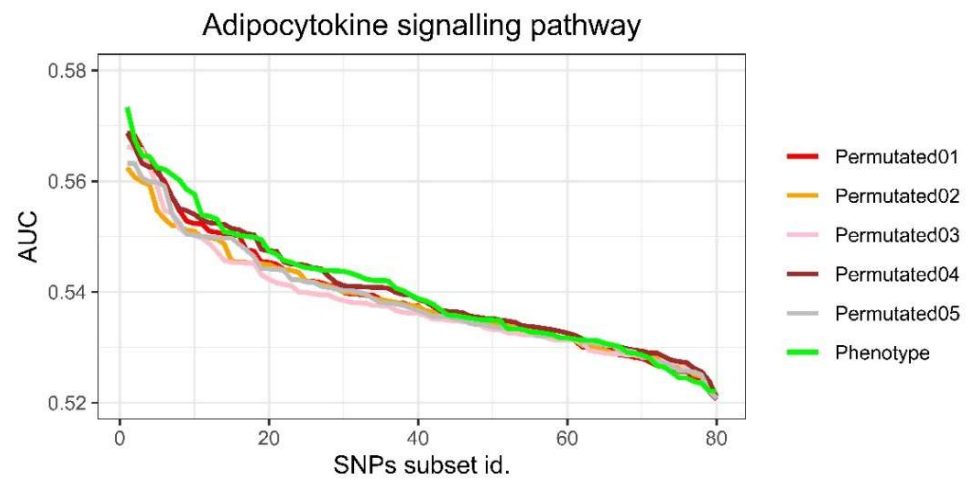
The reasons that lead the researchers to work in this way are twofold: on the one hand, the algorithm has a quite high computational cost, which means that repeating the loop while there are still SNPs available would be a highly time-consuming process and, also, that number of SNPs in the different pathways are not the same, and so stopping the process before running out of SNPs makes it easier to compare the results.

Once the algorithm execution was finished, it was run again, making use of the permutation of the cases and controls labels but preserving the number of individuals in each of these categories. Please note that the use of this methodology is very common for these kinds of studies [12].

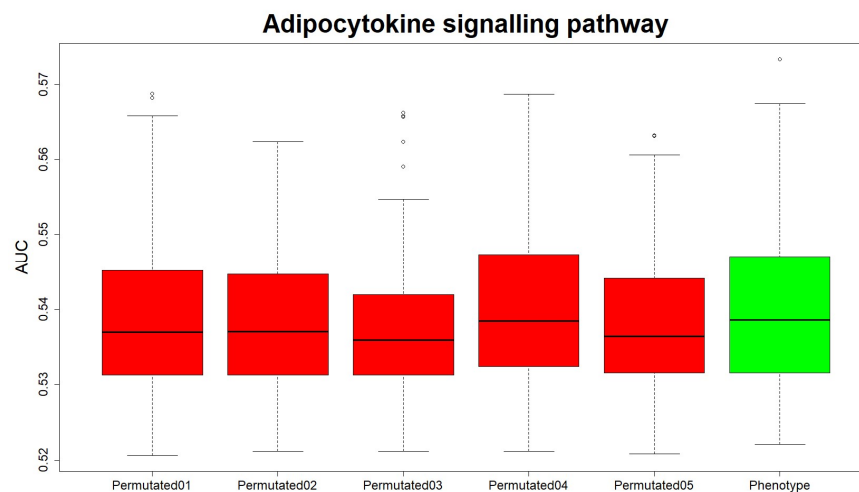
The numerical results produced are presented in Table 1. This table contains the following information on all the pathways under analysis: how many SNPs make up the pathway subset, how many different SNPs of the pathway employed in any of the models showed the best AUC performance, what AUC value the algorithm obtained on average, the AUC value when phenotypes were permuted, and what percentage of non-permuted AUC values were greater than the highest permuted AUC value that was obtained.

In this research, a figure for each pathway under analysis was created. Figure 2 shows the values that were obtained by applying the proposed algorithm to the adipocytokine-signaling pathway in six different cases, one of which is the one in which cases and controls were labeled correctly as cases and controls while the other five show the results achieved by five different permutations. In the six algorithm executions presented, the AUC values are ordered from higher to lower to make it easier to interpret the curves obtained. In the case of the pathway mentioned, the curve that represents data without cases and controls permutations is not higher than those curves obtained by making use of permuted cases and controls labels, as they are very close to each other. This result is confirmed by the boxplot presented in Figure 3, where the median AUC value obtained for cases and controls

correctly labeled is very similar to the median values achieved for the five algorithm executions with permuted labels represented in the same figure.

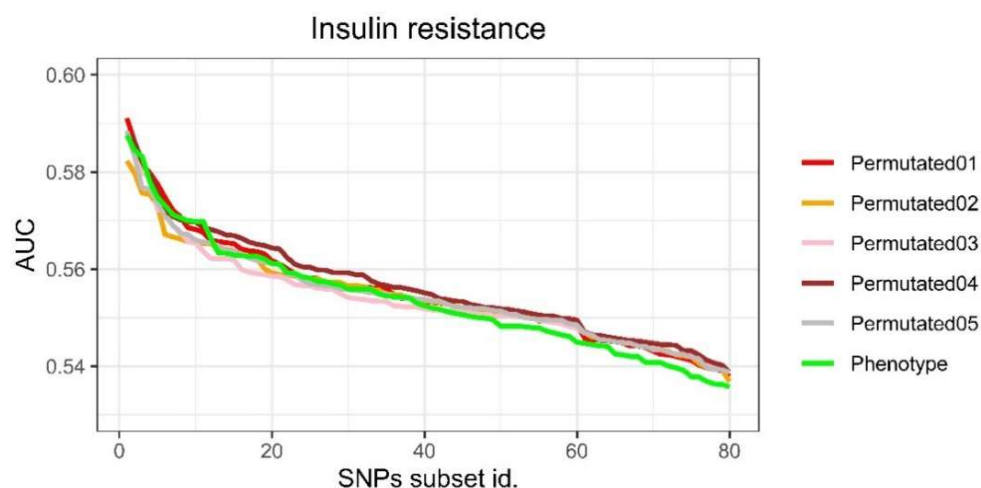


**Figure 2.** Adipocytokine signaling pathway: AUC values of 80 iterations for the execution of the algorithm with cases and controls correctly labeled and five different permutations.

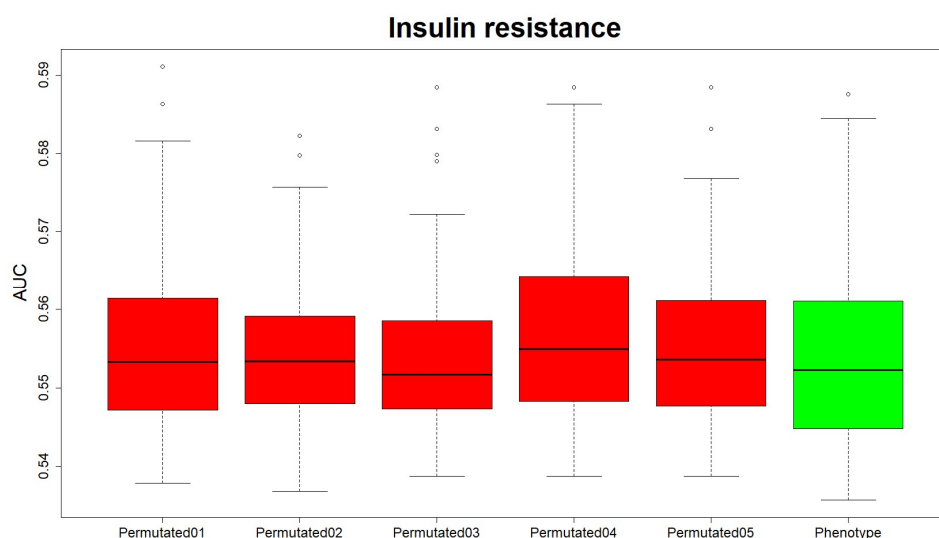


**Figure 3.** Boxplot of the adipocytokine signaling pathway AUC values of the 80 iterations of 5 executions of the algorithm with the labels of cases and controls permuted and one with cases and controls correctly labeled. Red color represents permuted cases, green color represents the phenotype case and circles are outliers.

In Figure 4 the results obtained for the insulin-resistance pathway can be seen. Figure 5 shows the boxplot of the insulin resistance AUC values of the 80 iterations of 5 executions of the algorithm with the labels of cases and controls permuted and one with cases and controls correctly labeled, while Figure 6 shows the same information that Figure 4 for the longevity pathway. In both cases, and this also occurs in Figure 2, for the green curve, called a phenotype and which refers to the results produced by the algorithm when applied to the data with the labels correctly assigned to cases and controls, the AUC values show a great similarity to those obtained with permuted labels. Please also note that the same effect can be noticed in Figures 5 and 7 as in both cases the median value of the AUCs for cases and controls is not higher than the median value of all the permuted executions. Therefore, it must be inferred that these three pathways are not linked to colorectal cancer.



**Figure 4.** Insulin resistance: AUC values of 80 iterations for the execution of the algorithm with cases and controls correctly labeled and five different permutations.



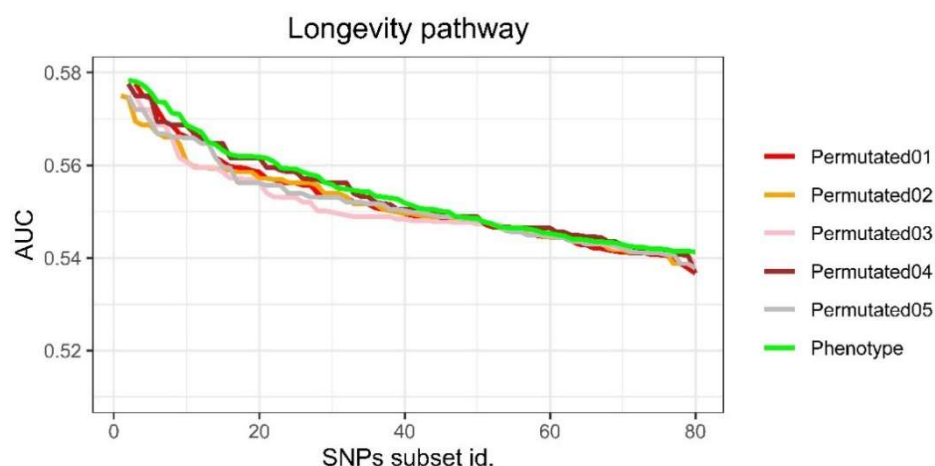
**Figure 5.** Boxplot of the insulin resistance AUC values of the 80 iterations of 5 executions of the algorithm with the labels of cases and controls permuted and one with cases and controls correctly labeled. Red color represents permuted cases, green color represents the phenotype case and circles are outliers.

Considering the results obtained by means of the algorithm, the AUC value obtained was on average a little greater when the algorithm was applied to those data sets in which cases and controls labels were permuted than the one achieved for cases and controls. Regarding the insulin resistance pathway, there was a greater value for cases and controls, by a difference of 0.003219 or 0.58%, while for the longevity pathway, this difference was 0.001549, or 0.29%. A summary of these results can be seen in Table 1.

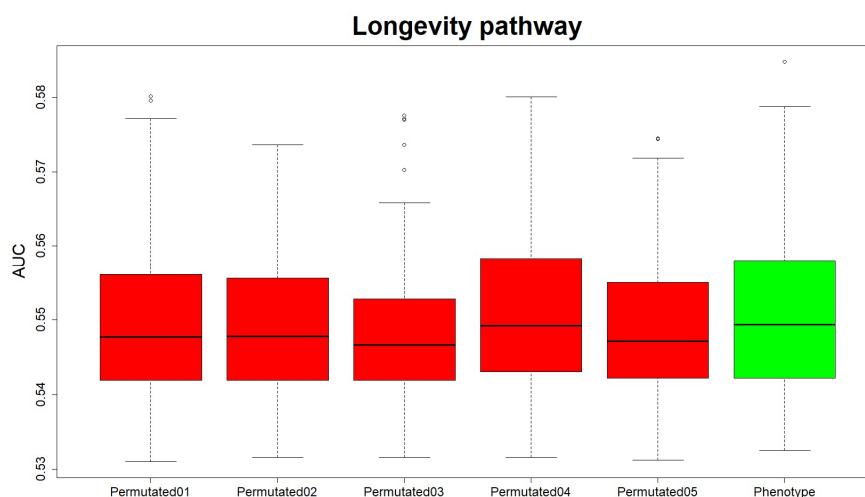
The three following pathways under study were the apelin-signaling pathway that is represented in Figures 8 and 9, the mitochondrial biogenesis pathway that is presented in Figures 10 and 11 and the colorectal cancer pathway that can be observed in Figures 12 and 13. In these three pathways, the curve that represents the AUC values obtained by the algorithm when applied to the subset in which labels correspond to cases and control has undoubtedly higher values than those obtained when permutation was applied. The same effect is observed in their corresponding boxplots, as the median value for the algorithm execution without permutation of cases and controls is higher than the five executions in which cases and controls are permuted. Even in the case of the apelin-signaling pathway and the



mitochondrial biogenesis pathway, the value that corresponds to the 25th percentile of the executions without permutation is higher than the 75th percentile of most of the permuted executions. These graphic results are in line with those presented in Table 1, where the AUC value for the apelin-signaling pathway was 0.578882, as compared to 0.542724 when it was permuted. Something similar occurred in the case of the mitochondrial biogenesis pathway, where the AUC value achieved was 0.578295, while with permuted labels, the average value was 0.550624. In the case of the colorectal cancer pathway, the AUC value was 0.583862, while in that of the permuted pathways, the average value was 0.569623. It is also of interest to highlight that in these three pathways, fully 100% of the permuted subsets obtained results under the subset with cases and controls correctly labeled.



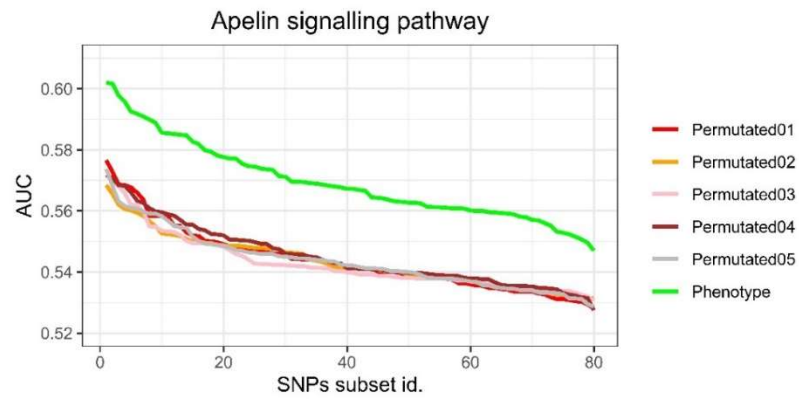
**Figure 6.** Longevity pathway: AUC values of 80 iterations for the execution of the algorithm with cases and controls correctly labeled and five different permutations.



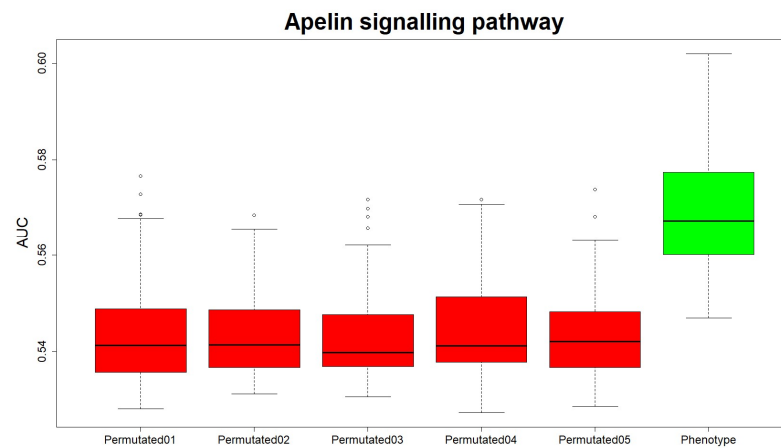
**Figure 7.** Boxplot of the longevity pathway AUC values of the 80 iterations of 5 executions of the algorithm with the labels of cases and controls permuted and one with cases and controls correctly labeled. Red color represents permuted cases, green color represents the phenotype case and circles are outliers.

In other words, these three pathways are the most likely of the 10 under analysis in the present research to influence over the colorectal cancer. This can be clearly noticed when their figures are compared with, for example, those that correspond to the results obtained for the adipocytokine-signaling pathway (Figure 2), insulin resistance pathway (Figure 4) or longevity pathway (Figure 6) or even with others that will be described afterwards in the present section. From a machine learning point of view, it can be interpreted that these

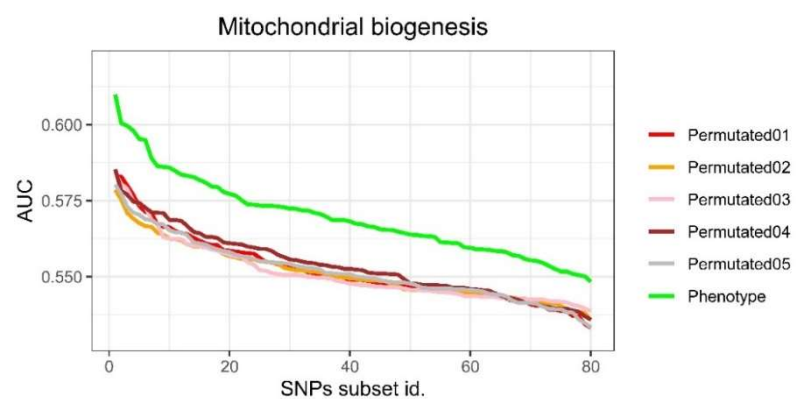
pathways are the ones that are able to separate in a clearer way the cases from controls, but from a genetic point of view, it is translated as a possible influence of the pathway under study on the trait of interest that in the present research is suffering or not from colorectal cancer. A detailed description of the possible biological reasons of the importance of the three referred pathways in the colorectal cancer can be found in the discussion section.



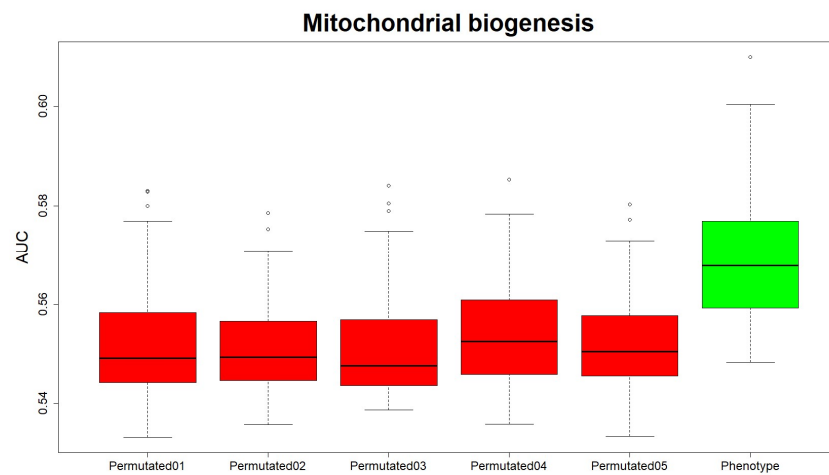
**Figure 8.** Apelin-signaling pathway: AUC values of 80 iterations for the execution of the algorithm with cases and controls correctly labeled and five different permutations.



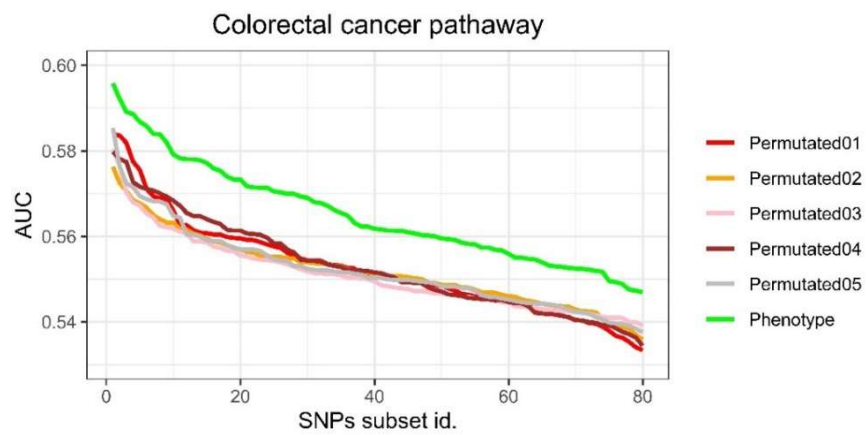
**Figure 9.** Boxplot of the apelin-signaling pathway AUC values of the 80 iterations of 5 executions of the algorithm with the labels of cases and controls permuted and one with cases and controls correctly labeled. Red color represents permuted cases, green color represents the phenotype case and circles are outliers.



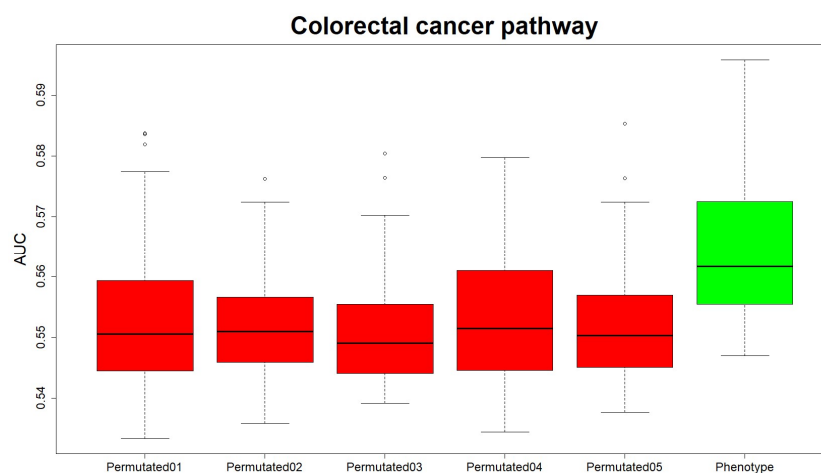
**Figure 10.** Mitochondrial biogenesis pathway: AUC values of 80 iterations for the execution of the algorithm with cases and controls correctly labeled and five different permutations.



**Figure 11.** Boxplot of the mitochondrial biogenesis pathway AUC values of the 80 iterations of 5 executions of the algorithm with the labels of cases and controls permuted and one with cases and controls correctly labeled. Red color represents permuted cases, green color represents the phenotype case and circles are outliers.

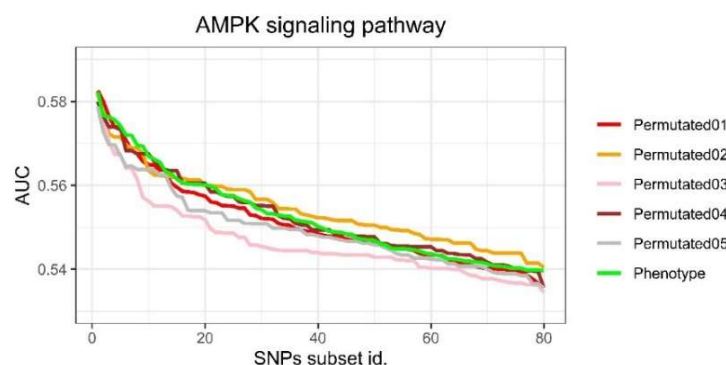


**Figure 12.** Colorectal cancer pathway: AUC values of 80 iterations for the execution of the algorithm with cases and controls correctly labeled and five different permutations.

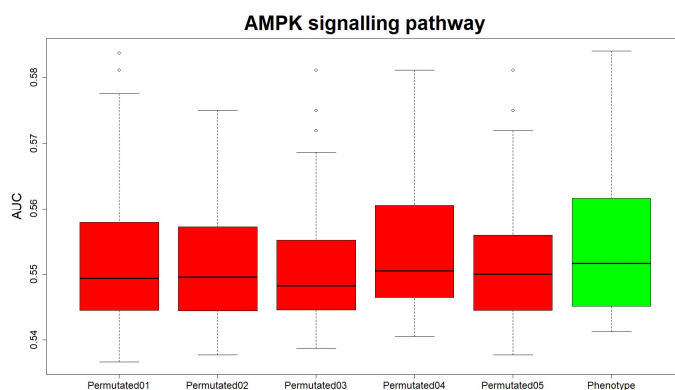


**Figure 13.** Boxplot of the colorectal cancer pathway AUC values of the 80 iterations of 5 executions of the algorithm with the labels of cases and controls permuted and one with cases and controls correctly labeled. Red color represents permuted cases, green color represents the phenotype case and circles are outliers.

Figure 14 represents how the algorithm behaves when applied to the AMPK-signaling pathway. For this pathway, the phenotype curve is close to the permuted ones. The same result is observed in the boxplot of Figure 15. The numerical results, presented in Table 1, where the AUC value of the phenotype curve is higher than the value obtained when the cases and controls were permuted, show a slightly better result in the case of no permutation. In this figure, the cases and control curves perform better than 89.45% of the permuted curves. In other words, there are some randomly labeled subsets where the algorithm achieves better AUC results than in the case of cases and controls.



**Figure 14.** AMPK-signaling pathway: AUC values of 80 iterations for the execution of the algorithm with cases and controls correctly labeled and five different permutations. Red color represents permuted cases, green color represents the phenotype case and circles are outliers.

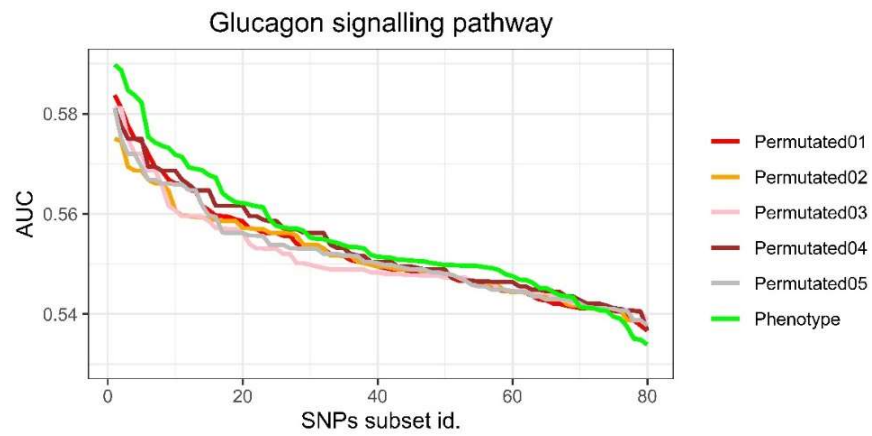


**Figure 15.** Boxplot of the AMPK-signaling pathway AUC values of the 80 iterations of 5 executions of the algorithm with the labels of cases and controls permuted and one with cases and controls correctly labeled. Red color represents permuted cases, green color represents the phenotype case and circles are outliers.

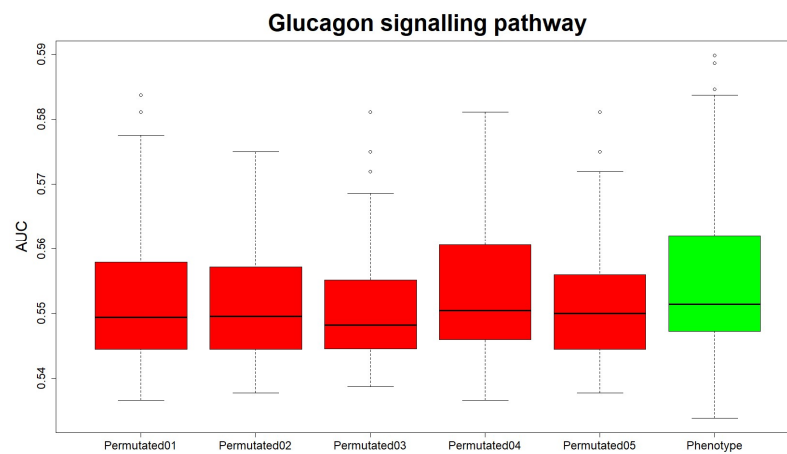
Figures 16 and 17 show the results obtained for the glucagon-signaling pathway, while Figures 18 and 19 do the same for the Huntington's disease pathway. In both cases, the phenotype curve is very close to the permuted curves, although its average value appears to be a little higher. Taking into account those values presented in Table 1, in both pathways, the AUC value of the cases and control is slightly higher than the permuted ones. In the case of the glucagon-signaling pathway, 81.95% of the permuted subsets obtained an AUC value lower than the ones not permuted, while in the case of the Huntington's disease pathway, these percentages rose to 85.15%.

Finally, Figure 20 shows the results obtained when the algorithm is applied to the insulin-signaling pathway. In this case, the phenotype curve is close to the permuted one, but at most of the points, the former is above the latter. In the case of the boxplot presented in Figure 21, the median of the AUC value of the execution without permutation is higher than the five permutations showed for the same pathway. Please also note that in this case,

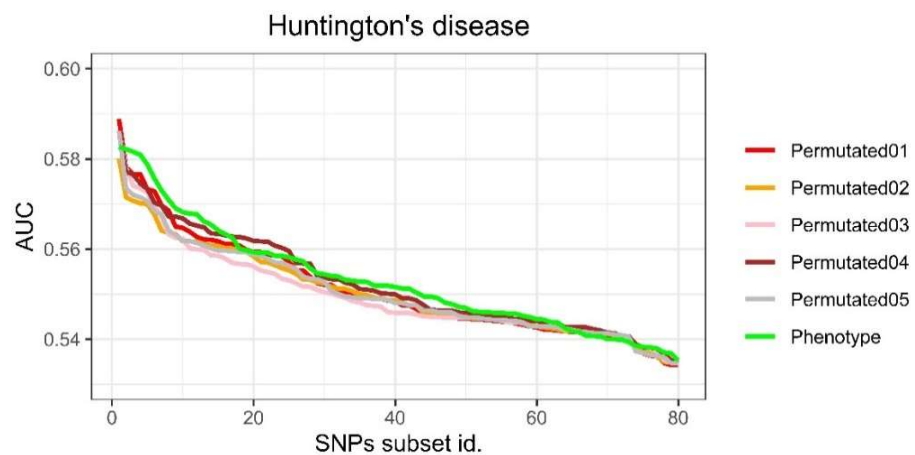
the AUC value for cases and controls is over the average permuted value and is higher than 96.5% of the permuted results obtained.



**Figure 16.** Glucagon-signaling pathway: AUC values of 80 iterations for the execution of the algorithm with cases and controls correctly labeled and five different permutations.

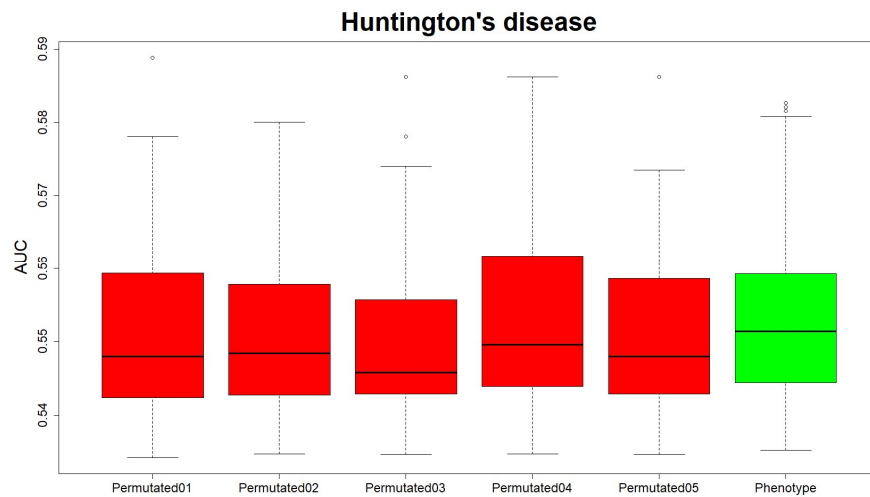


**Figure 17.** Boxplot of the glucagon-signaling pathway AUC values of the 80 iterations of 5 executions of the algorithm with the labels of cases and controls permuted and one with cases and controls correctly labeled. Red color represents permuted cases, green color represents the phenotype case and circles are outliers.

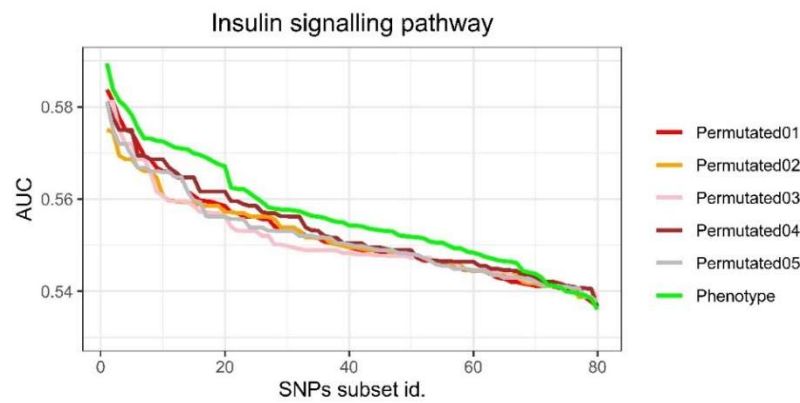


**Figure 18.** Huntington's disease pathway: AUC values of 80 iterations for the execution of the algorithm with cases and controls correctly labeled and five different permutations.

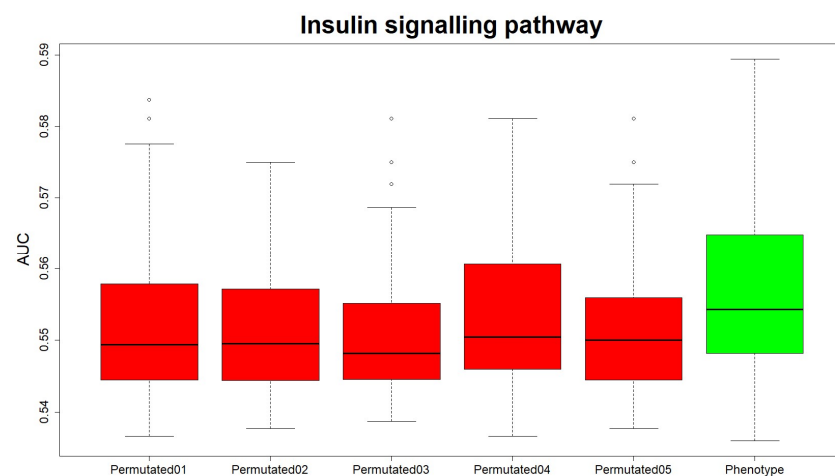




**Figure 19.** Boxplot of the Huntington’s disease pathway AUC values of the 80 iterations of 5 executions of the algorithm with the labels of cases and controls permuted and one with cases and controls correctly labeled. Red color represents permuted cases, green color represents the phenotype case and circles are outliers.



**Figure 20.** Insulin-signaling pathway: AUC values of 80 iterations for the execution of the algorithm with cases and controls correctly labeled and five different permutations.



**Figure 21.** Boxplot of the insulin-signaling pathway AUC values of the 80 iterations of 5 executions of the algorithm with the labels of cases and controls permuted and one with cases and controls correctly labeled. Red color represents permuted cases, green color represents the phenotype case and circles are outliers.

#### 4. Discussion

The results obtained by the algorithm proposed in the present research that makes use of ELM for classification are quite similar to those achieved in a study carried out by the authors in which SVM and genetic algorithms were employed [12]. The ELM performance was quite similar to that obtained by SVM, but the time required for the model training was considerably lower (times approximately divided by 47), which makes the use of ELM especially convenient considering the long time required for SVM model training.

As was already stated in the materials and methods section, ELM is a learning algorithm based on feed-forward neural networks that makes use of a single hidden layer [39]. Although ELM has a weaker generalization ability than SVM for a small sample, it can generalize as well as SVMs for large samples.

In our research, as in earlier works, it was found that ELM has advantages over SVM in the selection of parameters [40]. Like SVM, ELM is able to minimize the training errors as well as maximizing the separation margin [41].

Despite there being a great deal of existing literature, more profound research about the comparison of ELM and SVM is still necessary. The present research confirms a theoretical result found in the existing literature that indicates that ELM can achieve similar accuracy to SVM [40].

Previous research [39] made an experimental comparison of SVM and ELM for different training sample sizes. This research employed eight different data sets with samples ranging from 150 to 22,784 and the number of variables from 4 to 8. These results suggested that SVM has the strongest generalization behavior. This fact is important in the case of small sample sizes, but when the size of the training data set increases, the generalization ability of the ELM becomes closer to the SVM, making both have similar classification abilities for large sample sizes. According to previous research [42], DE outperforms GA on many optimization problems, both single [24] and multi-objective [43].

In the case of the present work, the sample size selected was the same as in previous research [12] to make it possible to compare the results. Please also note that in this case, instead of 1000 permutations, a number chosen due to the high computational cost set out in said research [12], 10,000 permutations were performed.

From our point of view, the results reached in this work are not only very similar to those obtained in previous work by [12] that made use of the same database, but also are in line with others available in existing literature. One of the main concerns of readers unfamiliar with GWAS studies is that the AUC values obtained are quite low, yet this is frequently the case [44] in this research area.

Taking these results into account, we concluded that some of the proposed pathways are obviously related to colorectal cancer, namely apelin signaling, colorectal cancer and mitochondrial biogenesis. Others present a probable, albeit weak, relationship with colorectal cancer, namely AMPK signaling, glucagon signaling, Huntington's disease and insulin signaling. We failed to find any relationship for adipocytokine signaling, insulin resistance or longevity regulating.

Regarding colorectal cancer, previous research has already highlighted the importance of the apelin-signaling pathway [45]. Apelin (AP) may potentially be a target for anticancer therapy [46,47]. Further research [48] studied the tumor tissues of 56 surgically treated colorectal adenocarcinoma patients and therein carried out an analysis of the apelin and its receptor mRNA as well as the levels of protein expression. The values obtained were compared with 27 healthy controls, finding that serum levels of apelin and its receptor were increased in colorectal cancer patients in comparison to controls. These results lead to the conclusion that apelin is an important factor in the progression of colorectal carcinoma. Finally, a recently published study came to the conclusion that the level of apelin and its receptors is closely linked to the regulation of migration and invasion of colon cancer cells [49]. The presence of the colorectal cancer pathway as being one of those linked to patients that suffer from colorectal cancer was expected, and it can be considered a basic test to guarantee the correct behavior of the algorithm.

As is well known and has already been stated in previous research, mitochondria are linked to the genesis of cancer [12]. In fact, cancer development in humans is closely linked to mitochondrial alteration, increased production of free mitochondrial oxide radicals and oxidative stress [50].

There are other pathways in which a certain degree of relationship with the labels of cases and controls of colorectal cancer was found. The first of these is the AMPK-signaling pathway. In existing literature, there are now some papers that link it with colorectal cancer, either considering that AMPK promotes the survival of colorectal cancer stem cells [51] or showing that the p-AMPK expression is more frequent in controls than in colorectal cases [52].

The fact that glucagon boosts the production of glucose in the liver by increasing glycogenesis and gluconeogenesis whilst also reducing glycogenesis glycolysis is now widely known. Research performed in 2008 [53] discovered expressions of the glucagon receptor in colon cancer cell lines and in colon cancer tissue obtained from patients. Furthermore, another study which came out the same year reported that the growth of colon cancer cells is promoted by glucagon through regulating AMPK and MAPK pathways [54].

In the case of Huntington's disease, previous studies have found [55] that those who suffer from this illness have up to 80% less cancer than the general population. The reason is that the mutated huntingtin (HTT) gene in those who suffer from Huntington's disease generates a class of small molecules that are highly toxic to cancer cells.

The relationship between the insulin-signaling pathway and the risk of colorectal cancer is backed by previous research [56] that found that genetic variations in the insulin-signaling pathways genes may affect the risk of colorectal cancer. It must also be taken into account that the modification in the individual values of plasma insulin levels due to diet may also affect the risk of suffering from colorectal cancer [56]. Finally, it must be highlighted that no relationship was found between the adipocytokine-signaling, insulin-resistance and longevity-regulating pathways. These results are in line with the lack of results about these possible relationships found in literature.

Although the algorithm presented in this research has proved a satisfactory predictive ability, this ability lacks easy biological interpretability [57]. In order to deal with this limitation, the use of explainable artificial intelligence techniques is required. As it was already pointed out by other authors [58], the main advantage of explainable artificial intelligence techniques is that they integrate interpretability and transparency into the machine learning models [59], which, in the case of the problem under study, means that the relevance of the different SNPs in a certain pathway would be known. Additionally, it must be highlighted that from a biological point of view, the use of explainable artificial intelligence techniques systems would help healthcare professionals in gaining a better understanding of the model and to make reasoned decisions.

Finally, from the point of view of the researchers, one of the first explainable artificial intelligence techniques that should be tested with the database of the present research is random forest. Random forest is a combination of predictor trees such that each tree depends on the values of a random vector tested independently and with the same distribution for each of them [60]. It can be considered a substantial modification of bagging that builds a large collection of uncorrelated trees and then averages them [61].

## 5. Conclusions

The research described in this paper presents a novel algorithm based on machine learning methodologies that has proved to give good performance in GWAS. This work continues a research line [12] that makes use of algorithms and combines different machine learning methodologies. One of the main drawbacks of these methodologies is the lack of a simple biological explanation for the results obtained, as, although there are many pathways whose relationships with illness and traits are well known, it is difficult to find how each of the SNPs that form the pathway behaves in the process and influences it. In

spite of this, it is possible to explain the influence of the different pathways on colorectal cancer by making use of the available literature.

As was already stated in a previous work [12], in our opinion, due to the current lack of a gold standard multivariate methodology for GWAS, all the algorithms, such as the one presented in this research, should be taken into account in GWAS. Additionally, when compared with the previous algorithm proposed by the authors [12], the fact that the computation time required for this one is about 1 divided by 47 must be considered as one of the main advantages of the algorithm proposed in this research. This result is mainly due to the fast training of the ELM that was already stated in the literature [62].

Furthermore, as it was proved in the present research, and in line with previous works, machine learning is a valuable tool for GWAS analysis, as it is able to find SNPs and the loci of interest. In spite of these, one of the main drawbacks of machine learning is the lack of a clear explanation for the models obtained with most of the methodologies. This fact is important in the field of GWAS where the relationships between genes and traits are often difficult to interpret. Considering what was mentioned above, one of our future research lines will consist of applying a machine learning explainable algorithm, such as random forest, to the problem under study in this paper. Additionally, and in order to contribute to consolidating the role of machine learning algorithms in GWAS, the authors will also focus on comparing the results obtained with other methodologies that are common in GWAS and do not belong to the machine learning field in order to close the gap between the different approaches to GWAS.

**Author Contributions:** Conceptualization, D.Á.G., V.M., V.M.S. and F.S.L.; data curation, F.M.-N. and A.J.M.d.l.T.; formal analysis, V.M.; methodology, F.S.L. and D.Á.G.; project administration, V.M.S.; resources, F.M.-N. and A.J.M.d.l.T.; software, S.L.S.G. and F.S.L.; validation, D.Á.G. and S.L.S.G.; visualization, F.S.L.; writing—original draft, D.Á.G. and F.S.L.; writing—review and editing, D.Á.G., F.S.L., V.M., F.M.-N., A.J.M.d.l.T., S.L.S.G. and V.M.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was partially funded by Agency for Management of University and Research Grants (AGAUR) of the Catalan Government, grant number 2017SGR723, Instituto de Salud Carlos III, co-funded by FEDER funds—a way to build Europe—grants and Spanish Association Against Cancer (AECC) Scientific Foundation grant GCTRA18022MORE.

**Institutional Review Board Statement:** The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Institutional Review Board of Instituto Municipal de Asistencia Sanitaria de Barcelona (Spain) with protocol code 2008/3123/I on date 3 September 2008 and approved the 29 May 2009 by the Ethical Committee of Leon Hospital (Spain) without any protocol number.

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

**Data Availability Statement:** The information presented in this study are available on request from the corresponding author. Please note that data are not publicly available due to privacy.

**Acknowledgments:** The authors would like to thank Anthony Ashworth for his revision of the English grammar and spelling in the manuscript. We thank CERCA Programme, Generalitat de Catalunya for institutional support.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A

Algorithms A1 shows the algorithm of the DE algorithm. This pseudocode requires 4 indexes, one of which is the target index  $i$ , while the other three are the vector indexes, called  $r_0$ ,  $r_1$ , and  $r_2$ . Please note that  $r_0 \neq r_1$  and  $r_1 \neq r_2$ . When the population is completed, the selection is performed. In this pseudocode,  $N_p$  represents the number of elements in the population.

**Algorithms A1.** Algorithm of the differential evolution (DE) algorithm

---

```

// initialize...
do // generate a trial population
{
  for (i = 0; i < Np; i++) // r0! = r1! = r2! = i
  {
    do r0 = floor(rand(0,1)*Np); while (r0 == i);
    do r1 = floor(rand(0,1)*Np); while (r1 == r0 or r1 == i);
    do r2 = floor(rand(0,1)*Np); while (r2 == r1 or r2 == r0 or r2 == i);
    jrand = floor(D*rand(0,1));
    for (j = 0; j < D; j++) // generate a trial vector
    {
      if (rand (0,1) <= Cr or j == jrand)
      {
         $u_{j,i} = x_{j,r0} + F * (x_{j,r1} - x_{j,r2});$  //check for out-of-bounds?
      }
      else
      {
         $u_{j,i} = x_{j,i};$ 
      }
    }
  }
  // select the next generation
  for (i = 0; i < Np; i++)
  {
    if ( $f(u_i) <= f(x_i)$ )  $x_i = u_i;$ 
  }
} while (termination criterion not met);

```

---

**References**

- Venter, J.C. The sequence of the human genome. *Science* **2001**, *291*, 1304–1351. [[CrossRef](#)] [[PubMed](#)]
- Gibbs, R.A.; Belmont, J.W.; Hardenbol, P.; Willis, T.D.; Yu, F.L.; Yang, H.M.; Ch'ang, L.Y.; Huang, W.; Liu, B.; Shen, Y. The International HapMap Project. *Nature* **2003**, *426*, 789–796.
- Manolio, T.A. Genomewide Association Studies and Assessment of the Risk of Disease. *N. Engl. J. Med.* **2010**, *363*, 166–176. [[CrossRef](#)] [[PubMed](#)]
- Nishino, J.; Ochi, H.; Kochi, Y.; Tsunoda, T.; Matsui, S. Sample Size for Successful Genome-Wide Association Study of Major Depressive Disorder. *Front. Genet.* **2018**, *9*, 227. [[CrossRef](#)]
- Hong, E.P.; Park, J.W. Sample Size and Statistical Power Calculation in Genetic Association Studies. *Genom. Inform.* **2012**, *10*, 117–122. [[CrossRef](#)]
- Ziyatdinov, A.; Kim, J.; Prokopenko, D.; Privé, F.; Laporte, F.; Loh, P.-R.; Kraft, P.; Aschard, H. Estimating the Effective Sample Size in Association Studies of Quantitative Traits. *G3* **2021**, *11*, jkab057. [[CrossRef](#)]
- Hellwege, J.N.; Keaton, J.M.; Giri, A.; Gao, X.; Velez Edwards, D.R.; Edwards, T.L. Population Stratification in Genetic Association Studies. *Curr. Protoc. Hum. Genet.* **2017**, *95*, 1.22.1–1.22.23. [[CrossRef](#)]
- Platt, A.; Vilhjálmsdóttir, B.J.; Nordborg, M. Conditions under Which Genome-Wide Association Studies Will Be Positively Misleading. *Genetics* **2010**, *186*, 1045–1052. [[CrossRef](#)]
- Shen, X.; Carlborg, O. Beware of Risk for Increased False Positive Rates in Genome-Wide Association Studies for Phenotypic Variability. *Front. Genet.* **2013**, *4*, 93. [[CrossRef](#)]
- Klein, R.J.; Zeiss, C.; Chew, E.Y.; Tsai, J.Y.; Sackler, R.S.; Haynes, C.; Henning, A.K.; SanGiovanni, J.P.; Mane, S.M.; Mayne, S.T. Complement factor H polymorphism in age-related macular degeneration. *Science* **2005**, *308*, 385–389. [[CrossRef](#)]
- DeWan, A.; Liu, M.; Hartman, S.; Zhang, S.S.; Liu, D.T.; Zhao, C.; Tam, P.O.; Chan, W.M.; Lam, D.S.; Snyder, M. HTRA1 promoter polymorphism in wet age-related macular degeneration. *Science* **2006**, *314*, 989–992. [[CrossRef](#)] [[PubMed](#)]
- Diez Díaz, F.; Sánchez Lasheras, F.; Moreno, V.; Moratalla-Navarro, F.; Molina de la Torre, A.J.; Martín Sánchez, V. GASVeM: A New Machine Learning Methodology for Multi-SNP Analysis of GWAS Data Based on Genetic Algorithms and Support Vector Machines. *Mathematics* **2021**, *9*, 654. [[CrossRef](#)]
- Ziegler, A.; Ghosh, S.; Dyer, T.D.; Blangero, J.; Maccluer, J.; Almasy, L. Introduction to genetic analysis workshop 17 summaries. *Gen. Epidemiol.* **2011**, *35*, S1–S4. [[CrossRef](#)]
- Lippert, C.; Listgarten, J.; Davidson, R.I.; Baxter, S.; Poon, H.; Cadie, C.M.; Heckerman, D. An exhaustive epistatic SNP association analysis on expanded Wellcome Trust data. *Sci. Rep.* **2013**, *3*, 1099. [[CrossRef](#)] [[PubMed](#)]



15. Ning, C.; Wang, D.; Zhou, L.; Wei, J.; Liu, Y.; Kang, H.; Zhang, S.; Zhou, X.; Xu, S.; Liu, J.F. Efficient multivariate analysis algorithms for longitudinal genome-wide association studies. *Bioinformatics* **2019**, *35*, 4879–4885. [[CrossRef](#)]
16. Schubach, M.; Re, M.; Robinson, P.N.; Valentini, G. Imbalance-Aware Machine Learning for Predicting Rare and Common Disease-Associated Non-Coding Variants. *Sci. Rep.* **2017**, *7*, 2959. [[CrossRef](#)]
17. Lin, H.; Hargreaves, K.A.; Li, R.; Reiter, J.L.; Wang, Y.; Mort, M.; Cooper, D.N.; Zhou, Y.; Zhang, C.; Eadon, M.T. RegSNPs-Intron: A Computational Framework for Predicting Pathogenic Impact of Intronic Single Nucleotide Variants. *Genome Biol.* **2019**, *20*, 254. [[CrossRef](#)]
18. Roshan, U.; Chikkagoudar, S.; Wei, Z.; Wang, K.; Hakonarson, H. Ranking Causal Variants and Associated Regions in Genome-Wide Association Studies by the Support Vector Machine and Random Forest. *Nucleic. Acids Res.* **2011**, *39*, e62. [[CrossRef](#)]
19. Isakov, O.; Dotan, I.; Ben-Shachar, S. Machine Learning-Based Gene Prioritization Identifies Novel Candidate Risk Genes for Inflammatory Bowel Disease. *Inflamm. Bowel Dis.* **2017**, *23*, 1516–1523. [[CrossRef](#)]
20. Deo, R.C.; Musso, G.; Tasan, M.; Tang, P.; Poon, A.; Yuan, C.; Felix, J.F.; Vasan, R.S.; Beroukhim, R.; De Marco, T. Prioritizing Causal Disease Genes Using Unbiased Genomic Features. *Genome Biol.* **2014**, *15*, 534. [[CrossRef](#)]
21. Maciukiewicz, M.; Marshe, V.S.; Hauschild, A.-C.; Foster, J.A.; Rotzinger, S.; Kennedy, J.L.; Kennedy, S.H.; Müller, D.J.; Geraci, J. GWAS-Based Machine Learning Approach to Predict Duloxetine Response in Major Depressive Disorder. *J. Psychiatr. Res.* **2018**, *99*, 62–68. [[CrossRef](#)]
22. Zhou, J.; Theesfeld, C.L.; Yao, K.; Chen, K.M.; Wong, A.K.; Troyanskaya, O.G. Deep Learning Sequence-Based Ab Initio Prediction of Variant Effects on Expression and Disease Risk. *Nat. Genet.* **2018**, *50*, 1171–1179. [[CrossRef](#)] [[PubMed](#)]
23. Storn, R.; Price, K. Differential Evolution—A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces. *J. Glob. Optim.* **1997**, *11*, 341–359. [[CrossRef](#)]
24. Price, R.; Storn, K.; Lampinen, R.M. *Differential Evolution: A Practical Approach to Global Optimization*; Springer: New York, NY, USA, 2005.
25. Huang, G.B.; Wang, D.H.; Lan, Y. Extreme Learning Machines: A Survey. *Int. J. Mach. Learn. Cybern.* **2011**, *2*, 107–122. [[CrossRef](#)]
26. Huang, G.B.; Zhu, Q.Y.; Siew, C.K. Extreme Learning Machine: A New Learning Scheme of Feedforward Neural Networks. In Proceedings of the 2004 IEEE International Joint Conference on Neural Networks, Budapest, Hungary, 25–29 July 2004; IEEE: Piscataway, NJ, USA, 2005.
27. Huang, G.B.; Zhu, Q.Y.; Siew, C.-K. Extreme Learning Machine: Theory and Applications. *Neurocomputing* **2006**, *70*, 489–501. [[CrossRef](#)]
28. Deng, W.; Zheng, Q.; Chen, L. Regularized Extreme Learning Machine. In Proceedings of the 2009 IEEE Symposium on Computational Intelligence and Data Mining, Nashville, TN, USA, 30 March–2 April 2009; IEEE: Piscataway, NJ, USA, 2009.
29. Joshi, G.P.; Alenezi, F.; Thirumoorthy, G.; Dutta, A.K.; You, J. Ensemble of Deep Learning-Based Multimodal Remote Sensing Image Classification Model on Unmanned Aerial Vehicle Networks. *Mathematics* **2021**, *9*, 2984. [[CrossRef](#)]
30. Gupta, U.; Gupta, D. Regularized Based Implicit Lagrangian Twin Extreme Learning Machine in Primal for Pattern Classification. *Int. J. Mach. Learn. Cybern.* **2021**, *12*, 1311–1342. [[CrossRef](#)]
31. Prakapenka, D.; Liang, Z.; Jiang, J.; Ma, L.; Da, Y. A Large-Scale Genome-Wide Association Study of Epistasis Effects of Production Traits and Daughter Pregnancy Rate in U.S. Holstein Cattle. *Genes* **2021**, *12*, 1089. [[CrossRef](#)]
32. Gondro, C.; van der Werf, J.; Hayes, B. *Genome-Wide Association Studies and Genomic Prediction*; Methods in Molecular Biology; Humana Press: New York, NY, USA, 2013.
33. Marozzi, M. A bi-aspect nonparametric test for the two-sample location problem. *Comput. Stat. Data. Anal.* **2002**, *64*, 639–648. [[CrossRef](#)]
34. Marozzi, M. Some remarks about the number of permutations one should consider to perform a permutation test. *Statistica* **2004**, *64*, 193–201.
35. Browning, B.L. PRESTO: Rapid calculation of order statistic distributions and multiple-testing adjusted P-values via permutation for one and two-stage genetic association studies. *BMC Bioinform.* **2008**, *9*, 309. [[CrossRef](#)] [[PubMed](#)]
36. Kanehisa, M.; Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic. Acids. Res.* **2000**, *28*, 27–30. [[CrossRef](#)] [[PubMed](#)]
37. Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. *Protein. Sci.* **2019**, *28*, 1947–1951. [[CrossRef](#)] [[PubMed](#)]
38. Kanehisa, M.; Furumichi, M.; Sato, Y.; Ishiguro-Watanabe, M.; Tanabe, M. KEGG: Integrating viruses and cellular organisms. *Nucleic. Acids. Res.* **2021**, *49*, D545–D551. [[CrossRef](#)]
39. Liu, X.; Gao, C.; Li, P. A comparative analysis of support vector machines and extreme learning machines. *Neural Netw.* **2012**, *33*, 58–66. [[CrossRef](#)] [[PubMed](#)]
40. Cheng, G.J.; Cai, L.; Pan, H.X. Comparison of extreme learning machine with support vector regression for reservoir permeability prediction. In Proceedings of the 2009 International Conference on Computational Intelligence and Security, Beijing, China, 11–14 December 2009; IEEE: Piscataway, NJ, USA, 2009; Volume 2, pp. 173–176.
41. Huang, G.B.; Ding, X.; Zhou, H. Optimization method based extreme learning machine for classification. *Neurocomputing* **2010**, *74*, 155–163. [[CrossRef](#)]
42. Price, K.V.; Storn, R. Differential evolution—A simple evolution strategy for fast optimization. *Dr. Dobbs. J.* **1997**, *22*, 18–24.

43. Tušar, T.; Filipi, B. Differential Evolution versus Genetic Algorithms in Multiobjective Optimization. In *Evolutionary Multi-Criterion Optimization, Matsushima, Japan, 2007*; Obayashi, S., Deb, K., Poloni, C., Hiroyasu, T., Murata, T., Eds.; Springer: Berlin/Heidelberg, Germany, 2007; Volume 4403.
44. Thomas, M.; Sakoda, L.C.; Hoffmeister, M.; Rosenthal, E.A.; Lee, J.K.; van Duijnhoven, F.J.B.; Platz, E.A.; Wu, A.H.; Dampier, C.H.; de la Chapelle, A. Genome-Wide Modeling of Polygenic Risk Score in Colorectal Cancer Risk. *Am. J. Hum. Genet.* **2020**, *107*, 432–444. [[CrossRef](#)]
45. Yang, Y.; Lv, S.Y.; Ye, W.; Zhang, L. Apelin/APJ system and cancer. *Clin. Chim. Acta* **2016**, *457*, 112–116. [[CrossRef](#)]
46. Picault, F.X.; Chaves-Almagro, C.; Progetti, F. Tumour co-expression of apelin and its receptor is the basis of an autocrine loop involved in the growth of colon adenocarcinomas. *Eur. J. Cancer* **2014**, *50*, 663–674. [[CrossRef](#)]
47. Mughal, A.; O'Rourke, S.T. Vascular effects of apelin: Mechanisms and therapeutic potential. *Pharmacol. Ther.* **2018**, *190*, 139–147. [[CrossRef](#)] [[PubMed](#)]
48. Podgórska, M.; Diakowska, D.; Pietraszek-Gremplewicz, K.; Nienartowicz, M.; Nowak, D. Evaluation of Apelin and Apelin Receptor Level in the Primary Tumor and Serum of Colorectal Cancer Patients. *J. Clin. Med.* **2019**, *8*, 1513. [[CrossRef](#)] [[PubMed](#)]
49. Podgórska, M.; Pietraszek-Gremplewicz, K.; Olszanska, J.; Nowak, D. The Role of Apelin and Apelin Receptor Expression in Migration and Invasiveness of Colon Cancer Cells. *Anticancer Res.* **2021**, *41*, 151–161. [[CrossRef](#)]
50. Sanchez Pino, M.J.; Moreno, P.; Navarro, A. Mitochondrial dysfunction in human colorectal cancer progression. *Front. Biosci.* **2007**, *12*, 1190–1199. [[CrossRef](#)] [[PubMed](#)]
51. Guo, B.; Han, X.; Tkach, D.; Huang, S.G.; Zhang, D. AMPK promotes the survival of colorectal cancer stem cells. *Anim. Models Exp. Med.* **2018**, *1*, 134–142. [[CrossRef](#)] [[PubMed](#)]
52. Khabaz, M.N.; Abdelrahman, A.S.; Al-Maghrabi, J.A. Expression of p-AMPK in colorectal cancer revealed substantial diverse survival patterns. *Pak. J. Med. Sci.* **2019**, *35*, 685–690. [[CrossRef](#)] [[PubMed](#)]
53. Wu, Z.; Liu, Z.; Ge, W.; Shou, J.; You, L.; Pan, H.; Han, W. Analysis of potential genes and pathways associated with the colorectal normal mucosa-adenoma-carcinoma sequence. *Cancer Med.* **2018**, *7*, 2555–2566. [[CrossRef](#)]
54. Yagi, T.; Kubota, E.; Koyama, H.; Tanaka, T.; Kataoka, H.; Imaeda, K.; Joh, T. Glucagon promotes colon cancer cell growth via regulating AMPK and MAPK pathways. *Oncotarget* **2018**, *9*, 10650–10664. [[CrossRef](#)]
55. Murmann, A.E.; Gao, Q.Q.; Putzbach, W.E.; Patel, M.; Bartom, E.T.; Law, C.Y.; Bridgeman, B.; Chen, S.; McMahan, K.M.; Thaxton, C.S.; et al. Small interfering RNA s based on huntingtin trinucleotide repeats are highly toxic to cancer cells. *EMBO Rep.* **2018**, *19*, e45336. [[CrossRef](#)]
56. Pechlivanis, S.; Pardini, B.; Bermejo, J.L.; Wagner, K.; Naccarati, A.; Vodickova, L.; Novotny, J.; Hemminki, K.; Vodicka, P.; Försti, A. Insulin pathway related genes and risk of colorectal cancer: INSR promoter polymorphism shows a protective effect. *Endocr.-Relat. Cancer* **2007**, *14*, 733–740. [[CrossRef](#)]
57. Carvalho, D.V.; Pereira, E.M.; Cardoso, J.S. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics* **2019**, *8*, 832. [[CrossRef](#)]
58. Aslam, N. Explainable Artificial Intelligence Approach for the Early Prediction of Ventilator Support and Mortality in COVID-19 Patients. *Computation* **2022**, *10*, 36. [[CrossRef](#)]
59. Goebel, R.; Chander, A.; Holzinger, K.; Lecue, F.; Akata, Z.; Stumpf, S.; Kieseberg, P.; Holzinger, A. Explainable AI: The New 42? In *Machine Learning and Knowledge Extraction*; Springer: Cham, Switzerland, 2018; pp. 295–303.
60. Kuncheva, L.I. *Combining Pattern Classifiers: Methods and Algorithms*; John Wiley & Sons: Hoboken, NJ, USA, 2014.
61. Iwendi, C.; Bashir, A.K.; Peshkar, A.; Sujatha, R.; Chatterjee, J.M.; Pasupuleti, S.; Mishra, R.; Pillai, S.; Jo, O. COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm. *Front. Public Health* **2020**, *8*, 357. [[CrossRef](#)] [[PubMed](#)]
62. Rezaei-Ravari, M.; Eftekhari, M.; Saberi-Movahed, F. Regularizing extreme learning machine by dual locally linear embedding manifold learning for training multi-label neural network classifiers. *Eng. Appl. Artif. Intell.* **2021**, *97*, 104062. [[CrossRef](#)]