Doctoral Programme in Theoretical Chemistry and Computational Modelling

Universitat Jaume I – Doctoral School

Modeling of proteins

Report submitted by Gerardo Alfonso Pérez in order to be eligible for a doctoral degree awarded by the Universitat Jaume I

Gerardo Alfonso Pérez	Dr. Raquel Castillo
Firmado por ALFONSO PEREZ GERARDO - 78559838A el día 07/09/2023 con un certificado emitido por AC FNMT Usuarios	RAQUEL| CASTILLO| SOLSONA Firmado digitalmente por RAQUEL|CASTILLO|SOLSONA Fecha: 2023.09.07 11:51:58 +02'00'
(PhD Candidate)	(Supervisor)

Castelló de la Plana (Spain), July 2023

Funding (grants):

- Spanish Ministerio de Ciencia, Innovación y Universidades (PGC2018-094852-B-C21)
- Spanish Ministerio de Ciencia y Tecnología (PID2021-1233320B-C21)
- Universitat Jaume I (UJI-B2019-43).
- Universitat Jaume I (UJI-B2022-12).

List of publications

Publications in indexes journals (three Q1 and one Q2).

1. Gerardo Alfonso Perez, Raquel Castillo.

   Categorical Variable Mapping Considerations in Classification Problems: Protein Application.

   Mathematics. 2022,11(279). I.F.=2.4. Q1.

   https://doi.org/10.339

2. Gerardo Alfonso Perez, Raquel Castillo.

   Nonlinear Techniques and Ridge Regression as a Combined Approach: Carcinoma Identification Case Study

   Mathematics. 2023,11(11795).I.F.=2.4. Q1.

   https://doi.org/10.3390/math11081

3. Gerardo Alfonso Perez, Raquel Castillo.

   Identification of Systemic Sclerosis through Machine Learning Algorithms and Gene Expression

Mathematics. 2022,10(4632). I.F.=2.4. Q1.

https://doi.org/10.3390/math102446

4. Gerardo Alfonso Perez, Raquel Castillo.

   Gene Identification in Inflammatory Bowel Disease via a Machine Learning Approach

   Medicina. 2023,59(1218). I.F.=2.6. Q2.

   doi.org/10.3390/medicina59071218

This thesis has been accepted by the co-authors of the publications listed above that have waved the right to present them as a part of another PhD thesis.

Abstract

In the first part of this dissertation a series of assumptions regarding the categorical mapping of amino acids in the context of protein classification using machine learning techniques are tested. The four assumptions tested: (1) translation, (2) permutation, (3) constant, and (4) eigenvalues were validated with experimental data. The first three assumptions relate to equivalent mappings, and the fourth involves a comparable mapping using a proposed eigenvalue-based matrix representation of the amino acid chain. These assumptions were tested across a range of 23 different machine learning algorithms. It is shown that the numerical simulations are consistent with the presented assumptions, such as translation and permutations, and that the eigenvalue approach generates classifications that are statistically not different from the base case or that have higher mean values while at the same time providing some advantages such as having a fixed predetermined dimensions regardless of the size of the analyzed protein.

Then, it was shown that it is possible to accurately distinguish, using non-linear techniques, between healthy patients and anal and carcinoma patients using DNA methylation data as input. The model selected 13 CpGs from a total of 450,000 CpGs available per patient with 171 patients in total. The model was also tested for robustness and compared to other more complex models that generated less precise classifications. The model obtained (testing dataset) an accuracy, sensitivity and specificity of 97.69%, 95.02% and 98.26%, respectively. The reduction of the dimensionality of the data, from 450,000 to 13 CpGs per patient, likely also reduced the likelihood of overfitting, which is a very substantial risk in this type of modelling.

Interstitial lung disease systemic sclerosis (ILD-SSc) was also studied. We present an algorithm (using machine learning techniques) that it is able to identify, with a 92.2% accuracy, patients suffering from ILD-SSc using gene expression data obtained from peripheral blood. The algorithm also identified 172 genes that might be involved in the illness. These 172 genes appeared in all the 20 most accurate classification models among a total of half a million models estimated. Their frequency might suggest that they are related to the illness to some degree. The proposed algorithm, besides differentiating between control and patients, was also able to distinguish among different variants of the illness (diffuse variants). This can have a significance from a treatment point of view. The different type of variants have a different associated prognosis.

In the last part of the dissertation Inflammatory bowel disease (IBD) was also analyzed. The illness is rather heterogeneous with different evolution among patients. A machine learning approach was followed to identify potential genes that are related to IBD. This is done by following a Monte Carlo simulation approach. In total, 23 different machine learning techniques were tested (in addition to a base level obtained using artificial neural networks). The best model identified 74 genes selected by the algorithm as being potentially involved in IBD. IBD seems to be a polygenic illness, in which environmental factors might play an important role. Following a machine learning approach, it was possible to obtain a classification accuracy of 84.2% differentiating between patients with IBD and control cases in a large cohort of 2490 total cases. The sensitivity and specificity of the model were 82.6% and 84.4%, respectively. It was also possible to distinguish between the two main types of IBD: (1) Crohn's disease and (2) ulcerative colitis.

Dedication

I need to thank a lot of people for their support and encouragement during these years. First I would like to thank my parents, Antonio and Maria del Pilar, for their support, not only during the PhD years but also as an undergraduate student and before. I also need to thank my brothers, Pilar, Javi and Jose. They have been also very supportive during this process. I need to specially thank my wife Wandi and kids that are always a source a support and have been very understanding.

Acknowledgements

I have been very lucky in this dissertation of having Prof. Raquel Castillo as the my PhD supervisor without her support and advice this dissertation would not have been possible.

Table of Contents

Chapter 1

Introduction

In recent years there has been a rapid increase in the amount of chemical [1–3] and biological data [4–6]. This new data has enable researches to advance their fields but it has also come with challenges [7–9], such as the need to develop techniques to handle this large volume of information. This would be rather challenging using traditional techniques. In this regard, artificial intelligence techniques present an interesting alternative [10]. Many artificial intelligence techniques require large amount of information in order to be able to produce accurate forecasts and hence seems a natural approach [11, 12]. There are many lines of investigation in which this approach can be followed, such as protein classification.

1.1 Protein classification using machine learning techniques

Machine learning techniques are having an increasingly important role in the computational chemistry [13–19] and other scientific related fields [20–24]. This is partially due to the capability of these techniques to model complex underlying processes [25–27] without the need to have a detailed understating of the underlying mechanics of the process as well as in situations were the underlying dynamics is well understood but the computational cost [28–30] of using a detailed models are too high. There is a rapidly increasing number of publications using machine learning techniques in the context of protein modeling and classification. For instance, there are some interesting articles, such as Xu et al. [31], describing sequence and activity relationships, particularly focusing on mutations. Another topic analyzed in the existing literature is the classification of cell decisions for protein Kinase B, Salau et al. [32], using neural networks such as radial basis functions (RBFs) and multi-layer perceptron (MLP). Neural networks are among some of the frequently used techniques in classification tasks. It should be noted that there is a large number of machine learning techniques [33, 34] that can be used for classification purposes, such as the above mentioned artificial neural networks (ANN), as well as other techniques such as support vector machines (SVM) [35, 36] and K-nearest neighbors (KNN) [37]. These tools will be explained in more detail later in this chapter.

The complexity of the protein classification task has been mentioned by authors such as McDowall and Hunter [38] and Nanni et al. [39]. Furthermore, some authors such as Diplaris et al. [40] mentioned that there is an actual need for machine learning techniques in this type of modeling exercises. This type of modeling is typically referees as “big data” [41, 42]. Experimental results have created very large databases containing increasingly large amount of information on proteins. In figure 1.1 it can be seen an example of a cyclic protein. This increasing amount of information is clearly a positive but it

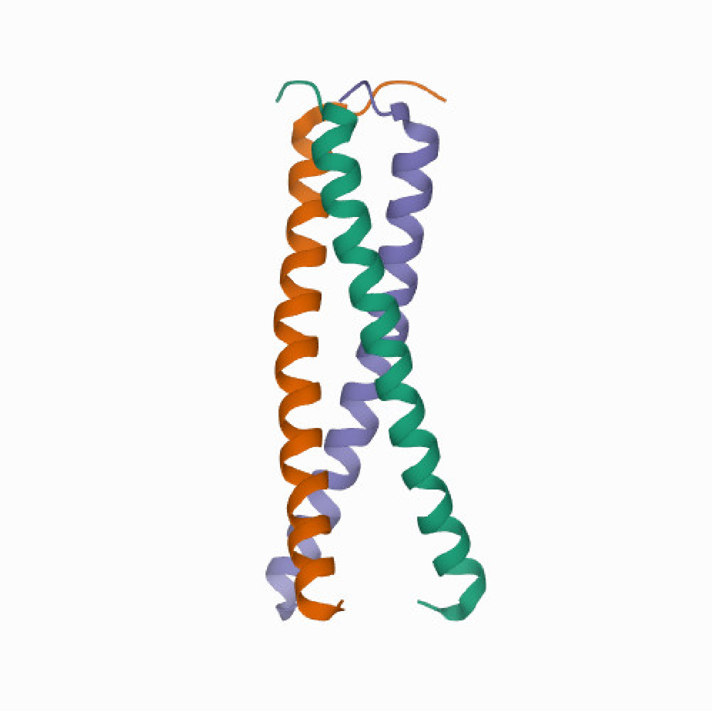
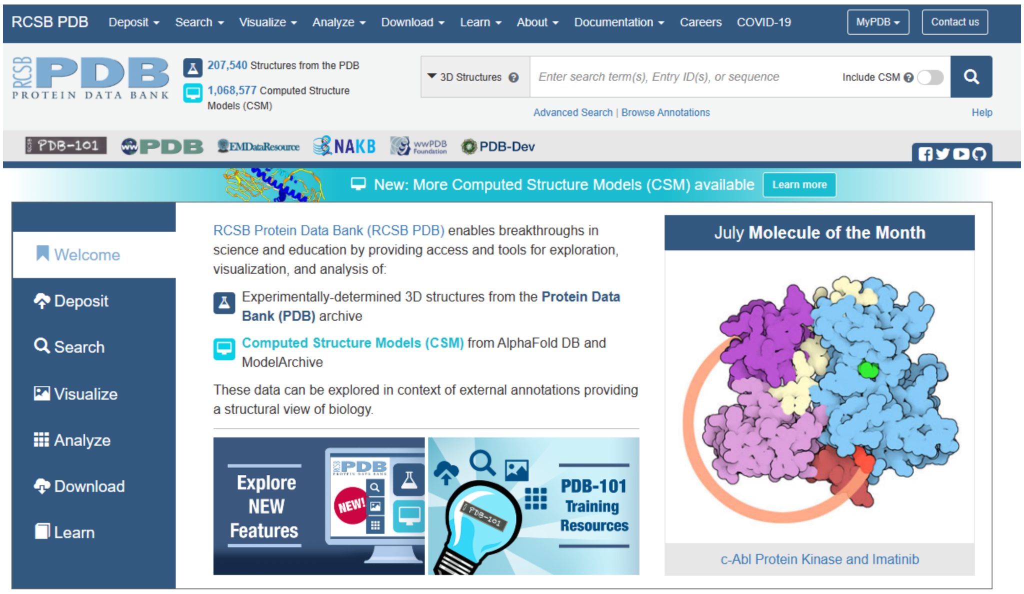


Figure 1.1: Example of a cyclic protein (4GIF). [43]

requires appropriate modeling techniques to model this vast amount of data.

In this dissertation the information regarding protein primary structure and classification was obtained from the Protein Data Bank (PDB) [43], which is a well-known data repository. An screenshot of the PBD website can be seen in figure 1.2.

Artificial intelligence techniques are not without its issues. One of the most frequently mentioned issue is that they tend to create models that are rather complex to interpret [44–47].



207,540 Structures from the PDB

1,068,577 Computed Structure Models (CSM)

New: More Computed Structure Models (CSM) available

RCSB Protein Data Bank (RCSB PDB) enables breakthroughs in science and education by providing access and tools for exploration, visualization, and analysis of:

- Experimentally-determined 3D structures from the **Protein Data Bank (PDB)** archive
- **Computed Structure Models (CSM)** from AlphaFold DB and ModelArchive

These data can be explored in context of external annotations providing a structural view of biology.

Figure 1.2: Protein Data Bank (PDB). [43]

The structure of an ANN can be composed by several layers of artificial neurons, with each neuron having an associated weight (which is the result of the training process). It might be difficult to interpret the chemical or biological meaning of each of these components of the mathematical model. Therefore, when using these models there tends to be a tradeoff between the accuracy of the model and the interpretability of the model.

One of the focus areas of this dissertation is the classification task of small proteins [48–52] in the context of numerical simulations regarding some assumptions in the mapping of categorical values [53, 54], which is an issue directly related to protein modeling [55], see figure 1.3. The input in this type of simulations is frequently the chain of amino acids. This is frequently expressed in the form of a letter, identifying each amino acid, in other words it is a categorical variable. The length of this variable will depend on the number of amino acids. The mapping of this categorical variables should respect some biological basic considerations, such as translation and permutation, which will be formally defined in later sections. An eigenvalue representation of this categorical variables was also tested numerically, obtaining accurate classification values. This approach has some computational advantages, which will be further explained. To the best of our knowledge this eigenvalue approach for categorical variables in protein modeling has never been done before.

1.1.1 Computational challenges

There are a large number of computational challenges [56] associated with protein classification [57, 58] using machine learning techniques. For instance, the length of the amino acid chain can wary substantially among different proteins. This

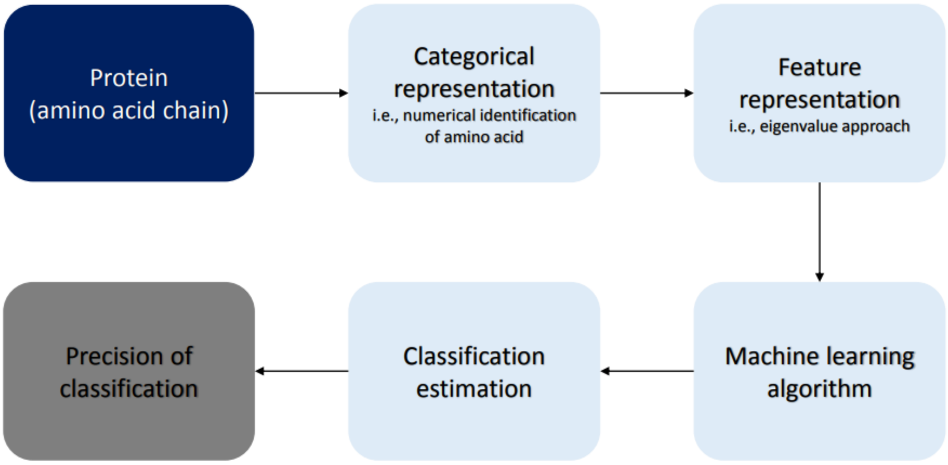

Protein
(amino acid chain)
Categorical representation
i.e., numerical identification of amino acid
Feature representation
i.e., eigenvalue approach
Machine learning algorithm
Classification estimation
Precision of classification

Figure 1.3: Approach flowchart.

represent a challenge from a computational point of view given that many models assume that the number of features per case (protein) is the same. As will be illustrated letter in this dissertation we have tried to go around this problem in several ways, such as an eigenvalue approach.

Other issue relates to computational power and time [59–61]. Many of the proposed models, such as artificial neural networks, require significant computational time for training purposes [62–64], particularly when handling large data bases [19] of proteins. This type of problem can be reduce, to some degree, by using parallel programming [65] when possible. It should be taken into consideration that some artificial intelligence techniques do not allow for straight forward parallelization. As previously mentioned, the issue of categorical variable [66] use (to describe the amino acids) in the context of protein classification will be thoroughly analyzed later in this dissertation.

1.2 Non-linear identification of carcinoma

Cancer is among the major causes of mortality [67–69] with carcinomas accounting for a substantial proportion of cancer related deaths. Eng et al. [70] calculated that the mortality rate is 3.1%. According to some estimates, Deshmukh et al. [71], the incidence of some types of carcinomas, such as anal carcinoma, is actually increasing at substantial annual rate (2.7% yearly). In recent years there has been a rapid increase in the amount of genetic information, such as DNA methylation levels [72, 73], available related to different types of cancers (including carcinomas).

Despite this increase in data available it is frequently mentioned in the existing literature, Monsrud et al. [74], that carcinomas, such as anal and cervical carcinomas, are not yet well understood. There are some articles in the existing literature, such as Zhang et al. [75], analyzing changes in DNA methylation levels in anal carcinoma patients. The authors concluded that resulted in abnormal DNA methylation profiles. Other articles, such as Siegel et al. [76] expanded this approach to cover both cervical and anal carcinomas.

DNA methylation [77–79] is an epigenetic process [80–82] in which a methyl group is added in a DNA chain. This is typically analyzed at a CpG level. i.e., a zone in the DNA chain in which a cytosine is followed by a guanine. DNA methylation levels are typically expressed as a percentage ranging from 0% (no methylation) to 100% (fully methylated), for each CpG. It should be mentioned that changes in DNA methylation are part of the natural process of aging [83, 84] and hence there is some variation of methylation levels across different age groups. This type of changes in DNA methylation levels have been used to estimate the biological each of individuals and also to differentiate between different healthy patients and patients with some illnesses such as different types of cancers. Some examples in the existing literature include glioblastoma, by Siegel et al. [85], and lung carcinomas, by Marchevsky [86].

1.3 Identification of systemic sclerosis

Systemic sclerosis (SSc) is an autoimmune chronic illness Sapadin and Fleischmajer [87]. Currently there is no curative treatment for SSc and it has an associated high morbidity and mortality [88, 89]. In line with other autoimmune illnesses it is more frequent in females [90]. Some studies, such as Zhong et al. [91], have estimated that the illness has a prevalence in the United Sates of approximately 50 cases per 100,000 individuals. Some ethnic groups, such a Native Americans, have a higher prevalence, Mayes et al. [92]. The typical onset age is between 30 to 60 years old, Hoffmann-Vold et al. [93]. SSc is characterized by excessive collagen content in tissue, as well as fibrosis and vascular damage [94–96].

The causes of SSc are not well understood, with mentions in the existing literature to both genetic factors, Ingegnoli et al. [97], and environmental triggers, Marie [98]. There is evidence of some professions, such as Silica miners [99], increasing the likelihood of having SSc. Patients with SSc frequently have complications. One of the most frequent severe complications is interstitial lung disease (ILD). Interstitial lung disease (ILD) substantially worsens the prognosis of SSc. While, as previously mentioned, there is no curative treatment for the illness there are some treatment options for some of the common complications for renal crisis (using for instance ACE inhibitors).

The evolution of patients with SSc can be substantially different [90]. It is not well understood the reason behind this heterogeneous evolution landscape of the illness across different patients and might be related to a combination of genetic and environmental factors. There are two main types of SSC: 1) Limited Cutaneous Systemic Sclerosis (also referred as CREST [100, 101]) and 2) Diffuse Systemic Sclerosis [102, 103]. It should be mentioned that there is some disagreement in the existing literature regarding the variants of the illness.

In this dissertation it is presented a machine learning algorithm that it is able to identify, with a 92.2% accuracy, patients ILD-SSc patients using gene expression data obtained from peripheral blood. This type of approach might we used in the future, when more data is available, to developed personalized treatments.

1.4 Gene Identification in inflammatory bowel disease

Inflammatory bowel disease (IBD) is a chronic inflammatory disease [104–106]. IBD remains as a not well understood illness [107–109] with patients showing a different array of symptoms and having different evolution. Some of the most common symptoms are pain, fatigue, cramps, diarrhea and blood in stools [110–112]. The illness appears to have a higher prevalence in urban areas [113]. The illness is becoming an increasing important health issue as its prevalence is increasing in newly industrialized countries [114]. The reasons behind this increase in prevalence in newly industrialized countries remains unclear but it might be related to changes in lifestyles as well as other environmental factors. It is also possible that the illness is detected more accurate in these countries as their economic development allows for better national healthcare systems.

One of the main hypothesis about the illness is that it is an abnormal immune response, triggered by some type of environ-

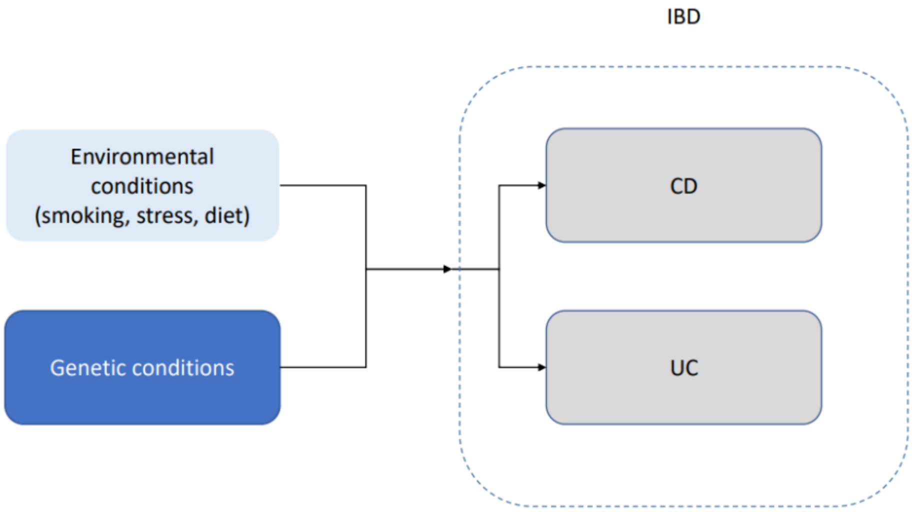


Figure 1.4: Schematic representation of the interaction between genetic predisposition and environmental factors in ulcerative colitis (UC) and Crohn's disease (CD).

mental factor, such as a bacteria or virus [115–118], in genetically predisposed individuals (see figure 1.4). Other lifestyle and environmental factors, such as smoking and diet, seem to also play a role in the illness [119]. IBD typically harms the mucosa [120–122]. There are two main types of IBD: Ulcerative Colitis (UC) [123–125] and Chron's Disease (CD) [126–128].

A genetic component is frequently mentioned in the existing literature. For example, Khor et al. [129], mentioned that genes help regulate the complex interaction between microbial and environmental factors. Other suggestion that there is a genetic component comes for some ethnic groups having a higher prevalence. An example are the Ashkenazim, which have a higher incidence and prevalence [130]. One of the objectives of the dissertation is to identify genes that play a role in IBD. This is done by using a machine learning approach and gene expression data.

1.5 Artificial intelligence

In this dissertation we do extensive use of artificial intelligence techniques. Artificial intelligence techniques can be broadly divided into the two categories of supervised [131–134] and unsupervised learning [135–138] typically targeting problems of:

1. Forecasting [139–141]
2. Classification [142, 142, 143]
3. Clustering [144–146]

- Forecasting

  In forecasting problems the objective of the artificial intelligence technique is to estimate the value of a signal. This could be for example, the heart rate of a patient [147, 148] or the average daily rainfall amount in a specific area [149, 150]. Forecasting task include time series analysis. Typically this is done, in the context of artificial intelligence techniques, by providing certain inputs to an algorithm that then generates a forecast [151–153]. For

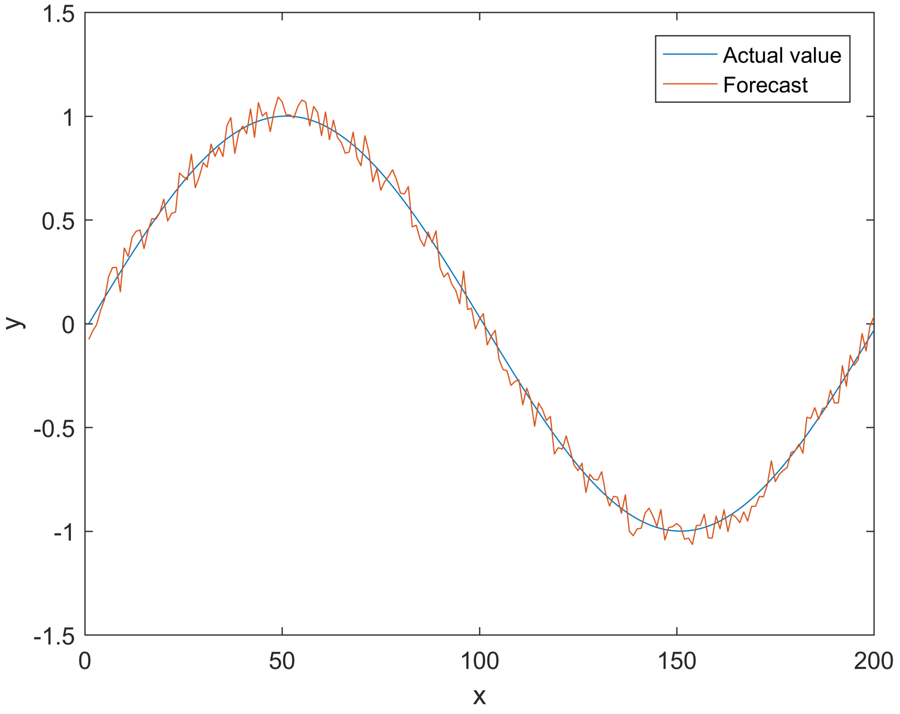


Figure 1.5: Graphical representation of forecasting model.

example, in the case of the heart rate of a patient we could use as inputs the heart rate of the patient in previous moments $(t-1, t-2, t-3, \ldots)$ to forecast the heart rate of the patient in the current moment (t). A graphical representation of a forecasting model can be seen in figure1.5 Forecasting is typically done using supervised learning.

- Classification

  In classification problems the objective is to identify members of different classes. For example, we might we interested in distinguishing between healthy patients and individuals suffering from a given illness using as an input certain medical parameters [154, 155]. Classification problems are also typically done using supervised learning. As will be explained later in the section, supervised learning assumes that we have a set of data in which we know the accurate (real) classification of the individuals.

- Clustering

  Clustering is a task in which cases are group into different categories [156, 157]. This is a different task from classification. Clustering problems are done typically using unsupervised learning [158, 159]. In other words, it is not know (before using the algorithm) which cases belong to which category. The objective of the algorithm is to group these data looking for some similarities of the attributes.

As previously mentioned, one of the main differences between supervised and unsupervised learning is that in supervised learning we have the correct classification or value for some of the cases analyzed and in unsupervised learning we do not have such information [160]. Hence, conceptually these are rather different approaches. In supervised learning the algorithm is trained by iterative altering of some parameters in order to generate forecasts as close as possible to the actual values. Then, after the algorithm is trained, new data is feed to the algorithm and it creates forecasts [161, 162]. In unsupervised learning the algorithm looks for the attributes of different cases and tries to group them (without knowing actual category information). For example, it can find clusters of people with similarities analyzing as inputs their weight, height and age. This type of clustering could be useful when there is no previous information.

All these techniques are quantitative techniques i.e., automated processes. This has advantages and disadvantages. One of the advantages is that the process is objective [163]. When the researcher has selected the technique to be used and the parameters for that technique the results are generated in an automated fashion. This could limit biases in the analysis. At the same time, this is also a disadvantage as it might be difficult

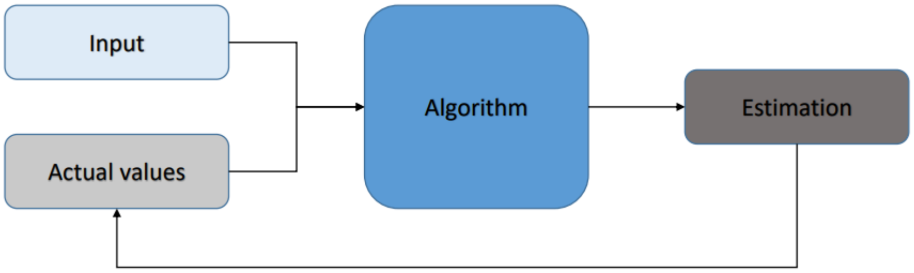


Figure 1.6: Supervised learning model.

to introduce expert classification views by the researcher.

In this dissertation we have used supervised learning techniques (mostly for classification purposes), see figure 1.6. This was possible because there is a large amount of publicly available information classifying different types of proteins and distinguishing between healthy patients and patients with some illness using genetic information. Given this large amount of information a supervised learning approach seemed reasonable. There are many different types of supervised learning tools.

Some of the supervised techniques that we have used in this dissertation include:

1. Artificial neural networks [164–166]

2. Support vector machines [167–170]

3. K-nearest neighbors [171–173]

- Artificial neural networks

  Artificial neural networks [174–178] are a set of well-known artificial intelligence techniques, inspired on the functioning of a brain. The basic idea is to conceptually replicate the functioning of a neuron. In this way the concept of an artificial neuron was created. An artificial neuron is a mathematical function with an associated weight. This function receives some numerical input and generates a numerical result. This numerical result can be altered down or up by modifying the associated weight. Typically an artificial neural network is composed by multiple artificial neuron arranged in layers.

  An artificial neural network requires a training algorithm [179, 180]. The task of the training algorithm is to modify (train) the algorithm to generate an output as close as possible to the actual value. In each iteration the training algorithm changes the weight to reduce the error between the forecast and the actual value. Clearly, there are many different types of training algorithm but conceptually their task is the same. There are some practical considerations to take into account. For instance, the time required to train a network might be substantial and it increases with the number of neurons. In simple terms the more neurons included in the model the more time that it would require for the model to be trained. Furthermore, it is not simple to estimate the number of iterations required by the model to achieve a certain precision. Therefore, usually a maximum number of iterations is defines (to avoid the algorithm entering into a loop that takes too long).

Another factor to take into considerations in that these models have some inherit randomness derived from the initialization of the weights. These models require an initial value for the weights of the neurons. This is done by using a random number generator. These initial values might have a substantial impact on the time required by the network to reach a certain error. Similarly, this also implies that the same network configuration will generate slightly different outputs because of the different initial values of the weights. A typical artificial neural network

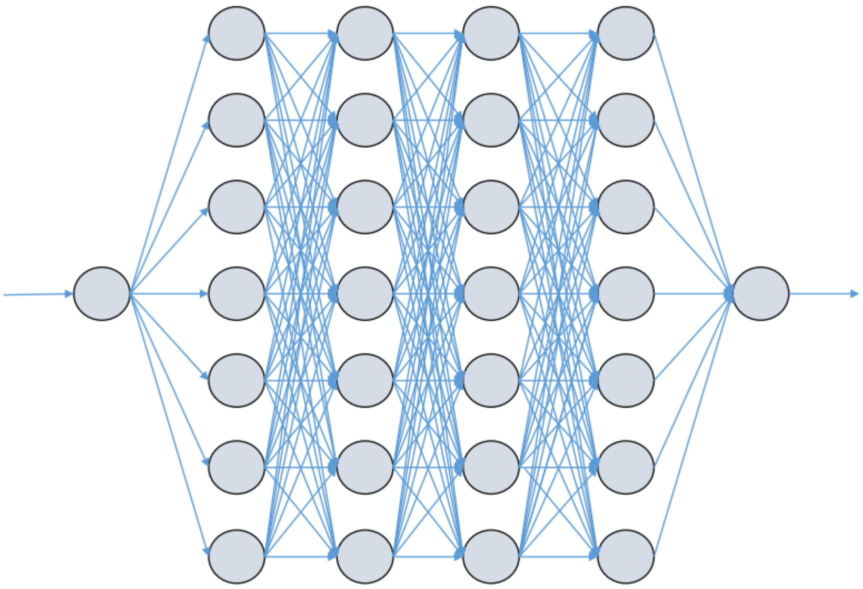


Figure 1.7: Artificial neural network (fully connected) example. For clarity purposes only the connections between the artificial neurons are shown (the weights of each neuron are not shown).

can be seen in figure 1.7.

In the example illustrated in figure 1.7 the artificial neural network is fully connected. In other words each artificial neuron in one layer is connected to all the artificial neurons in the following layers. This is a common network configuration but there are many alternative, such as partial connections. In this type of configurations the neurons in one layers are only connected to some of the neurons in the next layer, see figure 1.8, 1.9 and 1.10 as examples of

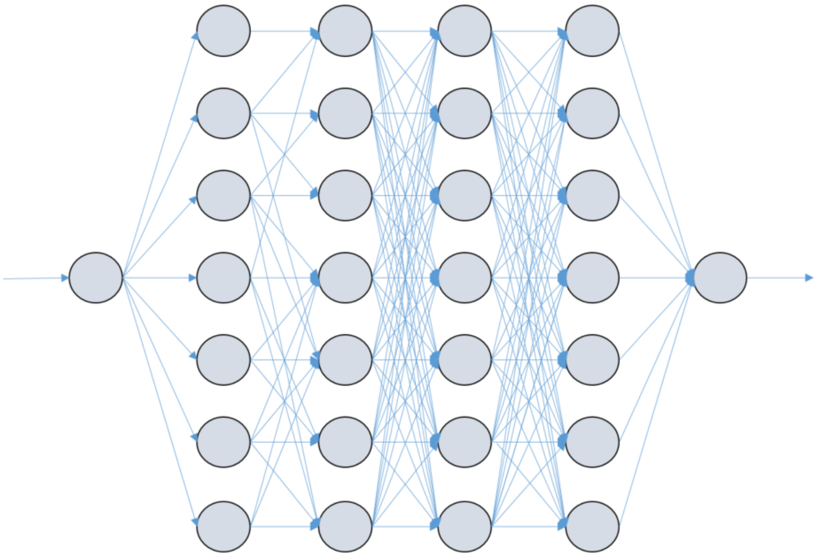


Figure 1.8: Artificial neural network (partially connected) example.

partially connected networks.

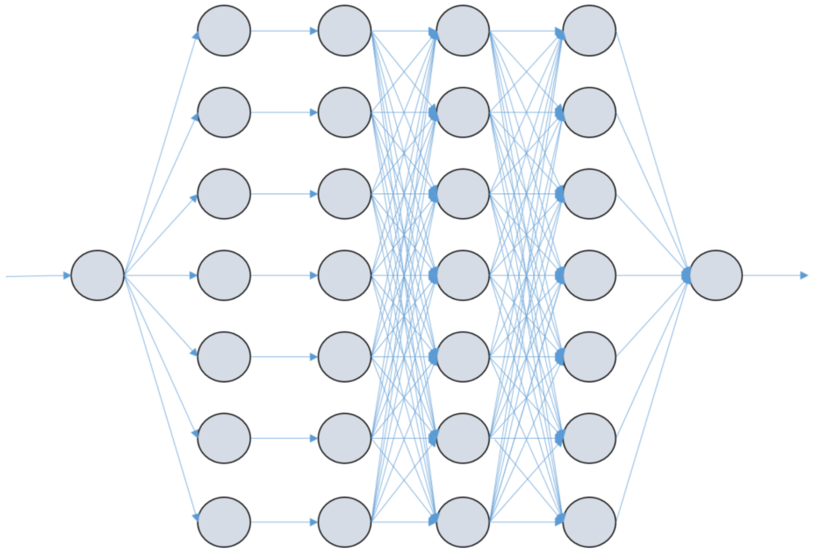


Figure 1.9: Artificial neural network (partially connected) example.

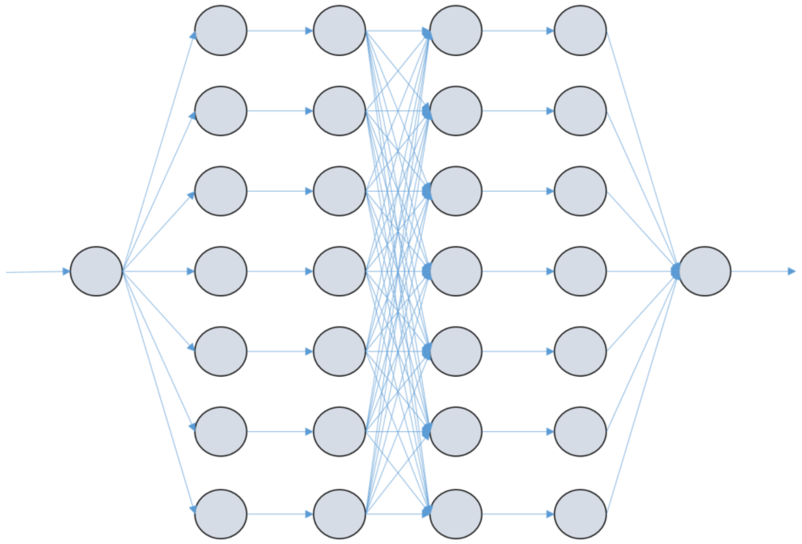


Figure 1.10: Artificial neural network (partially connected) example.

- Support vector machines

  Support vector machines [181–183] is another well-known set of supervised learning techniques. In this case the approach followed is to divide the data into different hyperplanes [184, 185]. This is easy to visualize in a 2D example (see figure 1.11). For simplicity purposes we can visualize a set of points in a sheet of paper. It would be straight forward to see if those points are inside or outside a given circle in the paper of radius r. This would be the case in which the data only have one feature but usually the data have multiple features. Hence, hyperplanes are required to separate the data. There might be situations in which it is not possible to separate the data using hyperplanes.

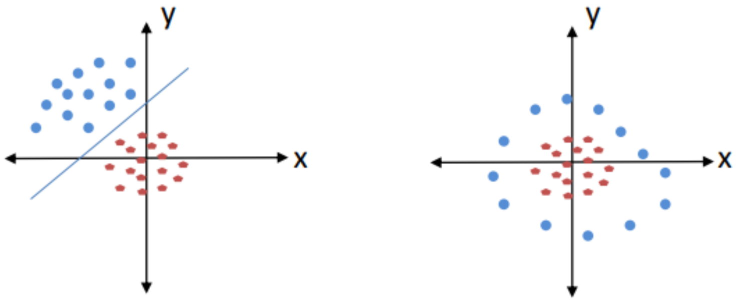


Figure 1.11: Dividing data into hyperplanes (2D model).

- K-nearest neighbors

  This is another useful technique [186–188]. It is based in the idea of the distance between the cases. The algorithm estimates the distance between the cases and the grouping the neighbors according to this distance. While this approach seems straightforward there are multiple considerations to take into account. Firstly, there exist many different types of distance measures [189]. Perhaps the most common is the Euclidean distance but this might not be always the best option. There are other alternatives such as the Minkowski distance [190–192]. Also, the number of k neighbors to be taken into account needs to be considered.

It is a difficult task to know, before doing the analysis, which one of these techniques will generate better results [193–195]. It is then common in practice to test different techniques. This is however very time consuming as there are multiple techniques with multiple configurations to be modeled.

1.6 Software

In order to tackle the previously mentioned problems, such as protein classification and carcinoma detection, via artificial intelligence techniques we have extensively used different coding languages including:

- Matlab [196]
- Phyton [197]
- R (Bioconductor) [198]

As well as many software packages, such as:

- Orca [199]
- Gaussian [200]
- Gauss View [201]
- PyMol [202]
- Avogadro [203]
- NWChem [204]
- Galaxy (FastQC) [205]

- AmberTools (LeaP) [206]
- BLAST [207]
- MEGA [208]
- Open Babel [209]
- Chemdraw [210]
- CP2K [211]
- UCSC genome browser [212]

CHAPTER 2

OBJECTIVES

These are the main objectives of this PhD dissertation:

1. One of the main objectives is testing a series of four assumptions related to the mapping of categorical variables describing the amino acid chains in the context of protein classification problem. These four assumptions are:

    - Translation,
    - Permutation,
    - Constant
    - Eigenvalues

This objective is achieved in chapter 3 – Section 3.1.

2. Another main objective is to distinguish between healthy control patients and patients with anal or cervical carcinoma using DNA methylation data and an algorithm combining ridge regression with nonlinear techniques, such as artificial neural networks. This objective is achieved in chapter 3 – Section 3.2

3. To be able to distinguish between control and SSc patients using gene expression data analyzed with machine learning techniques as well as to differentiate between different variants of the illness using the same approach. This objective is achieved in chapter 3 – Section 3.3.

4. Try to identify genes that are relevant in the context of inflammatory bowel disease using machine learning techniques. The process is based on using different machine learning techniques (classification purposes) in combination with Monte Carlo simulations for the selection of genes. This objective is achieved in chapter 3 – Section 3.4.

5. The final objective is to be able to identity appropriate

genes differentiating between Crohn's disease and ulcerative colitis using a similar approach than when distinguishing between healthy and IBD patients. This objective is achieved in chapter 3 – Section 3.4.

CHAPTER 3

PAPERS

In this chapter it can be seen three (Q1) papers already published in peer review journals.

3.1 Paper I

Categorical Variable Mapping Considerations in Classification Problems: Protein Application

Authors: Gerardo Alfonso Perez, Raquel Castillo

Mathematics. 2022, 11(279).

https://doi.org/10.339

mathematics

MDPI

Article

Categorical Variable Mapping Considerations in Classification Problems: Protein Application

Gerardo Alfonso Perez * and Raquel Castillo

Biocomp Group, Institute of Advanced Materials (INAM), Universitat Jaume I, 12071 Castello, Spain
* Correspondence: ga284@cantab.net

Abstract: The mapping of categorical variables into numerical values is common in machine learning classification problems. This type of mapping is frequently performed in a relatively arbitrary manner. We present a series of four assumptions (tested numerically) regarding these mappings in the context of protein classification using amino acid information. This assumption involves the mapping of categorical variables into protein classification problems without the need to use approaches such as natural language process (NLP). The first three assumptions relate to equivalent mappings, and the fourth involves a comparable mapping using a proposed eigenvalue-based matrix representation of the amino acid chain. These assumptions were tested across a range of 23 different machine learning algorithms. It is shown that the numerical simulations are consistent with the presented assumptions, such as translation and permutations, and that the eigenvalue approach generates classifications that are statistically not different from the base case or that have higher mean values while at the same time providing some advantages such as having a fixed predetermined dimensions regardless of the size of the analyzed protein. This approach generated an accuracy of 83.25%. An optimization algorithm is also presented that selects an appropriate number of neurons in an artificial neural network applied to the above-mentioned protein classification problem, achieving an accuracy of 85.02%. The model includes a quadratic penalty function to decrease the chances of overfitting.

Keywords: categorical variables; numerical variables; mappings

MSC: 97-04

check for updates

Citation: Alfonso Perez, G.; Castillo, R. Categorical Variable Mapping Considerations in Classification Problems: Protein Application. *Mathematics* **2023**, *11*, 279. https://doi.org/10.3390/math11020279

Academic Editors: Hongying Liu and Fanhua Shang

Received: 24 November 2022
Revised: 15 December 2022
Accepted: 4 January 2023
Published: 5 January 2023

1. Introduction

Machine learning applications have been successful in different classification tasks in areas such as physics [1–3], chemistry [4–9], and engineering [10–12], and many different algorithms currently exist, such as Trees [13–15], K-Nearest Neighbors (KNNs) [16–18], or Support Vector Machines (SVMs) [19–21]. The internal logic of these machine learning algorithms can substantially vary among the different types of models. A machine learning approach might be advantageous in a situation in which more traditional models do not exist or when these models are too complex to be efficiently implemented. Typically, machine learning models do not require a detailed understanding of the underlying problem that they are trying to model (requiring only some input and output data) or when such detailed modeling is too costly from a computational (or economic) point of view. Therefore, machine learning techniques might be suitable for modeling some complex processes [22–25] such as protein classification. In this article, we focused on the classification task of small proteins and numerical simulations regarding some assumptions regarding the mapping of categorical values, which is an issue directly related to protein modeling, as the input is typically a chain of amino acids, with each amino acid designated with a given letter. A frequently mentioned drawback of this type of approach is that machine learning techniques tend to be black boxes [26–28]. In other words, even if the classification estimations are accurate, the underlying logic is not easily explainable. In this type of modeling, some categorical variables are commonly mapped into numerical values, as it is frequently more convenient to use numerical data in the simulations [29,30]. In this paper, we present a series of four mapping assumptions in the context of protein classification [31,32]. There is a relatively high degree of arbitrariness in the way in which these categorical variables are mapped into numerical values, and it appears interesting to test a series of assumptions about these mapping with numerical simulations. This, in fact, is one of the motivations of this article, as the issue of categorical mapping in the context of protein classification could have some modeling implications.

An algorithm is also presented for the optimization of artificial neural networks [33–35] for the classification problem, including a penalty function [36,37]. The objective of the penalty function is to favor simpler models among classification models that have similar precision. Simpler models, for example, neural networks with fewer neurons, have the advantage of being less prone to overfitting [38,39]. Overfitting is a relatively common issue in which the selected model fits rather accurately to the training data but is not properly generalized when faced with new data. Optimization approaches are commonly used in many diverse fields, such as machine learning applications in the context of ambient music in gyms [40]. There are many different proteins and many different classifications of proteins, thus making this type of analysis challenging, which is a potential limitation of the analysis. In order to minimize this risk, a well-known database of proteins (Protein Data Bank (PDB)) was selected, and we used its standard classification of proteins.

This paper is structured into five sections. An introduction in which some of the basic concepts and the main theme of the article are presented, a literature review in which related works are reported, a materials and methods section in which the four mathematical assumptions about the categorical mapping are stated, as well as the optimization algorithm, data, and general procedures. The last two sections are the results and the conclusions and recommendations, in which we analyze the results, propose potential areas of future research, and suggest some recommendations in this type of analysis.

2. Literature Review

The field of protein modeling using machine learning techniques is rapidly expanding [41]. For instance, Xu et al. used machine learning techniques to describe sequence/activity relationship [42]. This article also focused on mutations, which is an area out of the scope of our analysis. Another interesting article in the field is that of Salau and Jain [43]. In this article, the authors used machine learning techniques for the classification of cell decisions for AKT proteins. The authors used, among other techniques, neural networks such as radial basis functions (RBFs) and multi-layer perceptron (MLP). The importance of feature extraction in this context is frequently mentioned in the literature [44], and it is not exclusive to neural networks, as other popular machine learning techniques such as KNN and SVMs are also mentioned in the literature [45]. There are also some articles such as Hancock et al. [46], highlighting the importance of categorical information in machine learning techniques. More precisely, this article is a survey of categorical information in neural network applications. Some authors, such as Ofer et al. [47], followed a different approach by using natural language processing (NLP) for this type of protein classification task, which avoids the issue of categorical classification (mapping from a categorical value to a numerical value). This, however, remains an approach not followed by the majority of researchers. A potential reason for this is that a numerical approach facilitates the application of some well-known machine learning techniques, and there is so far no indication that this type of NLP approach can generate more accurate results than more traditional machine learning approaches.

Many authors, such as McDowall and Hunter [48] and Nanni et al. [49], revealed the complexity of manually performing protein classification, which is probably one of the reasons for the increasing number of applications of machine learning techniques in this field. Diplaris et al. [50] explicitly mentioned the need for automated tools that can classify new proteins. Data availability has also increased [31]. Using SVMs, Cai et al. managed to achieve an accuracy ranging from 69.1% to 99.6% [51]. Another related field in protein classification that has raised interest from a machine learning approach is the field of protein–protein interaction. In this type of research, the objective is not to identify the type of protein but to forecast the protein–protein interaction. To some degree, the classification of the type of protein could have some impact on the interaction, but it is likely not the determining factor. Bock et al. [52] achieved an 80% success rate in this type of protein–protein interaction analysis. More recently, Das and Chakrabarti [53] followed a similar approach, achieving comparable results. For the special case of G-protein-coupled receptors, Karchin et al. [54] also followed a machine learning approach, using several techniques and obtaining an error ranging from 13.7% to 49%.

In this article, we focus on the analysis of small proteins, which is an area of increasing medical interest [55–57]. We also focus on the analysis of the classification of categorical variables, i.e., amino acids (in different representations), without the need to use NLP approaches. As previously mentioned, the importance of the process of categorical mapping into numerical values has been frequently mentioned in the existing literature. There are some articles using machine learning applications in the field of small proteins. For example, Ernest et al. [58] used this approach to study antimicrobial peptides. This research in antimicrobial peptides is actually one of the subfields that have received more interest among researchers [59,60], but there is some existing research in other areas as well, for example, regarding antifungal peptides [61].

3. Materials and Methods

Mapping variables is a common practice in machine learning applications such as classification problems [62], particularly in situations in which it is necessary to model a process using categorical variables, for example, a protein classification task using their amino acid chains. There is a certain degree of arbitrariness in this process. A protein P can be described by its amino acid chain. This can be seen with an example, as illustrated in Equation (1).

$$P = \{AC...A\} \tag{1}$$

where each amino acid is defined with its standard letter. Note that the letter B is not typically associated with an amino acid. It is usually more convenient in machine learning applications to map into numerical values. A common practice is to map it using alphabetical order and increasing numbers (Equation (2)).

$$\{A, C, ...\} \to f_1\{A, C, ...\} = \{1, 2, ...\} \tag{2}$$

As previously mentioned, this type of mapping is a bit arbitrary, as other numbers could have been used. For example, this should be equivalent to a mapping function that is identical to the previous, but a constant α is added to all the values.

$$f_2\{A, C, ...\} = \{1 + \alpha, 2 + \alpha, ...\} \tag{3}$$

This could be noted as (Equation (4))

$$f_1 \leftrightarrow f_2 \tag{4}$$

Assumption 1 (translation)**.** *A mapping function (Equation (5))*

$$f_1\{C_1, C_2, ...\} = \{a_1, a_2, ...\} \tag{5}$$

where $\{C_1, C_2, ...\}$ are categorical values, and $\{a_1, a_2, ...\}$ are numerical values, should be equivalent ($f_1 \leftrightarrow f_2$) to a mapping function f_2 such that (Equation (6))

$$f_2\{C_1, C_2, ...\} = \{a_1 + \alpha, a_2 + \alpha, ...\} \tag{6}$$

with $\alpha \geq 0$ as a constant.

Assumption 1 could be understood as a translation of mapping with a constant α. Similarly, there is no reason in principle to assume that the numerical values shown in Equation (2) are specifically representative of the related amino acid; hence, a permutation of these values (assigned to each amino acid) should generate another equivalent mapping. For example, mapping according to Equation (7)

$$\{A, C, ...\} \rightarrow f_3\{A, C, ...\} = \{2, 1, ...\} \tag{7}$$

should be equivalent to f_1 $(f_1 \leftrightarrow f_3)$. In this example, the amino acids A and C are mapped into the values 1 and 2, respectively, in mapping f_1, and to the values 2 and 1 in mapping f_3. This change should have no effect on the output of a machine learning classification analysis.

Assumption 2 (permutation). *A mapping function (Equation* (8)*)*

$$f_1\{C_1, C_2, ...\} = \{a_1, a_2, ...\} \tag{8}$$

with $\{C_1, C_2, ...\}$ *as categorical values and* $\{a_1, a_2, ...\}$ *as numerical values should be equivalent to mapping according to* f_3*, as described in Equation* (9)*.*

$$f_3\{C_1, C_2, ...\} = \{a_1, a_2, ..., a_j, a_{j-1}, ...\} \tag{9}$$

where we have a permutation of the numerical values of f_1 *in* f_3*.*

Another common situation in machine learning classification analysis is having data vectors of different lengths, for example, a group of proteins with different numbers of amino acids. These types of data are frequently stored in a matrix for easy use. It is more practical to use a square matrix, and hence a common practice is to add additional zeros (or other numerical values) to the amino acid chains to make them all of the same dimensions. We can define an operator $L()$ such that (Equation (10))

$$L(P^i) = L(\{a_1, a_2, ...\}) = l \tag{10}$$

where P^i is a given protein, and l is the length of the vector (number of amino acids) in this protein. Given a set of k proteins, the maximum size $(\bar{l})$ can be defined as Equation (11):

$$\bar{l} = sup(L(P^1), ..., L(P^k)) \tag{11}$$

Hence, $\forall P^i,\ L(P^i) \leq \bar{l}$. The set of these proteins can be represented as Equation (12):

$$X = (P^1, P^2, ...P^l) = \begin{pmatrix} a_1^1 & a_1^2 & \cdots & a_1^k \\ a_2^1 & a_2^2 & \cdots & a_2^k \\ \vdots & \vdots & \vdots & \vdots \\ \beta & \beta & \cdots & a_{\bar{l}}^k \end{pmatrix} \tag{12}$$

where β is a constant (usually set equal to zero or to a positive value) added in order to make the dimensions of the data vector containing each protein the same. Through this process, we ordered the proteins for clarity purposes (Equation (13)), but this is not a requirement.

$$L(P^1) \leq L(P^1) \leq \cdots \leq L(P^k) \tag{13}$$

As previously mentioned, the constant $(\beta \geq 0)$ added to the data is arbitrary, and hence it should not impact the output of a machine learning classification estimation.

Assumption 3 (constant)**.** *A mapping function (Equation (14))*

$$f_1\{C_1, C_2, ...\} = \{a_1, a_2, ..., \beta_1\} \tag{14}$$

where $\beta_1 \geq 0$ *is an added constant to fit the required dimensions is equivalent to a mapping function* f_4 *(Equation (15)).*

$$f_4\{C_1, C_2, ...\} = \{a_1, a_2, ..., \beta_2\} \tag{15}$$

$\forall\ \beta_2 \geq \beta_1\ \geq 0.$

3.1. Comparable Mappings

In Assumptions 1–3, the mappings are presumed to be equivalents. A less strict requirement (comparable mapping) can be also assumed on similar (but not strictly equivalent) mapping representations. For example, a protein can be described by the number of each type of amino acid and some other indicators such as the length of the chain. In this case, the mapping function would be (Equation (16)):

$$f_5\{C_1, C_2, ...\} = \{na_1, na_2, ..., na_{20}, ...\} \tag{16}$$

where na_i is the number of amino acids of type a_i contained in the chain. It should be noted that the information contained in this mapping is less than in f_1, as it is typically assumed that the order in which the amino acids appear is an important factor in determining the shape and function of the protein [63–65]. Therefore, it cannot be claimed that f_1 and f_5 are equivalent. We denote this as a comparable (but not equivalent) mapping, as expressed in Equation (17).

$$f_1 \leftrightsquigarrow f_5 \tag{17}$$

Note that in f_5, some additional terms, such as the length of the chain, are not explicitly shown for simplicity. A potential full depiction of f_5 could be (Equation (18)).

$$f_5\{C_1, C_2, ...\} = \{l^*, na_1, na_2, ..., na_{20}, \overline{M}, \underline{M}, (\overline{M} - \underline{M}), l^*(\overline{M} - \underline{M})\} \tag{18}$$

with the terms l^*, $\overline{M}$, and $\underline{M}$ defined in Equations (19)–(21).

$$l^* = card\{C_1, C_2, ...\} \tag{19}$$

$$\overline{M} = sup\{na_1, na_2, ..., na_{20}\} \tag{20}$$

$$\underline{M} = inf\{na_1, na_2, ..., na_{20}\} \tag{21}$$

with this mapping, the information for each protein is represented with a vector of length 25. This information can be also represented by a 5x5 matrix.

$$A = \begin{pmatrix} l^* & na_1 & na_2 & na_3 & na_4 \\ na_5 & na_6 & na_7 & na_8 & na_9 \\ na_{10} & na_{11} & na_{12} & na_{13} & na_{14} \\ na_{15} & na_{16} & na_{17} & na_{18} & na_{19} \\ na_{20} & \overline{M} & \underline{M} & (\overline{M} - \underline{M}) & l^*(\overline{M} - \underline{M}) \end{pmatrix} \tag{22}$$

A comparable representation (Equation (23)) would be the eigenvalues of this matrix $|A - \lambda I| = 0$.

$$f_6\{C_1, C_2, ...\} = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5\} \tag{23}$$

Hence, k proteins could be represented as (Equation (24)):

$$\begin{pmatrix} \lambda_1^1 & \lambda_1^2 & \cdots & \lambda_1^k \\ \lambda_2^1 & \lambda_2^2 & \cdots & \lambda_2^k \\ \lambda_3^1 & \lambda_3^2 & \cdots & \lambda_3^k \\ \lambda_4^1 & \lambda_4^2 & \cdots & \lambda_4^k \\ \lambda_5^1 & \lambda_5^2 & \cdots & \lambda_5^k \end{pmatrix} \quad (24)$$

Assumption 4 (eigenvalues)**.** *For some applications, mappings f_1 and f_6 are comparable $(f_1 \leftrightsquigarrow f_6)$.*

When using only five variables per protein, f_6 is more compact than f_1 compared with an arbitrary, large amount for f_1 (depending on the length of the amino acid chain). Assumptions 1–4 will be tested in a later section.

The eigenvalue approach could be considered a feature selection approach. Feature selection is an important component in machine learning approaches [66]. A simplified flowchart can be seen in Figure 1.

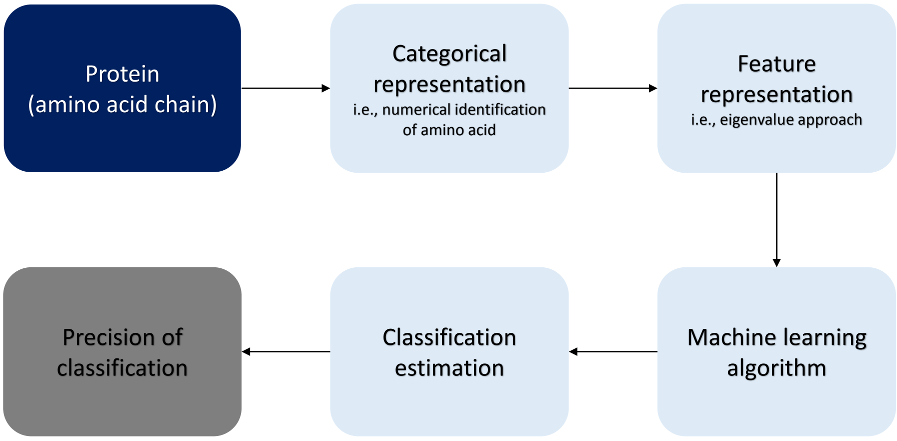


Figure 1. Simplified flowchart diagram.

3.2. Optimization

In this section, we present an algorithm for the optimization of the structure of an artificial neural network. The steps are as follows:

1. Chose the number of simulations (k), the required accuracy C_m, the maximum number of iterations (j_m), and the maximum number of neurons $\bar{a}$.
2. Define a penalty function P. For example,

$$P = \omega a^2 \quad (25)$$

where a is the number of neurons, and ω is a constant.
3. Obtain a randomly generated number of neurons (a), with $1 \le a \le \bar{a} \in I$.
4. Store a classification vector $Y = \{y_1, y_2, \ldots.\}$ (target vector) with $y_i = \{0, 1\}$ and the mapping into a matrix X.
5. Divide the data into a training dataset $\{X_T, Y_T\}$ and a testing dataset $\{X_E, Y_E\}$ [67–69].
6. Train the network (ϕ) with the training dataset $(\phi(X_T, Y_T))$.
7. Estimate the classification estimations $(Y_T^F = Y_T^F(\phi(X_T, Y_T)))$.
8. Estimate b_i as follows:

$$If \begin{cases} Y_{T,i}^F = Y_{T,i} & \Rightarrow b_i = 0 \\ Y_{T,i}^F \neq Y_{T,i} & \Rightarrow b_i = 1 \end{cases} \quad (26)$$

9. Estimate the accuracy Ac (Equation (27)) and calculate the additional metrics of precision (Pr), recall (Rec), and F1-score (F1) using Equations (28)–(30), respectively.

$$Ac = \frac{\sum b_i}{i} \tag{27}$$

$$Pr = \frac{TP}{TP + FP} \tag{28}$$

$$Rec = \frac{TP}{TP + FN} \tag{29}$$

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{30}$$

In this notation, TP, FP, and FN are the true-positive, false-positive, and false-negative values, respectively.

10. The estimated adjusted accuracy Ac^* is expressed in Equation (31):

$$Ac^* = Ac - P(a) \tag{31}$$

This term penalizes an overly complex model with too many neurons.

11. Compare the results of iterations and choose the model.
    - $j = 0 \;\Rightarrow\; MN(j) = Ac^*$
    - $j \neq 0$

$$If \begin{cases} Ac^* > MN(j-1) \;\Rightarrow MN(j) = Ac^* \\ Ac^* \leq MN(j-1) \;\Rightarrow MN(j) = MN(j-1) \end{cases} \tag{32}$$

12. Iterate until $(j = j_m)$ or $MN(j) \geq C_m$.
13. Repeat k times generating MN = $\{MN^1, MN^2, ..., MN^k\}$.
14. Select $\overline{MN} = sup\{MN^1, MN^2, .., MN^k\}$.
15. Calculate the classification estimations (Equation (33)) with the testing dataset for the mode $\overline{MN}$.

$$Y_E^F = Y_E^F(\phi(x_E, Y_E)) \tag{33}$$

16. Repeat step 7 with Y_E^F to obtain the testing dataset accuracy.

3.3. Data

A total of 307 small proteins were analyzed using their amino acid sequence. The data were obtained from the Protein Data Bank (PDB) [70–72]. This database is a frequently used database for protein information [73–77]. For the numerical simulations, we used the protein classification used in PDB. All the analyzed molecules were either classified as asymmetric or cyclic. A categorical variable was assigned to these two types of proteins. The dataset was composed of 254 asymmetric and 53 symmetric small proteins. The median and average number of amino acids were 84 and 81, respectively, and the amino acid chain ranged from 26 to 225 amino acids.

$$Y = Y\{0,1\} = \begin{pmatrix} Asymmetric \\ Cyclic \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \tag{34}$$

The full list of the analyzed molecules can be seen in the Supplementary Material file. All the results shown were estimated using only the testing dataset. In machine learning, it is often not difficult to create a model that accurately describes the training dataset but fails to generalize when faced with new (unseen) data. The training dataset contains approximately 66.6% of the proteins, and the testing dataset contains the remainder 33.3%. Examples of cyclic (Figure 2) and asymmetric (Figure 3) are shown below.

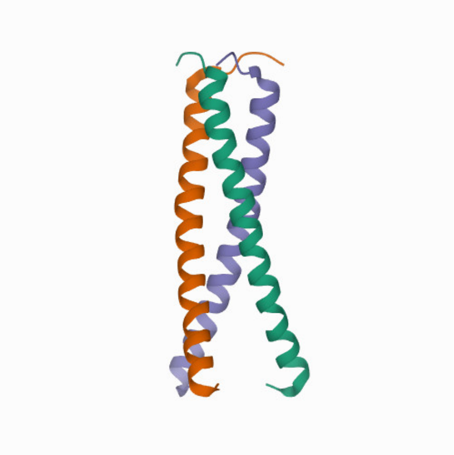


Figure 2. Cyclic molecule example (PKD2L1, Polycystin-L). Extracted from the Protein Data Bank (4GIF).

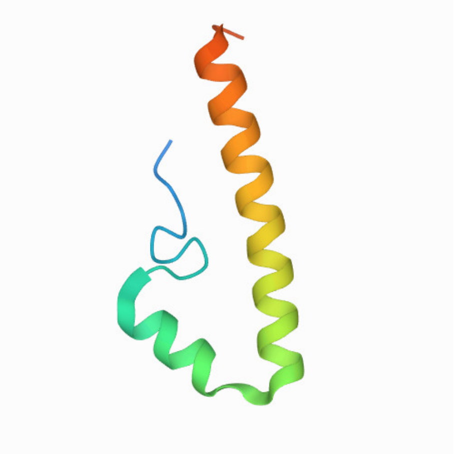


Figure 3. Asymmetric molecule example (S. cerevisiae Rtf1). Extracted from the Protein Data Bank (5EMX).

3.4. Numerical Simulations

Numerical simulations were carried out to test Assumptions 1 to 4. There could be a sizeable difference in accuracy in classification results when using different machine learning algorithms. In order to account for this, a relatively large number (23) of classification algorithms were used. The list of the algorithms used in this study can be found in Table 1. Each model was simulated q times in order to obtain a mean value for accuracy.

Table 1. List of classification algorithms (and related Matlab libraries).

N.	Algorithm	N.	Algorithm
1	Complex Tree (fitctree)	13	Fine KNN (fitcknn)
2	Medium Tree (fitctree)	14	Medium KNN (fitcknn)
3	Simple Tree (fitctree)	15	Coarse KNN (fitcknn)
4	Linear Discriminant (fitcdiscr)	16	Cosine KNN (fitcknn)
5	Quadratic Discriminant (fitcdiscr)	17	Cubic KNN (fitcknn)
6	Logistic Regression (fitglm)	18	Weighted KNN (fitcknn)
7	Linear SVM (fitcsvm)	19	Boosted Trees (fitctree)
8	Quadratic SVM (fitcsvm)	20	Bagged Tress (fitctree)
9	Cubic SVM (fitcsvm)	21	Subspace Discriminant (fitcdiscr)
10	Fine Gaussian SVM (fitcsvm)	22	Subspace KNN (fitcknn)
11	Medium Gaussian SVM (fitcsvm)	23	RUSBoosted Trees (fitctree)
12	Coarse Gaussian SVM (fitcsvm)		

The optimization algorithm was applied to neural networks. The training algorithm selected was the scaled conjugate gradient with the number of neurons selected in an automated way using the optimization algorithm. The optimization algorithm was run for one million iterations, with a constant ω in the penalty function equal to 0.0001.

4. Results

4.1. Assumption 1

In addition to the base case, 4 different models, each with 23 algorithms, were used to test Assumption 1 The difference between these four models resides in the value of the translation constant α ranging from 1000 to 1,000,000 ($\{\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5\}$ = 0, 1000, 10,000, 100,000, 1,000,000). Model 1 was the base case with $\alpha = 0$. The results showing the accuracy can be seen in Figure 4, while the results showing the precision, recall, and F1-score can be seen in Appendix A in Figures A1–A3. A Kolmogorov–Smirnov test [78] was carried out comparing the base model (Model 1 with $\alpha = 0$) with the other models for each of the 23 algorithms (see Table A1 in Appendix A). The test shows that, for the majority of models and algorithms, it cannot be concluded that there is a statistically significant difference between these distributions (accuracy value).

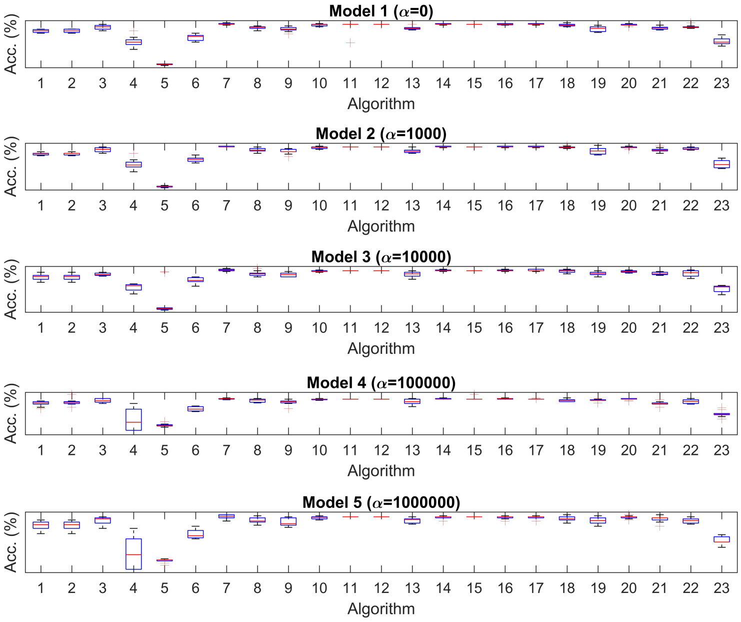


Figure 4. Numerical simulation Assumption 1. Accuracy of models after increasing the translation constant α for all the 23 algorithms. The 23 algorithms are represented in the x-axis, and the accuracy is shown in the y-axis.

4.2. Assumption 2

Five models with different permutations of the numerical values were created for all twenty-three algorithms. The number of permutations for each model was selected randomly. No additional restrictions were introduced in the permutations. The results showing the accuracy can be seen in Figure 5, while the results showing the precision, recall, and F1-score can be seen in Appendix A in Figures A4–A6. As in the previous assumption, the results for the majority of cases suggest no statistically significant difference among the majority of models and algorithms. This was also the result when using a Kolmogorov–Smirnov test comparing Model 1 with Models 6 to 9 (see Table A2 in Appendix A).

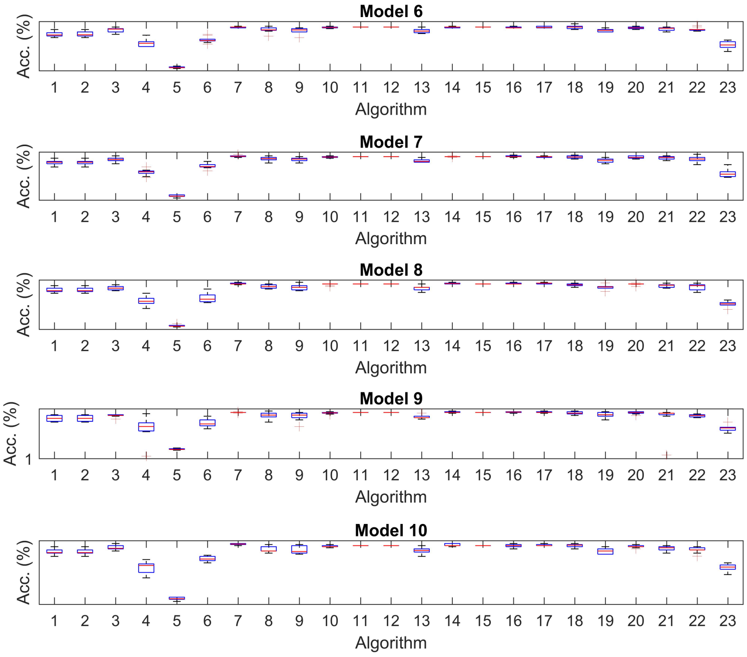


Figure 5. Numerical simulation Assumption 2. Accuracy of various models after permutations in the numerical values of the mapping. The 23 algorithms are represented in the x-axis, and the accuracy is shown in the y-axis.

4.3. Assumption 3

In this section, a variable β was added to each vector to make their length equal. The base case continued to be Model 1 with a $\beta = 0$. Four models were tested with four different betas ($\{\beta_1, \beta_2, \beta_3, \beta_4\}$ = {1000, 10000, 100,000, 1,000,000}). It is worth noting that, in this case, the constant β was added in order to make the dimensions equal, and hence the existing data were not altered as in the case of Assumption 1, in which all data increased by a certain amount α. Figure 6 shows the results (accuracy) of the numerical simulations, indicating, as in the previous cases, that for most of the models and algorithms, there are no statistically significant differences. The results showing the precision, recall, and F1-Score can be seen in Appendix A in Figures A7–A9. Kolmogorov–Smirnov test comparing Model 1 with Models 11 to 14 (see Table A3 in Appendix A) generated similar results.

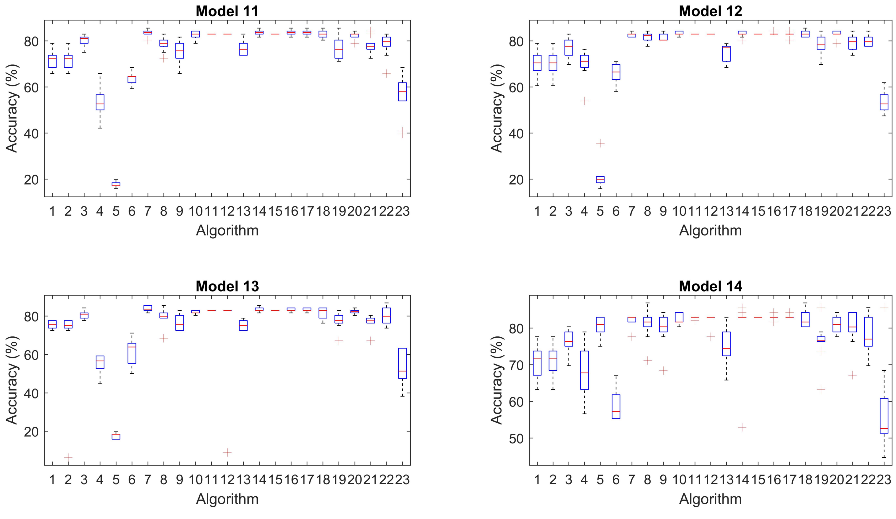


Figure 6. Numerical simulation Assumption 3 (Constant). Accuracy of various models after permutations in the numerical values of the mapping. The 23 algorithms are represented in the x-axis, and the accuracy is shown in the y-axis.

4.4. Assumption 4

In Model 15, rather than the full sequence of amino acids, the input for the classification models was the number of times that a given amino acid appeared in the amino acid chain. Hence, the information about the order of the amino acids was lost. The length of the amino acid chain was also included $(\{l^*, na_1, na_2, \ldots na_{20}\})$. Model 16 was similar to Model 15 but without the length (l^*) of the protein $(\{na_1, na_2, \ldots na_{20}\})$, see the results (accuracy) in Figures 7 and 8. The results for the precision, recall, and F1-score are shown in Appendix A in Figures A10–A15. The results of the Kolmogorov–Smirnov tests, comparing the base model (Model 1) with Models 15 and 16, showed that for the majority of the algorithms, there is no statistically significant difference, as shown in Table A4 in Appendix A.

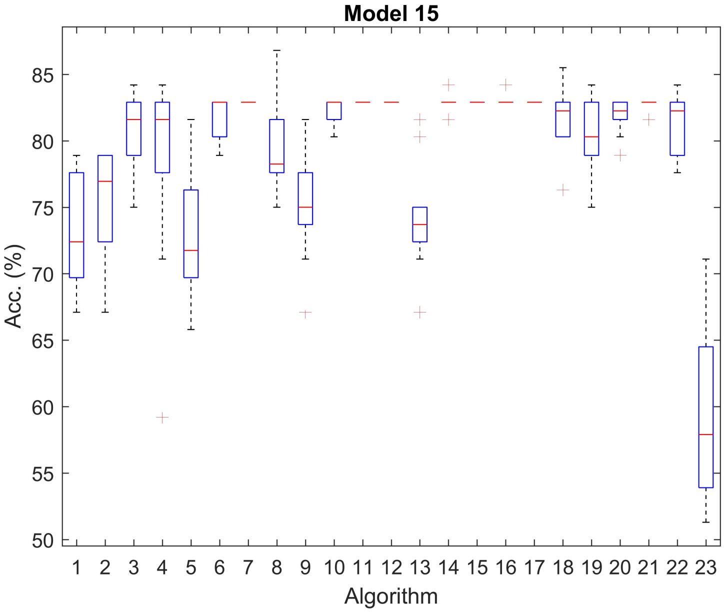


Figure 7. Accuracy of Model 15 for the different algorithms. The 23 algorithms are represented in the *x*-axis, and the accuracy is shown in the *y*-axis.

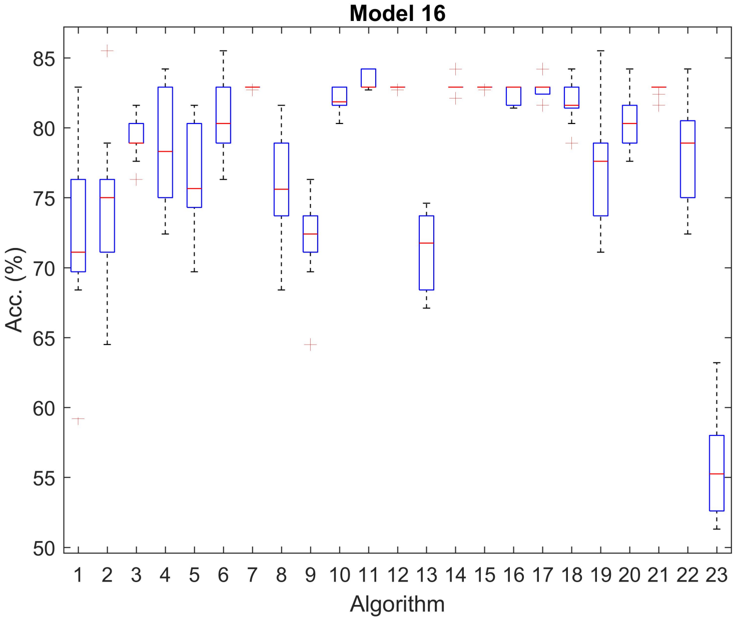


Figure 8. Accuracy of Model 16 for the different algorithms. The 23 algorithms are represented in the *x*-axis, and the accuracy is shown in the *y*-axis.

The next step entailed using the eigenvalues and some additional terms such as the l^*, as defined in expression (22). In many numerical simulations, SVM failed to generate an estimation, and they were thus excluded from the analysis. Interestingly, the rest of the models were accurate or led to better results than the base case (Model 1). The only exception to this trend was the case of the linear discriminant, in which the eigenvalue approach was statistically significantly less accurate. For all the other cases, there was no statistically significant difference, or the mean accuracy of the eigenvalue approach was higher than the value obtained in the base case. Most of the models achieved a mean accuracy above 80%. In Table 2, the mean accuracy values are shown for the eigenvalue approach and Model 1, as well as the p-values for the Kolmogorov–Smirnov test comparing these two approaches. In this table, it can be seen that the best model is the Weighted KNN model, with a mean accuracy of 83.25%, closely followed by the Subspace KNN, Simple Tree, Medium Tree, Complex Tree, and Logistic Regression, with an accuracy of 82.59%, 82.75%, 82.83%, 82.89%, and 82.95% respectively.

Table 2. Mean values of the accuracy for the eigenvalue approach and Model 1 as well as *p*-values of the Kolmogorov–Smirnov test comparing these two approaches. T refers to the average computational time required to train the algorithm.

Algorithm	Eig. Acc.	M1 Acc.	*p*-Value (ks)	T (s)
Complex Tree	82.89	72.04	0.0001	1.34
Medium Tree	82.83	72.13	0.0001	1.79
Simple Tree	82.75	78.11	0.0002	0.60
Linear Discriminant	17.08	54.44	0.0001	1.70
Quadratic Discriminant	30.37	17.47	0.6751	1.82
Logistic regression	82.95	61.87	0.0001	4.32
Fine KNN	79.88	76.53	0.0002	7.96
Medium KNN	81.59	83.18	0.6751	7.85
Coarse KNN	80.65	82.88	1.0000	7.69
Cosine KNN	81.83	82.95	0.6751	8.52
Cubic KNN	81.75	83.14	0.6751	7.10
Weighted KNN	83.25	81.83	0.0069	6.90
Boosted Trees	80.65	75.64	0.0001	7.64
Bagged Trees	82.83	81.6	0.0069	9.71
Subspace Discriminant	81.91	77.03	0.0001	10.98
Subspace KNN	82.59	78.91	0.0002	11.84
RUSBoosted Trees	81.55	54.74	0.0001	13.32

4.5. Optimization

We also used an algorithm for the optimization of the classification using neural networks, as described in Section 3.2. The algorithm was the scale conjugate gradient, and the process involved one million iterations. This model achieved an 85.02% out-of-sample classification accuracy with 215 neurons, suggesting that model parameter optimization plays an important role in improving classification accuracy. In the context of protein classification, it is important to carry out parameter optimization in a consistent way to improve the chances of the model to generalize (classify new data) with a reasonable level of accuracy. Randomly selecting the parameter could potentially lead to biases in the model or poor generalization. Figure 9 shows that the classification accuracy improves as the number of iterations increase, initially very rapidly and then more slowly as the model approaches its upper limit. There are several potential ways of performing data validation [79]. In this article, we performed cross-validation of the data in the training dataset 10 times, and then the results were tested with the testing dataset (not used during the training phase).

A limitation in this article, and a potential area of future work, is increasing the number of analyzed proteins. In this article, we analyzed 307 proteins for classification purposes, but this number could be further increased. This type of analysis could also be parallelized, which could enable a larger number of simulations to be performed while potentially not substantially increasing computational time.

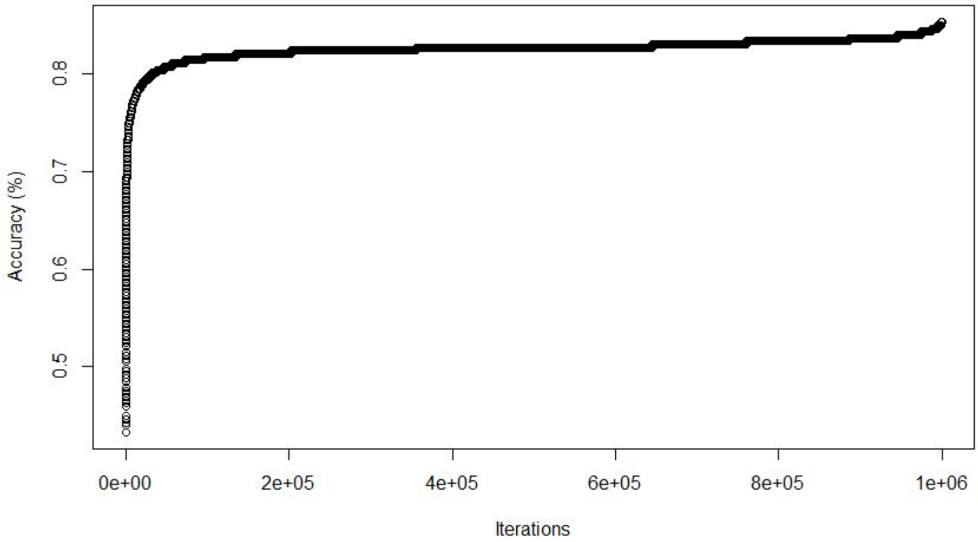


Figure 9. Improvement in accuracy as the number of iterations increases in the optimization algorithm.

5. Conclusions and Recommendations

Mapping categorical variables into numerical variables is a common practice in many machine learning classification tasks, and it is frequently carried out in an arbitrary matter. In this paper, we proposed four different assumptions related to this topic in the context of protein classification: (1) translation, (2) permutation, (3) constant, and (4) eigenvalues. Assumptions 1–3 are related to the concept of equivalent mappings in which changes to the mapping should, in principle, not alter the results of a classification analysis (for instance, adding a constant to all the input parameters). Assumption 4 relates to a less strict requirement in which the mappings are not in principle strictly equivalent, but they are comparable. An example is the eigenvalue mapping approach in which the information about the order of the amino acids (present in the initial mapping f_1) is not contained in this new mapping (f_6). The results for Assumptions 1–3 showed that, in the majority of the cases, no statistically significant difference exists between the mappings when we compared their mean accuracy. The case of Assumption 4 is different, and we see that using the eigenvalue approach generates similar or more accurate classifications than the base case model. All these numerical simulations were carried out for 23 different classification algorithms, including KNN, Tress, and SVMs. As previously mentioned, the eigenvalue approach (related to Assumption 4) generated accurate estimations for most algorithms. One noticeable exception was SVM, which, in many cases, failed to generate a classification estimation and was, therefore, excluded from the analysis. For the majority of the other algorithms, the eigenvalue approach generated results that were not statistically significantly different from the base case or that had higher mean accuracy than the base case. The best model obtained a mean classification accuracy of 83.25%. While direct comparisons are challenging, this result is 14.15% better than the lower-bound result obtained by Cai et al. [51] but lower than the upper bound. This is consistent with the idea of focusing the analysis on the stability of results rather than only focusing on increasing accuracy. This result is also substantially higher than the lower bound achieved by Karchin et al. [54], in which the authors focused on a specific subset of proteins (G-protein-coupled receptors).

An optimization analysis algorithm was also presented for the automated selection of the number of neurons in a classification model using only the frequency of the occurrence of amino acid in the amino acid chain as input (no order information), as well as the length of the chain. The model included a quadratic penalty function to try to decrease the chance of overfitting. This approach generated an accuracy of 85.02% percent. This result is even closer to the upper bound (and substantially higher than the lower bound) of Cai et al. [51] even after accounting for the penalty function introduced to avoid an overly complex model, which potentially could impact the generalization capabilities of the model, i.e., the accuracy of the classification when faced with new data. Furthermore, this approach does not require the use of techniques such as NLP [47], which could be beneficial from an implementation point of view, as there is a large number of machine learning applications that can be easily and accurately applied to numerical values, and there is no indication that an NLP approach will generate more accurate results.

It should be noted that this accuracy is not directly comparable with the accuracy obtained in the previous sections, as there was no additional algorithm optimization. The focus of the previous section was on the comparability of the models, and hence it did not appear appropriate to add additional optimization techniques that differ in the different algorithms. For instance, an optimization process based on finding an appropriate number of neurons, as shown in the optimization section, cannot be performed for other classification techniques such as KNN, SVM, or Trees, as they do not use artificial neurons.

This type of big data analysis is challenging and can be computationally expensive, depending on the type of machine learning applied and/or the optimization algorithm followed. As an area of future research, it would be interesting to use genetic algorithms or particle algorithms as potential optimization strategies. There is a wide range of options to optimize this type of analysis. There is, however, the risk of overfitting the model, and some measures should be taken to minimize that risk, such as using a penalty function, as we used in this article, to penalize the accuracy of overly complex models. Arguably, an overly complex model is more likely to result in an overfitting issue than a simpler model.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/math11020279/s1.

Author Contributions: Conceptualization, G.A.P. and R.C.; methodology, G.A.P. and R.C.; software, G.A.P.; validation, G.A.P. and R.C.; formal analysis, G.A.P. and R.C.; investigation, G.A.P. and R.C.; resources, G.A.P. and R.C.; data curation, G.A.P. and R.C.; writing—original draft preparation, G.A.P.; writing—review and editing, G.A.P. and R.C.; visualization, G.A.P. and R.C.; supervision, G.A.P. and R.C.; project administration, G.A.P. and R.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Spanish Ministerio de Ciencia, Innovación y Universidades (PGC2018-094852-B-C21), and Universitat Jaume I (UJI-B2019-43).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All the data used in this article can be accessed from the Protein Data Bank https://www.rcsb.org/ accessed on 14 December 2022.

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

Appendix A

Table A1. *p*-values of the Kolmogorov–Smirnov test Assumption 1 (translation).

M1–M2	M1–M3	M1–M4	M1–M5
0.97479	0.97479	0.31285	0.67508
0.97479	0.97479	0.031047	0.67508
1.00000	0.67508	0.67508	0.97479

Table A1. *Cont.*

M1–M2	M1–M3	M1–M4	M1–M5
1.00000	0.67508	0.11084	0.11084
1.00000	0.31285	1.89×10^{-5}	1.89×10^{-5}
0.97479	0.31285	0.97479	0.67508
1.00000	0.67508	0.31285	0.97479
0.97479	0.97479	0.11084	0.31285
0.97479	0.97479	0.67508	0.67508
1.00000	0.97479	0.97479	1.00000
0.97479	0.97479	0.97479	0.97479
1.00000	1.00000	1.00000	1.00000
1.00000	0.31285	0.67508	0.11084
1.00000	0.97479	1.00000	0.67508
1.00000	1.00000	1.00000	1.00000
1.00000	0.97479	0.67508	0.97479
1.00000	0.67508	0.97479	0.97479
0.97479	0.67508	0.67508	0.67508
1.00000	0.67508	0.031047	0.67508
1.00000	0.97479	0.11084	0.31285
1.00000	0.97479	0.11084	0.03105
0.67508	0.67508	0.31285	0.67508
0.97479	0.67508	0.11084	0.67508

Table A2. *p*-values of the Kolmogorov–Smirnov test Assumption 2 (permutation).

M1–M6	M1–M7	M1–M8	M1–M9	M1–M10
0.67508	0.67508	0.67508	0.67508	0.31285
0.97479	0.67508	0.67508	0.67508	0.31285
1.00000	0.97479	0.67508	0.31285	0.31285
0.31285	0.11084	0.67508	0.31285	0.31285
1.00000	0.67508	1.00000	1.00000	0.67508
0.67508	0.031047	0.97479	0.67508	0.31285
0.67508	0.97479	1.00000	0.97479	0.11084
0.31285	0.67508	0.67508	0.67508	0.67508
0.31285	0.11084	0.31285	0.31285	0.67508
1.00000	0.97479	0.67508	1.00000	0.97479
0.97479	0.97479	0.97479	0.97479	0.97479
1.00000	1.00000	1.00000	1.00000	1.00000
0.97479	0.97479	0.67508	0.31285	0.97479
0.67508	0.67508	0.97479	1.00000	0.67508
1.00000	1.00000	1.00000	1.00000	1.00000
0.31285	1.00000	0.67508	0.97479	0.97479
0.31285	0.97479	0.67508	0.97479	0.97479
0.67508	0.97479	1.00000	0.97479	0.67508
0.11084	0.0068986	0.11084	0.031047	0.31285
1.00000	0.67508	0.31285	0.31285	0.97479
1.00000	1.00000	0.31285	0.67508	0.67508

Table A3. *p*-values of the Kolmogorov–Smirnov test Assumption 3 (constant).

M1–M11	M1–M12	M1–M13	M1–M14
1.00000	0.67508	0.11084	0.67508
1.00000	0.67508	0.31285	0.97479
0.67508	0.67508	0.31285	0.67508
1.00000	0.0012162	0.67508	0.0068986
0.97479	0.0068986	0.97479	1.89×10^{-5}
0.67508	0.31285	0.97479	0.31285
0.97479	0.67508	0.97479	0.31285
0.67508	0.0068986	0.11084	0.031047
0.97479	0.00017012	0.67508	0.0068986
0.97479	0.31285	0.67508	0.97479
0.97479	0.97479	0.97479	1.00000
1.00000	1.00000	1.00000	1.00000
0.97479	0.67508	0.67508	0.31285
0.97479	1.00000	1.00000	0.97479
1.00000	1.00000	1.00000	1.00000
0.97479	0.67508	0.97479	0.97479
0.97479	0.97479	1.00000	0.97479
0.67508	0.31285	0.67508	1.00000
1.00000	0.67508	0.67508	0.97479
0.31285	0.11084	0.97479	0.67508
0.97479	0.31285	0.97479	0.031047
0.67508	0.67508	0.67508	0.31285
0.31285	0.67508	0.67508	0.67508

Table A4. *p*-values of the Kolmogorov–Smirnov test Models 15 and 16.

M1–M15	M1–M16
0.31285	0.67508
0.031047	0.67508
0.11084	0.31285
0.0012162	0.00017012
1.89×10^{-5}	1.89×10^{-5}
1.89×10^{-5}	1.89×10^{-5}
0.31285	0.31285
0.67508	0.31285
0.97479	0.11084
0.97479	0.97479
0.97479	0.31285
1.00000	1.00000
0.31285	0.0012162
0.67508	0.67508
1.00000	1.00000
0.67508	0.67508
0.67508	0.97479
0.97479	1.00000
0.031047	0.67508
0.97479	0.11084
0.00017012	0.00017012
0.11084	0.67508
0.31285	0.97479

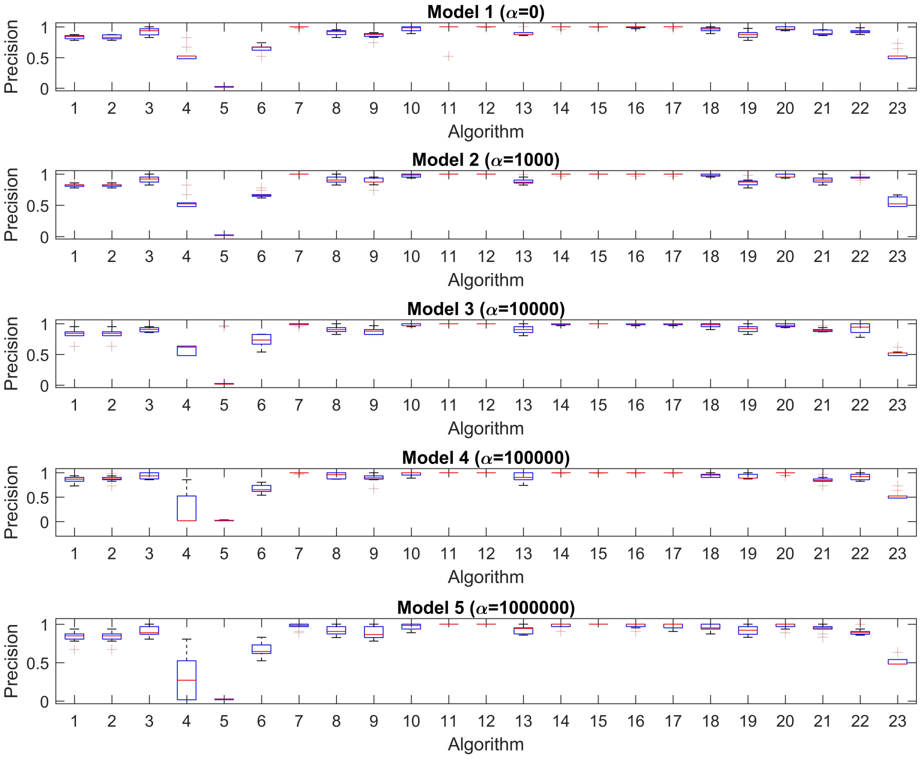


Figure A1. Numerical simulation Assumption 1. Precision of models after increasing the translation constant α for all the 23 algorithms. The 23 algorithms are represented in the x-axis, and the accuracy is shown in the y-axis.

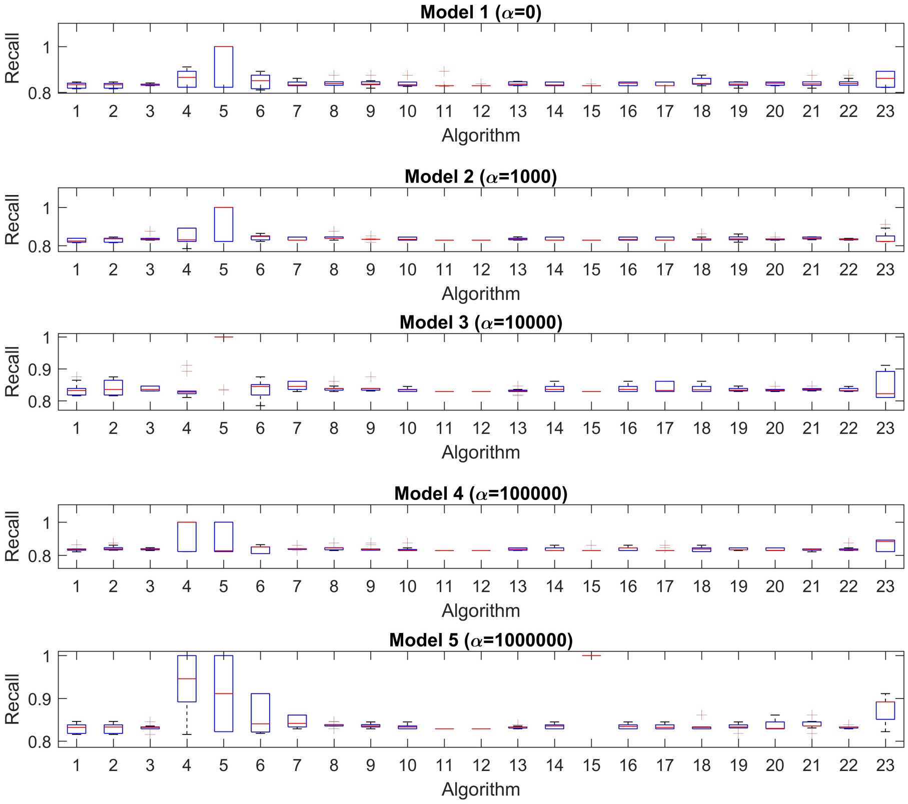


Figure A2. Numerical simulation Assumption 1. Recall of models after increasing the translation constant α for all the 23 algorithms. The 23 algorithms are represented in the x-axis, and the accuracy is shown in the y-axis.

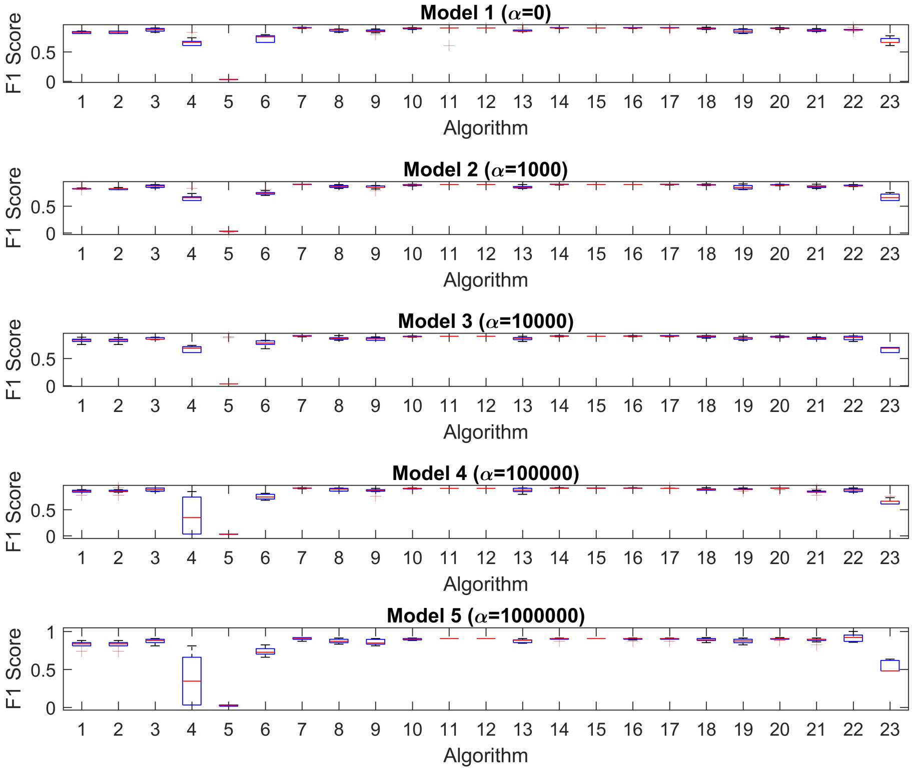


Figure A3. Numerical simulation assumption 1. F1-score of models after increasing the translation constant α for all the 23 algorithms. The 23 algorithms are represented in the x-axis, and the accuracy is shown in the y-axis.

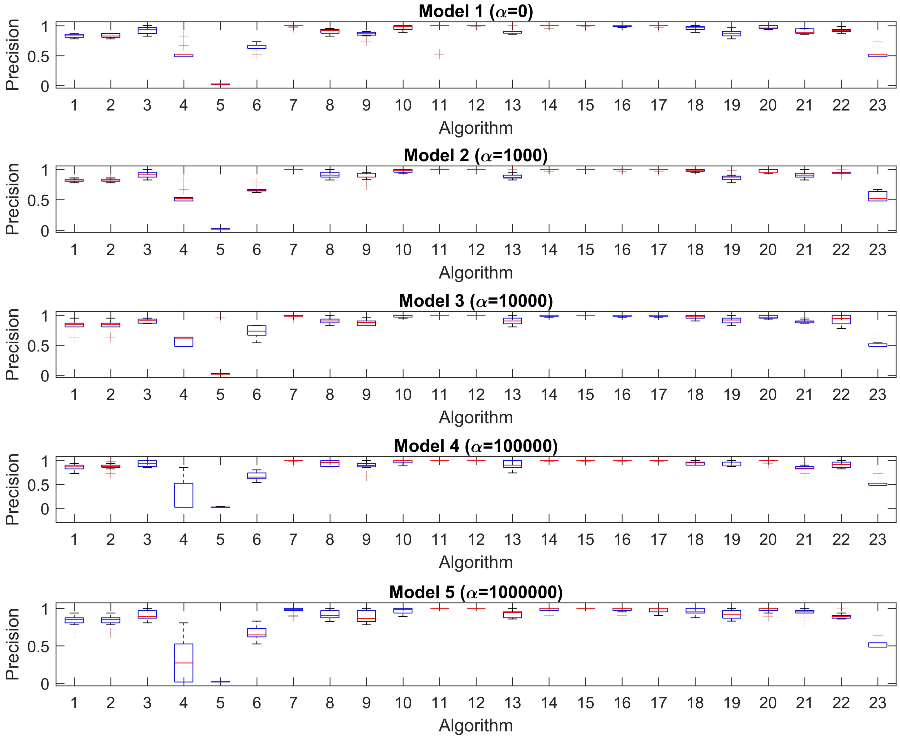


Figure A4. Numerical simulation Assumption 2. Precision of various models after permutations in the numerical values of the mapping. The 23 algorithms are represented in the x-axis, and the accuracy is shown in the y-axis.

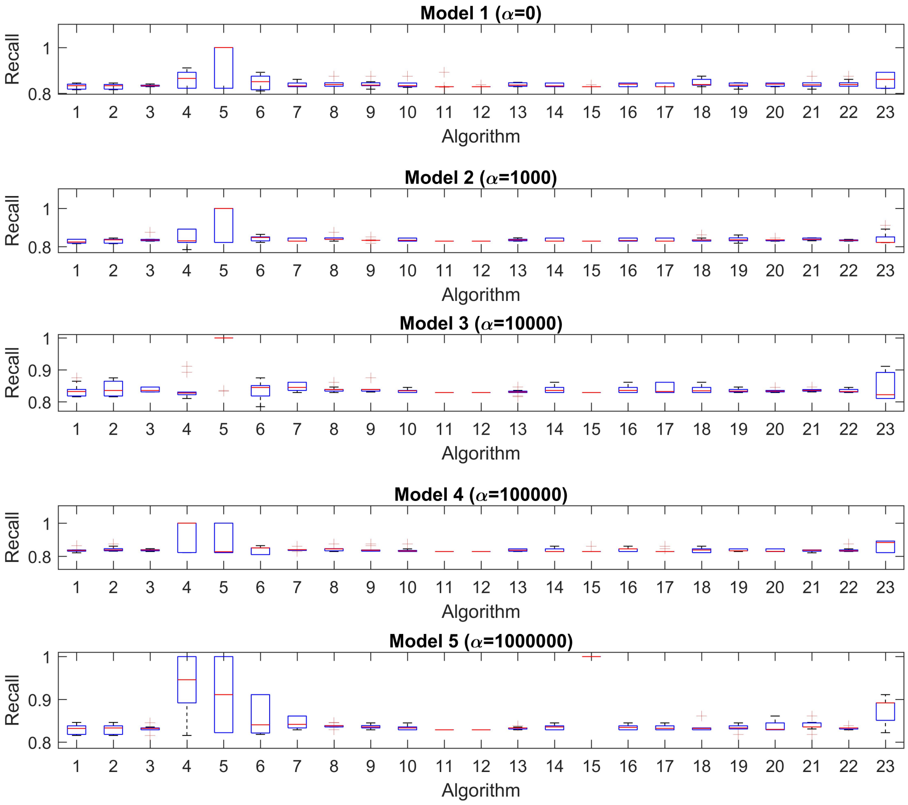


Figure A5. Numerical simulation Assumption 2. Recall of various models after permutations in the numerical values of the mapping. The 23 algorithms are represented in the *x*-axis, and the accuracy is shown in the *y*-axis.

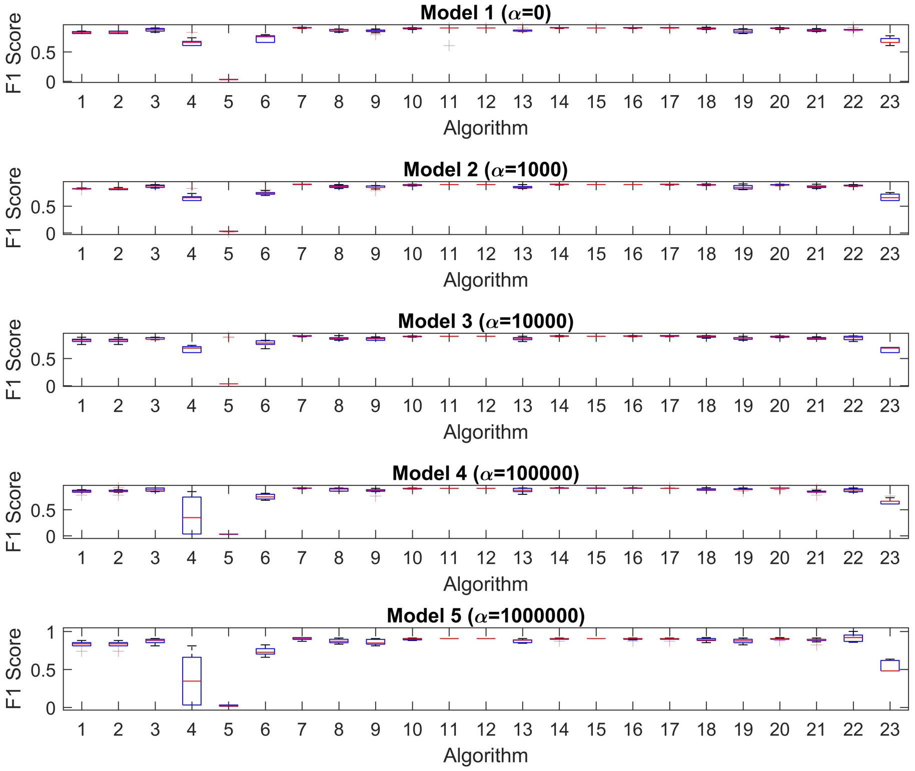


Figure A6. Numerical simulation Assumption 2. F1-score of various models after permutations in the numerical values of the mapping. The 23 algorithms are represented in the *x*-axis, and the accuracy is shown in the *y*-axis.

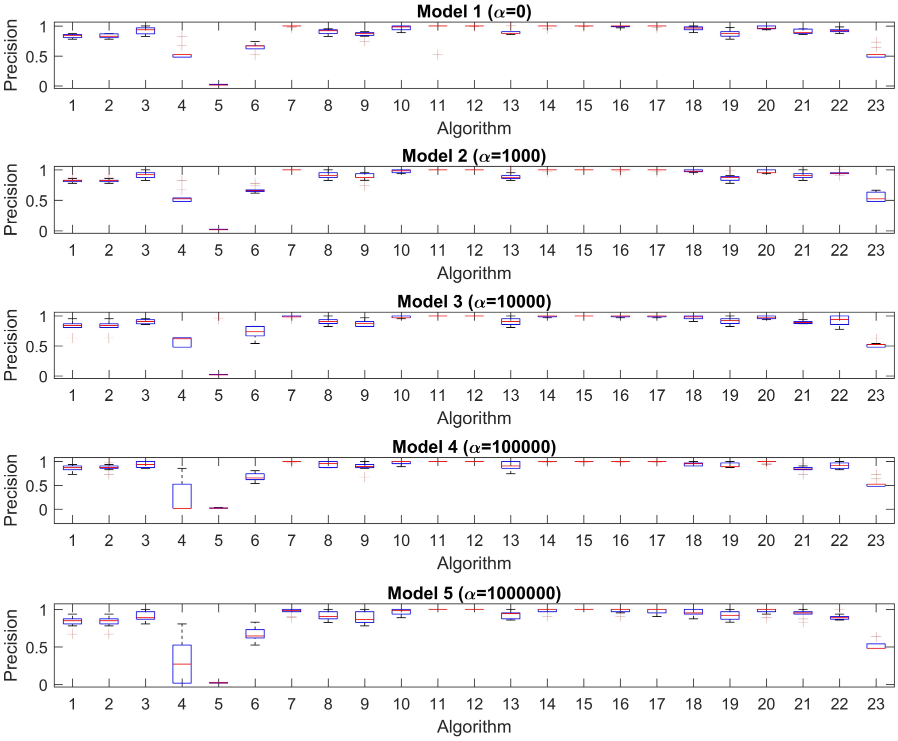


Figure A7. Numerical simulation Assumption 3 (constant). Precision of various models after permutations in the numerical values of the mapping. The 23 algorithms are represented in the *x*-axis, and the accuracy is shown in the *y*-axis.

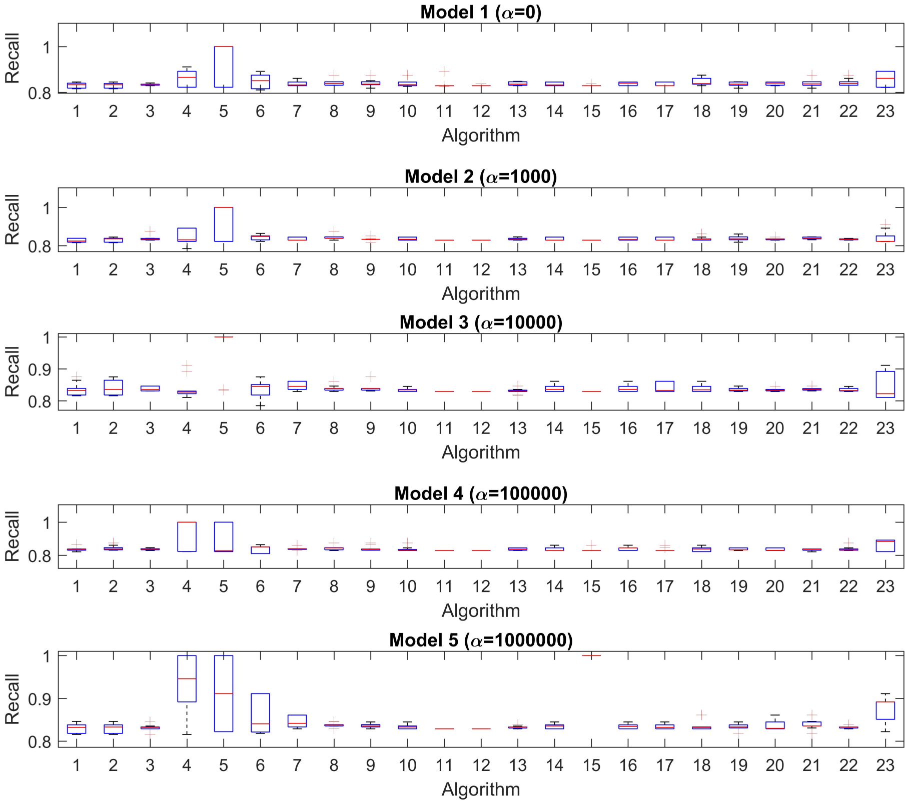


Figure A8. Numerical simulation assumption 3 (Constant). Recall of various models after permutations in the numerical values of the mapping. The 23 algorithms are represented in the *x*-axis, and the accuracy is shown in the *y*-axis.

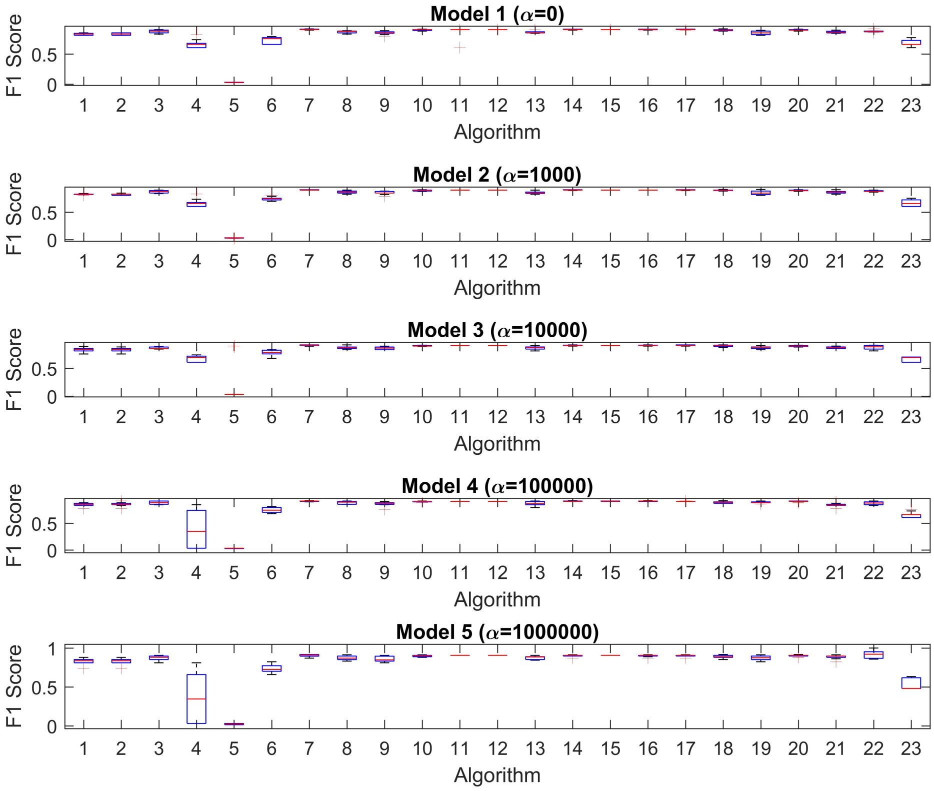


Figure A9. Numerical simulation Assumption 3 (constant). F1-score of various models after permutations in the numerical values of the mapping. The 23 algorithms are represented in the x-axis, and the accuracy is shown in the y-axis.

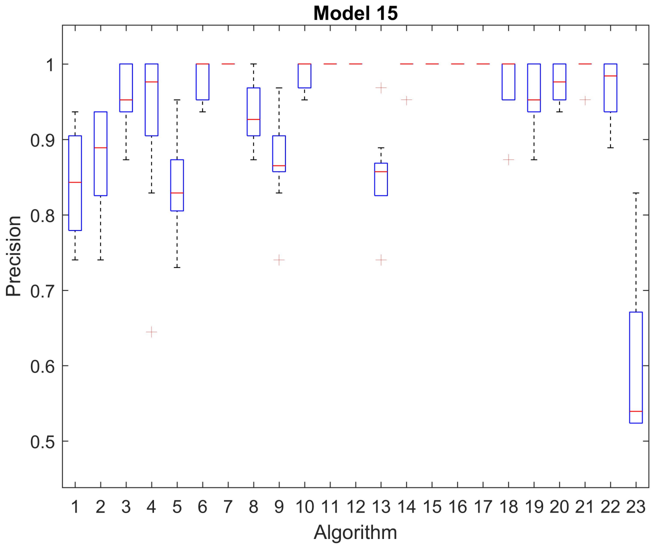


Figure A10. Precision of Model 15 for the different algorithms. The 23 algorithms are represented in the x-axis, and the accuracy is shown in the y-axis.

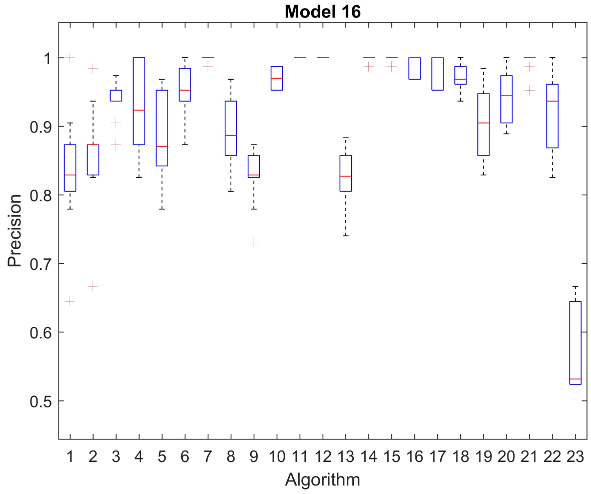


Figure A11. Precision of Model 16 for the different algorithms. The 23 algorithms are represented in the *x*-axis, and the accuracy is shown in the *y*-axis.

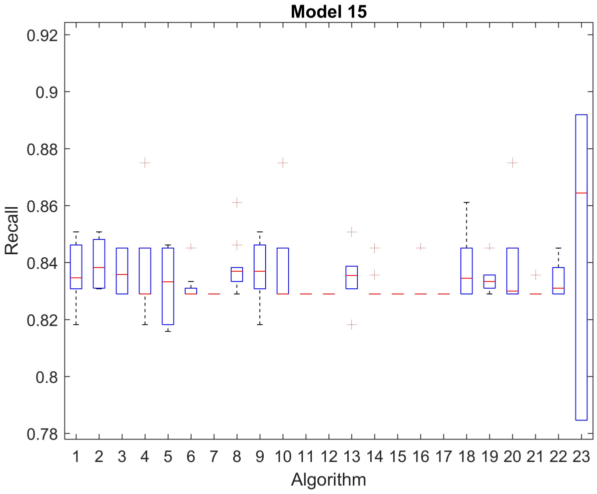


Figure A12. Recall of Model 15 for the different algorithms. The 23 algorithms are represented in the *x*-axis, and the accuracy is shown in the *y*-axis.

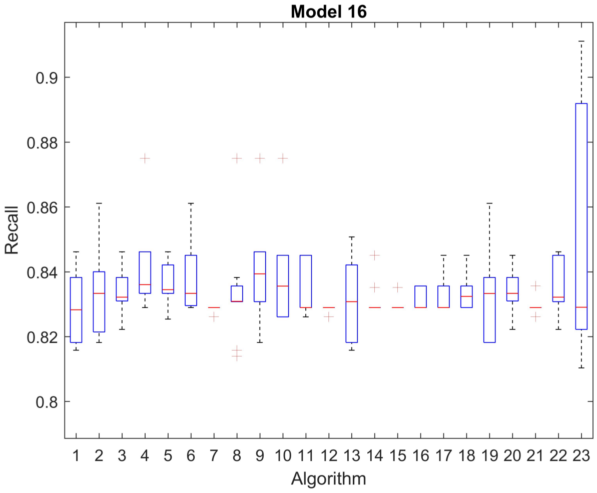


Figure A13. Recall of Model 16 for the different algorithms. The 23 algorithms are represented in the *x*-axis, and the accuracy is shown in the *y*-axis.

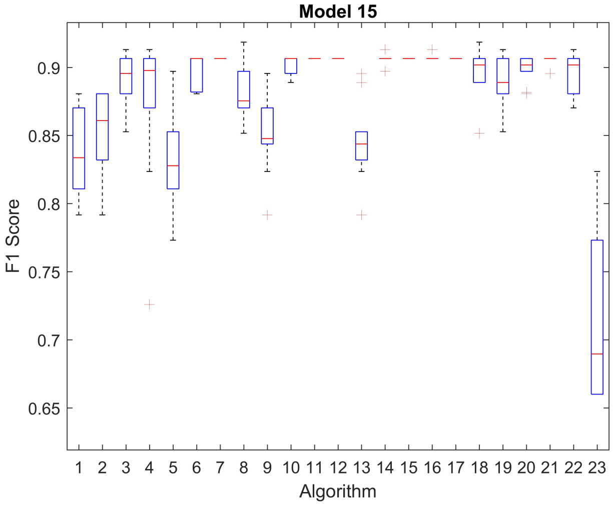


Figure A14. F1-score of Model 15 for the different algorithms. The 23 algorithms are represented in the x-axis, and the accuracy is shown in the y-axis.

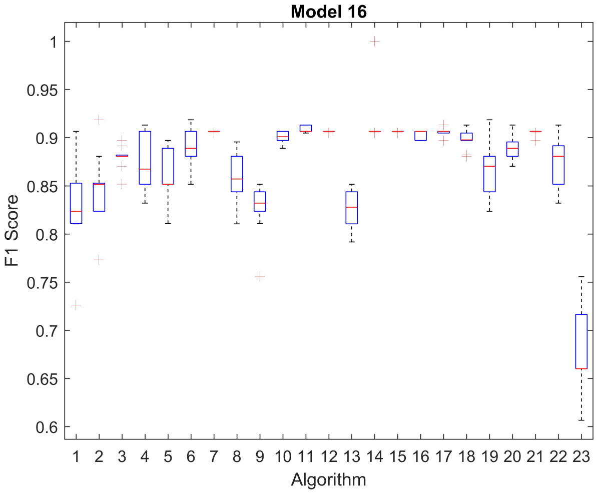


Figure A15. F1-score of Model 16 for the different algorithms. The 23 algorithms are represented in the x-axis, and the accuracy is shown in the y-axis.

References

1. Carleo, G.; Cirac, I.; Cranmer, K.; Daudet, L.; Schuld, M.; Tishby, N.; Vogt-Maranto, L.; Zdeborová, L. Machine learning and the physical sciences. *Rev. Mod. Phys.* **2019**, *91*, 45002–45041 . [CrossRef]
2. Radovic, A.; Williams, M.; Rousseau, D.; Kagan, M.; Bonacorsi, D.; Himmel, A.; Aurisano, A.; Terao, K.; Wongjirad, T. Machine learning at the energy and intensity frontiers of particle physics. *Nature* **2018**, *560*, 41–48. [CrossRef] [PubMed]
3. Karniadakis, G.E.; Kevrekidis, I.G.; Lu, L.; Perdikaris, P.; Wang, S.; Yang, L. Physics-informed machine learning. *Nat. Rev. Phys.* **2021**, *3*, 422–440. [CrossRef]
4. Jimenez, J.; Doerr, S.; Martinez-Rosell, G.; Rose, A.S.; DeFabritis, G. Deepsite: Protein-binding site predictor using 3D-convolutional neural networks. *Bioinformatics* **2017**, *19*, 3036–3042. [CrossRef] [PubMed]
5. Pages, G.; Charmettant, B.; Grudinin, S. Protein model quality assessment using 3D oriented convolutional neural networks. *Bioinformatics* **2019**, *35*, 3313–3319. [CrossRef] [PubMed]
6. Wang, X.; Terashi, G.; Christoffer, C.W.; Zhu, M.; Kihara, D. Protein docking model evaluation by 3D deep convolutional neural network. *Bioinformatics* **2020**, *36*, 2113–2118. [CrossRef] [PubMed]
7. Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes, D.R. Protein-ligand scoring with convolutional neural networks. *J. Chem. Inf. Model.* **2017**, *57*, 942–957. [CrossRef]
8. Keith, J.A.; Vassilev-Galindo, V.; Cheng, B.; Chmiela, S.; Gastegger, M.; Muller, K.; Tkatchenko, A. Combining machine learning and computational chemistry for predictive insights into chemical systems. *Chem. Rev.* **2021**, *121*, 9816–9872. [CrossRef]
9. Artrith, N.; Butler, K.T.; Coudert, F.; Han, S.; Isayev, O.; Jain, A.; Walsh, A. Best practices in machine learning for chemistry. *Nat. Chem.* **2021**, *13*, 505–508. [CrossRef]

10. Amershi, S.; Begel, A.; Bird, C.; DeLine, R.; Gall, H.; Kamar, E.; Nagappan, N.; Nushi, B.; Zimmermann, T. Software engineering for machine learning: A case study. In Proceedings of the 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP), Montreal, QC, Canada, 25–31 May 2019; pp. 291–300.
11. Park, C.; Took, C.C.; Seong, J. Machine learning in biomedical engineering. *Biomed. Eng. Lett.* **2018**, *8*, 1–3. [CrossRef]
12. Zhang, D.; Tsai, J. Machine learning and software engineering. *Softw. Qual. J.* **2003**, *11*, 87–119. [CrossRef]
13. Rodriguez-Galiano, V.; Sanchez-Castillo, M.; Chica-Olmo, M.; Chica-Rivas, M. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geol. Rev.* **2015**, *71*, 804–818. [CrossRef]
14. Blanco-Justicia, A.; Domingo-Ferrer, J. Machine learning explainability through comprehensible decision trees. In Proceedings of the International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Canterbury, UK, 26–29 August 2019; pp. 15–26.
15. Allen, J.P.; Snitkin, E.; Pincus, N.B.; Hauser, A.R. Forest and trees: Exploring bacterial virulence with genome-wide association studies and machine learning. *Trends Microbiol.* **2021**, *29*, 621–633. [CrossRef] [PubMed]
16. Guo, G.; Wang, H.; Bell, D.; Bi, Y.; Greer, K. *OTM Confederated International Conferences "On the Move to Meaningful Internet Systems"*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 986–996.
17. Lee, T.; Wood, W.T.; Phrampus, B.J. A machine learning (kNN) approach to predicting global seafloor total organic carbon. *Wiley Online Libr.* **2019**, *33*, 37–46. [CrossRef]
18. Zhang, S.; Li, X.; Zong, M.; Zhu, X.; Cheng, D. Learning k for knn classification. *ACM Trans. Intell. Syst. Technol.* **2017**, *8*, 1–19. [CrossRef]
19. Noble, W.S. What is a support vector machine? *Nat. Biol.* **2006**, *24*, 1565–1567. [CrossRef] [PubMed]
20. Cortes, C.; Vapnik, V. Support vector machine. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
21. Pisner, D.A.; Schnyer, D.M. Support vector machine. In *Chapter 6—Machine Learning*; Academic Press: Cambridge, MA, USA, 2020; pp. 101–121.
22. Qi, D.; Majda, A.J. Using machine learning to predict extreme events in complex systems. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 52–59. [CrossRef]
23. Qi, D.; Majda, A.J. Introduction to Focus Issue: When machine learning meets complex systems: Networks, chaos, and nonlinear dynamics. *Chaos Interdiscip. J. Nonlinear Sci.* **2020**, *30*, 063151.
24. Wood, D. A transparent open-box learning network provides insight to complex systems and a performance benchmark for more-opaque machine learning algorithms. *Adv. Geo-Energy Res.* **2018**, *2*, 148–162. [CrossRef]
25. Qin, J.; Hu, F.; Liu, Y.; Witherell, P.; Wang, C.; Rosen, D.W.; Simpson, T.; Lu, Y.; Tang, Q. Research and application of machine learning for additive manufacturing. *Addit. Manuf.* **2022**, *52*, 102691. [CrossRef]
26. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [CrossRef] [PubMed]
27. McGovern, A.; Lagerquist, R.; Gagne, D.J.; Jergensen, G.E.; Elmore, K.L.; Homeyer, C.R.; Smith, T. Making the black box more transparent: Understanding the physical implications of machine learning. *Nat. Mach. Intell.* **2019**, *100*, 2175–2199. [CrossRef]
28. Zhou, Z. Learnware: On the future of machine learning. *Front. Comput. Sci.* **2016**, *10*, 589–590. [CrossRef]
29. Cerda, P.; Varoquaux, G.; Kégl, B. Similarity encoding for learning with dirty categorical variables. *Mach. Learn.* **2018**, *8*, 1477–1494. [CrossRef]
30. Cerda, P.; Varoquaux, G.; Kégl, B. Encoding high-cardinality string categorical variables. *IEEE Trans. Knowl. Data Eng.* **2020**, *34*, 1164–1176. [CrossRef]
31. Sonego, P.; Pacurar, M.; Dhir, S.; Kertesz-Farkas, A.; Kocsor, A.; Gáspári, Z.; Leunissen, J.A.M.; Pongor, S. A protein classification benchmark collection for machine learning. *Nucleic Acids Res.* **2007**, *35*, 232–236. [CrossRef]
32. Jain, P.; Garibaldi, J.M.; Kirst, J.D. Supervised machine learning algorithms for protein structure classification. *Comput. Biol. Chem.* **2009**, *33*, 216–223. [CrossRef]
33. Muller, B.; Joachim, R.; Strickland, M.T. *Neural Networks an Introduction*; Springer Science & Business Media: Abingdon, UK, 1995.
34. Anderson, J.A. *An Introduction to Neural Networks*; MIT Press: Cambridge, MA, USA, 1995.
35. Miller, W.T.; Werbos, P.J.; Sutton, R.S. *Neural Networks for Control*; MIT Press: Cambridge, MA, USA, 1995.
36. Le, T.; Hoai, A.; Le, H.M.; Pham, D.T. Feature selection in machine learning: An exact penalty approach using a difference of convex function algorithm. *Mach. Learn.* **2015**, *101*, 163–186.
37. Jiang, M.; Meng, Z.; Shen, R. Partial Exactness for the Penalty Function of Biconvex Programming. *Entropy* **2021**, *23*, 132. [CrossRef]
38. Roelofs, R.; Shankar, V.; Recht, B.; Fridovich-Keil, S.; Hardt, M.; Miller, J.; Schmidt, L. A meta-analysis of overfitting in machine learning. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 1–11.
39. Peng, Y.; Nagata, M.H. An empirical overview of nonlinearity and overfitting in machine learning using COVID-19 data. *Chaos Solitons Fractals* **2020**, *139*, 110055. [CrossRef]
40. De Prisco, R.; Guarino, A.; Lettieri, N.; Malandrino, D.; Zaccagnino, R. Providing music service in ambient intelligence: Experiments with gym users. *Expert Syst. Appl.* **2021**, *177*, 114951. [CrossRef]
41. Kamerzell, T.J.; Middaugh, C.R. Prediction machines: Applied machine learning for therapeutic protein design and development. *J. Pharm. Sci.* **2021**, *110*, 665–681. [CrossRef]

42. Xu, Y.; Verma, D.; Sheridan, R.P.; Liaw, A.; Ma, J.; Marshall, N.M.; McIntosh, J.; Sherer, E.C.; Svetnik, V.; Johnston, J.M. Deep Dive into Machine Learning Models for Protein Engineering. *J. Chem. Inf. Model.* **2020**, *60*, 2773–2790. [CrossRef]
43. Salau, A.O.; Jain, S. Adaptive diagnostic machine learning technique for classification of cell decisions for AKT protein. *Inform. Med. Unlocked* **2021**, *23*, 100511. [CrossRef]
44. Salau, A.O.; Jain, S. Computational modeling and experimental analysis for the diagnosis of cell survival/death for Akt protein. *J. Genet. Eng. Biotechnol.* **2020**, *18*, 1–10. [CrossRef]
45. Jain, S.; Salau, A.O. An image feature selection approach for dimensionality reduction based on kNN and SVM for AkT proteins. *Cogent Eng.* **2019**, *6*, 1599537. [CrossRef]
46. Hancock, J.T.; Khoshgoftaar, T.M. Survey on categorical data for neural networks. *J. Big Data* **2020**, *7*, 1–41. [CrossRef]
47. Ofer, D.; Brandes, N.; Linial, M. The language of proteins: NLP, machine learning & protein sequences. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 1750–1758.
48. McDowall, J.; Hunter, S. InterPro protein classification. *Bioinform. Comp. Proteom.* **2011**, *694*, 37–47.
49. Nanni, L.; Lumini, A.; Brahnam, S. An empirical study of different approaches for protein classification. *Sci. World J.* **2014**, *2014*, 236717. [CrossRef]
50. Diplaris, S.; Tsoumakas, G.; Mitkas, P.A.; Vlahavas, I. Protein classification with multiple algorithms. *Panhellenic Conf. Inform.* **2005**, *7*, 448–456.
51. Cai, C.Z.; Han, L.Y.; Ji, Z.L.; Chen, X.; Chen, Y.Z. SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res.* **2003**, *31*, 3692–3697. [CrossRef]
52. Bock, J.R.; Gough, D.A. Predicting protein–protein interactions from primary structure. *Bioinformatics* **2001**, *17*, 455–460. [CrossRef]
53. Das, S.; Chakrabarti, S. Classification and prediction of protein–protein interaction interface using machine learning algorithm. *Sci. Rep.* **2021**, *11*, 1761. [CrossRef]
54. Karchin, R.; Karplus, K.; Haussler, D. Classifying G-protein coupled receptors with support vector machines. *Bioinformatics* **2002**, *18*, 147–159. [CrossRef]
55. Chen, X.; Gao, Y.; Wang, Y.; Pan, G. Mussel-inspired peptide mimicking: An emerging strategy for surface bioengineering of medical implants. *Smart Mater. Med.* **2021**, *2*, 26–37. [CrossRef]
56. Kazemzadeh-Narbat, M.; Cheng, H.; Chabok, R.; Alvarez, M.M.; De La Fuente-Nunez, C.; Phillips, K.S.; Khademhosseini, A. Strategies for antimicrobial peptide coatings on medical devices: A review and regulatory science perspective. *Crit. Rev. Biotechnol.* **2021**, *41*, 94–120. [CrossRef]
57. Apostolopoulos, V.; Bojarska, J.; Chai, T.-T.; Elnagdy, S.; Kaczmarek, K.; Matsoukas, J.; New, R.; Parang, K.; Lopez, O.P.; Parhiz, H. A global review on short peptides: Frontiers and perspectives. *Molecules* **2021**, *26*, 430. [CrossRef]
58. Charoenkwan, P.; Chiangjong, W.; Hasan, M.M.; Nantasenamat, C.; Shoombuatong, W. Review and Comparative Analysis of Machine Learning-based Predictors for Predicting and Analyzing Anti-angiogenic Peptides. *Curr. Med. Chem.* **2022**, *29*, 849–864. [CrossRef] [PubMed]
59. Fjell, C.D.; Jenssen, H.; Hilpert, K.; Cheung, W.A.; Pante, N.; Hancock, R.E.W.; Cherkasov, A. Identification of novel antibacterial peptides by chemoinformatics and machine learning. *J. Med. Chem.* **2009**, *52*, 2006–2015. [CrossRef] [PubMed]
60. Rondon-Villarreal, P.; Sierra, D.; Torres, R. Machine learning in the rational design of antimicrobial peptides. *Curr. Comput.-Aided Drug Des.* **2014**, *10*, 183–190. [CrossRef] [PubMed]
61. Mousavizadegan, M.; Mohabatkar, H. An evaluation on different machine learning algorithms for classification and prediction of antifungal peptides. *Med. Chem.* **2016**, *12*, 795–800. [CrossRef]
62. Sen, P.C.; Hajra, M.; Ghosh, M. Supervised classification algorithms in machine learning: A survey and review. *Emerg. Technol. Model. Graph.* **2020**, *937*, 99–111.
63. Ivankov, D.N.; Finkelstein, A.V. Prediction of protein folding rates from the amino acid sequence predicted secondary structure. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 8942–8944. [CrossRef]
64. Kunt, I.D.; Crippen, G.M.; Kollman, P.A. Calculation of protein tertiary structure. *J. Mol. Biol.* **1976**, *106*, 983–994.
65. Hagler, A.T.; Barry, H. On the formation of the protein tertiary structure on a computer. *Proc. Natl. Acad. Sci. USA* **1978**, *75*, 554–558. [CrossRef]
66. Salau, A.O.; Jain, S. Feature Extraction: A Survey of the Types, Techniques, Applications. In Proceedings of the 2019 International Conference on Signal Processing and Communication (ICSC), Noida, India, 7–9 March 2019; Volume 75, pp. 158–164.
67. Zur, R.M.; Jiang, Y.P.; Pesce, L.L. Noise injection for training artificial neural networks. A comparison with weight decay and early stopping. *Med. Phys.* **2009**, *36*, 4810–4818. [CrossRef]
68. Lu, H.; Setiono, R.; Huan, L. Effective data mining using neural networks. *IEE Trans. Knowl. Data Eng.* **1996**, *8*, 957–961.
69. Torgyn, S.; Khovanova, N.A. Handling limited datasets with neural networks applications: A small data approach. *Artif. Intell. Med.* **2017**, *75*, 51–63.
70. Rose, P.W.; Prlic, A.; Altunkaya, A.; Bi, C.; Bradley, A.R.; Christie, C.H. The RCSB protein data bank: Integrative view of protein, gene and 3D structural information. *Nucleic Acids Res.* **2016**, *gkw1000*, 271–281 .
71. Rose, P.W.; Bi, C.; Bluhm, W.F.; Christie, C.H.; Dimitropoulus, D.; Dutta, S.; Bourne, P.E. The RCSB Protein data bank: New resources for research and education. *Nucleic Acids Res.* **2012**, *41*, 475–482. [CrossRef]
72. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Bourne, P.E. The protein data bank. *Nucleic Acids Res.* **2000**, *28*, 235–242. [CrossRef]

73. Springs, R.V.; Artymiuk, P.J.; Willet, W. Searching for 3D patterns of amino acids in 3D protein structures. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 412–421. [CrossRef]
74. Abola, E.E.; Bernstein, F.C.; Frances, C.; Koetzle, T.F. The protein data bank. *Neutroms in Biology*; Springer: Boston, MA, USA, 1984.
75. Berman, H.M.; Henrick, K.; Nakamura, H. Announcing the worldwide protein data bank. *Nat. Struct. Mol. Biol.* **2003**, *10*, 980. [CrossRef]
76. Parasuraman, S. Protein data bank. *J. Pharmacol. Pharmacother.* **2012**, *3*, 351 [CrossRef]
77. Sussman, J.L.; Abola, E.E.; Lin, D.; Jiang, J.; Manning, N.O. The protein data bank. *Struct. Biol. Funct. Genom.* **1999**, *54*, 251–264.
78. Fauman, E.B.; Hyde, C. An optimal variant to gene distance window derived from an empirical definition of cis and trans protein QTLs. *BMC Bioinform.* **2022**, *23*, 1–11. [CrossRef]
79. Guarino, A.; Malandrino, D.; Zaccagnino, R. An automatic mechanism to provide privacy awareness and control over unwittingly dissemination of online private information. *Comput. Netw.* **2022**, *202*, 108614. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

3.2 Paper II

Nonlinear Techniques and Ridge Regression as a Combined Approach: Carcinoma Identification Case Study

Authors: Gerardo Alfonso Perez, Raquel Castillo

Mathematics. 2023, 11(1795).

https://doi.org/10.3390/math11081

mathematics

MDPI

Article

Nonlinear Techniques and Ridge Regression as a Combined Approach: Carcinoma Identification Case Study

Gerardo Alfonso Perez * and **Raquel Castillo**

Biocomp Group, Institute of Advanced Materials (INAM), Universitat Jaume I, 12071 Castelló de la Plana, Spain
* Correspondence: ga284@cantab.net

Abstract: As more genetic information becomes available, such as DNA methylation levels, it becomes increasingly important to have techniques to analyze such data in the context of cancers such as anal and cervical carcinomas. In this paper, we present an algorithm that differentiates between healthy control patients and individuals with anal and cervical carcinoma, using as an input DNA methylation data. The algorithm used a combination of ridge regression and neural networks for the classification task, achieving high accuracy, sensitivity and specificity. The relationship between methylation levels and carcinoma could in principle be rather complex, particularly given that a large number of CpGs could be involved. Therefore, nonlinear techniques (machine learning) were used. Machine learning techniques (nonlinear) can be used to model linear processes, but the opposite (linear techniques simulating nonlinear models) would not likely generate accurate forecasts. The feature selection process is carried out using a combination of prefiltering, ridge regression and nonlinear modeling (artificial neural networks). The model selected 13 CpGs from a total of 450,000 CpGs available per patient with 171 patients in total. The model was also tested for robustness and compared to other more complex models that generated less precise classifications. The model obtained (testing dataset) an accuracy, sensitivity and specificity of 97.69%, 95.02% and 98.26%, respectively. The reduction of the dimensionality of the data, from 450,000 to 13 CpGs per patient, likely also reduced the likelihood of overfitting, which is a very substantial risk in this type of modelling. All 13 CpGs individually generated classification forecasts less accurate than the proposed model.

Keywords: anal cancer; cervical cancer; algorithm

MSC: 65F30

check for updates

Citation: Alfonso Perez, G.; Castillo, R. Nonlinear Techniques and Ridge Regression as a Combined Approach: Carcinoma Identification Case Study. *Mathematics* **2023**, *11*, 1795. https://doi.org/10.3390/math11081795

Academic Editor: Francesco Calimeri

Received: 1 March 2023
Revised: 3 April 2023
Accepted: 8 April 2023
Published: 10 April 2023

1. Introduction and Literature Review

Some recent articles, such as Deshmukh et al. [1], have estimated that the incidence of anal carcinoma is increasing at 2.7% per year. They also estimated a similar trend for mortality. Similar results were found by Eng et al. [2]. They estimated a 3.1% increase in the mortality rate. Anal and cervical carcinomas are not yet well understood [3–5]. Articles, such as Melbye and Sprogel [6] and Rabkin et al. [7], have mentioned that anal and cervical cancers have common risk factors and other similarities. Parallels between these two illnesses have been mentioned in the existing literature for decades [8–10]. There is increasing research pointing to a link between anal and cervical carcinomas and the human papillomavirus (HPV) with causal relationship or a strong link mentioned in several articles, such as Darrangh and Winkler [11], Franceschi and De Vuyst [12], Škamperle et al. [13] and Ryan et al. [14]. De Sanjose et al. [15] mentions that HPV has been established as a "central and necessary cause of cervical cancer". Immunosuppressed patients, such as HIV patients, have a higher likelihood of developing this type of cancer [16]. Cancer is in fact a common comorbidity in HIV patients [17–19].

Varnani et al. [20] found a sensitivity and specificity of 93.6% and 80.0%, respectively, in a histological analysis of biopsies of suspected anal carcinoma patients. Van der Zee et al. [21] found a similar specificity (79%) when modeling the risk of anal carcinoma in

HIV-positive patients using as an input DNA methylation data. Other authors have found similar results [22,23].

Alterations of DNA methylation in anal carcinoma have been mentioned in several articles, such as Zhang et al. [24]. The authors of this paper concluded that aberrant methylation is frequent in anal carcinomas. Some articles, such as Siegel et al. [25], have studied changes in methylation levels in both cervical and anal carcinomas, finding also changes in the methylation patterns. Machine learning techniques [26] are an increasingly important tool in many non-medical [27,28] and medical research areas [29–31], and cancer research in no exception [32–34]. Some authors, such as Cuocolo et al. [35], have mentioned that machine learning "could become an essential part of. . . oncological screening". Other authors such as Forsch et al. [36] and Kourou et al. [37] have concluded similarly. There are some interesting articles applying machine learning techniques in the context of carcinomas. For instance, Huang et al. [38] used deep neural networks applied to DNA methylation data aiming to predict outcomes for patients. Nartowt et al. [39] applied an artificial neural network approach for scoring colorectal cancer using self-reported personal health data, achieving a sensitivity and specificity of 57% and 89%, respectively. Methylation data have been used in the analysis of other cancers, such as lung carcinomas (Marchevsky [40], Ligor et al. [41]), glioblastoma (Calabrese et al. [42]), endometrial cancer (Pergialiotis et al. [43]) and gastric cancer (Zhang et al. [44]). Lin et al. [45] used a LASSO approach, which is a special case of ridge regression, in the analysis of the relationship between the expression of m6A RNA methylation and hepatocellular carcinoma prognosis. Butcher and Beck [46] also used a LASSO approach in the context of colon cancer (but no machine learning techniques such as neural networks). Zhong et al. [47] also used the LASSO approach and concluded that this approach with linear regression models has limited prediction power. Cancer screening methods for anal carcinoma (e.g., occult blood test) and cervical carcinoma (e.g., pap smear) are well established. Methylation changes might be able to be detected (but this would need to be tested by further experimental data) before there is occult blood. It can also potentially be used for targeted medicine, i.e., DNA methylation profiles can potentially be used to try to assign more suitable treatment options, according to their methylation profile, to patients.

There are several articles in the existing literature highlighting the applicability of artificial neural networks in the context of nonlinear processes. For example, Zhang et al. [48] applied this technique to nonlinear time series. Liu et al. [49] proposed a multilevel artificial neural network nonlinear equalizer for millimiter-wave mobile fronthaul systems. There are also several papers related to nonlinear control processes, see for instance Cong et al. [50].

There are other ways to carry out this type of analysis. For instance, it is possible to use logistic regression [51] instead of artificial neural networks. There are advantages and disadvantages of using these techniques. Tu [52] mentioned that one of the main advantages of artificial neural networks is their ability to implicitly detect complex nonlinear relationships as well as the ability to detect all possible combinations between predictor variables. One of the disadvantages mentioned by Tu when comparing artificial neural networks and logistic regression was the black box behavior of artificial neural networks with some of the models created being potentially very complex and difficult to interpret.

Objectives

The main objective of this paper is to distinguish between healthy control patients and patients with anal or cervical carcinoma using DNA methylation data and an algorithm combining ridge regression with nonlinear techniques, such as artificial neural networks.

2. Materials and Methods

2.1. Data

The data were obtained from the GEO database with accession code GSE 186859 (publicly available), containing 171 samples of genomic DNA, of which 152 are anal and cervical carcinomas as well as pre-tumours (AIN3 with 13 cases and CIN3 with 9 cases), and the rest are control healthy patients. The dataset consists of 28 cervical samples and 143 anal samples. The data were obtained using the standard illumina protocol, and the chips were scanned on a HiScanSQ System. The researchers that collected the data preprocessed it by performing background correction and normalization using the minfi Bioconductor software in R. Given the relatively low number of pre-tumor cases, the tumor and pre-tumor cases were combined into a single category, which assumes that pre-tumors and tumors have altered DNA methylation levels compared to a healthy individual. There are approximately 450,000 CpGs per patient.

2.2. Notation

The CpG methylation data (X) are represented (Equation (1)) in a matrix form [53]:

$$X = \begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots & \vdots \\ x_{m1} & x_{m2} & x_{m3} & \cdots & x_{mn} \end{pmatrix} \quad (1)$$

Each column represents the methylation data for a given patient, with each row representing the same CpG across different patients. The methylation level is a percentage value expressed as a number ranging from 0 (not methylated) to 1 (fully methylated). As an example of this nomenclature, X_{21} represents the methylation data for patient 1 in CpG 2. It is also convenient to have a vector (Equation (2)) distinguishing between control and patients $(Y_i = \{0, 1\})$.

$$Y = \{y_1, y_2, \ldots, y_n\} \quad (2)$$

2.3. Preliminary Filtering

As usual in nonlinear models, the data need to be divided into a training and a testing dataset. The testing dataset contains approximately 20% of the total data. Furthermore, 10% of the data (training dataset) were used as validation data. We carried out cross-validation 10 times. There are several interesting papers covering validation, see for instance [54,55]. The testing dataset was not used during the training phase. The reported measures, such as accuracy, sensitivity and specificity, are those obtained in the testing dataset (unused duting the trainign phase). A preliminary step consists in filtering each CpG $(X^t = \{x_{t1}, x_{t2}, x_{t3}, \ldots, x_{tn}\})$ individually using binomial regression. In this regression, the independent variable is the methylation level for each CpG across all patients in the training dataset, and the dependent variable is Y (Equation (2)). In this first step, all the CpGs with a p-value bigger than 0.05 were excluded from the analysis.

2.4. Variance Filtering

After this preliminary filtering, an additional filtering was carried out. In this step, the k CpGs with the highest variance were selected. The idea behind this approach is that in the extreme, a CpG that does not have any variation would not be useful as an input for an algorithm that tries to distinguish between control cases and patients.

2.5. Combined Ridge Regression and Nonlinear Modeling

It is possible to further reduce the dimensionality of the data using an approach such as ridge regression [56–58]. This approach automatically reduces the dimensionality of the data by making some of the coefficients in the regression equal to zero. The number of coefficients made equal to zero depends on the parameter α in the ridge regression. In principle, there is no indication that the relationship between the level of DNA methylation and the presence or absence of a tumor should follow a linear relationship. Hence, a nonlinear approach (artificial neural networks) was followed. In this way, the ridge regression selects the CpGs that are then used as inputs in the nonlinear model. The neural network accuracy will depend on factors such as the number of neurons (l) used. Hence, we have the following optimization problem (Equations (3)–(5)). The artificial neural network uses a scaled conjugate gradient backpropagation as a training algorithm, a hidden layer consisting of a hyperbolic tangent sigmoid and an output layer with a softmax transfer function. The training algorithm was used only with the training dataset.

$$\max_{l^*,\alpha^*} \quad f(l,\alpha) \tag{3}$$

$$\text{s.t.} \quad l \leq l_{max}, \tag{4}$$

$$0 < \alpha \leq 1. \tag{5}$$

where l is the number of neurons in the artificial neural network, α is the $\alpha - parameter$ in the ridge regression, and f is a function measuring the goodness of the binary forecast (patient vs. control) of the model output compared to the actual values. This function (f) can be for example the accuracy of the model or the sensitivity or specificity of the model. This task can be performed following a grid approach (Algorithm 1):

Algorithm 1 Grid approach optimization (l_i, α_j)

```
Input: $l_i,\ \alpha_j$
Output: $f_{ij}(l_i,\alpha_j)$
1. Create a grid of values for $l_i = \{l_1, l_2, l_3, \ldots, l_{max}\}$
2. Create a grid of values for $\alpha_j = \{\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_{max}\}$
3. Estimate forecast $(F)$ of the status of patients $F = F_{ij}(l_i,\alpha_j)$
4. Estimate goodness of fit to the binary classification $f = f_{ij}(l_i,\alpha_j)$
5. Repeat steps 3 and 4 q times and obtain mean values
6. Select
       $\sup \left( \frac{1}{q} \sum_{s=1}^{q} f_{ij}^{s} - g(i) \right) = \bar{f}_{ij}^{*}$
       s.t. $l \leq l_{max},$
            $0 < \alpha \leq 1.$
where $g(i)$ is a penalty function of the type $g(i) = \beta \cdot i$
```

With the type of approach presented, it is also necessary to carry out a robustness analysis in which, after $\bar{f}_{ij}^{*}$ is obtained (and hence i and j fixed), the modeling needs to be repeated r times. This step is necessary given the random initialization of the weights in neural networks that result in different outputs, even if the inputs and the structure of the neural network remain unchanged. The value k (variance filtering) needs to be chosen in order for the grid approach to be computationally feasible. Another important step is modelling each CpG individually $(x^t = \{x_{t1}, x_{t2}, x_{t3}, \ldots, x_{tn}, \})$ to study the potential case in which any of the CpGs individually might generate results comparable to the previously generated model.

An alternative to Algorithm 1, in which the optimization is carried out on the number of neurons (l_i) and the $\alpha - factor$ (α_i) of the ridge regression, would be to expand it to include a variable number of layers (κ) as well as adding different types of penalty functions. This can be seen in Algorithm 2.

The purpose of the penalty function is to penalize overly complex model structures that could potentially reduce the generalization capability of the model.

Algorithm 2 Grid approach optimization (l_i, α_j,κ_u)

```
Input: $l_i$, $\alpha_j$,$\kappa_u$
Output: $f_{iju}(l_i, \alpha_j, \kappa_u)$
1. Create a grid of values for $l_i = \{l_1, l_2, l_3, \ldots, l_{max}\}$
2. Create a grid of values for $\alpha_j = \{\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_{max}\}$
3. Create a grid of values for $\kappa_u = \{\kappa_1, \kappa_2, \kappa_3, \ldots, \kappa_{max}\}$
4. Estimate forecast $(F)$ of the status of patients $F = F_{iju}(l_i, \alpha_j, \kappa_u)$
5. Estimate goodness of fit to the binary classification $f = f_{iju}(l_i, \alpha_j, \kappa_u)$
6. Repeat steps 4 and 5 q times and obtain mean values
7. Select
    $$\begin{aligned} \sup \quad & \left( \frac{1}{q} \sum_{s=1}^{q} f_{iju}^{s} - g(i) \right) = \bar{f}_{iju}^{*} \\ \text{s.t.} \quad & l \leq l_{max}, \\ & 0 < \alpha \leq 1, \\ & \kappa \leq \kappa_{max}. \end{aligned}$$
where in this case, the penalty function can be $g(i, \kappa) = \beta_1 \cdot i + \beta_2 \cdot \kappa$ or a quadratic expression $g(i, \kappa) = \beta_1 \cdot i^2 + \beta_2 \cdot \kappa^2$
```

3. Results

After the initial pre-filtering (excluding CpGs with a p value bigger than 0.05), the 200 CpGs ($\kappa = 200$) with the highest variance were selected. As previously mentioned, the assumption is that CpGs with no or very little variance will be of limited use as an input for a classification algorithm. The value of $\kappa = 200$ was selected in order to make the calculations computationally feasible while at the same time maintaining a relatively high number of CpGs. Then, Algorithm 1 was applied to the filtered data (containing 200 CpGs per patient). As described in Section 2, the algorithm tries to find a suitable combination of number of CpGs, which are a function of the α parameter in the ridge regression, and the number of neurons. For clarity purposes, in Figure 1, a graph can be seen showing the results for a given number of neurons and the accuracy at the different α values. A sample of the goodness of the model for a specific configuration can be seen in the ROC curves in Figure 2.

Algorithm 1 then expands this approach for a grid of different numbers of neurons, as can be seen in Figure 3. This approach resulted in a model with only 13 CpGs selecting an accuracy of 97.69%. The specificity and sensitivity of the model were 98.26% and 95.02%, respectively. The number of neurons (l) selected was 790. The average methylation level for these 13 CpGs (for control and patients) can be seen in Figure 4. The list of these 13 CpGs can be found in the Appendix A (Table A1).

It is important to obtain a robust model in which the results are hopefully repeatable. In order to test the robustness of the model, the simulation was repeated 1000 times with the same inputs and network structure. The random initialization of the weights leads to changes in the classification forecast of the model even with the same inputs and network structure. In Figure 5, a histogram can be found showing the resulting accuracy of these simulations. It can seem that it is relatively tightly centered with no frequent outliers. It is also important to analyze each of these CpGs individually. No single CpG has a mean accuracy above 88.94%. Accuracy for each CpG (individually) can be seen in Figure 6.

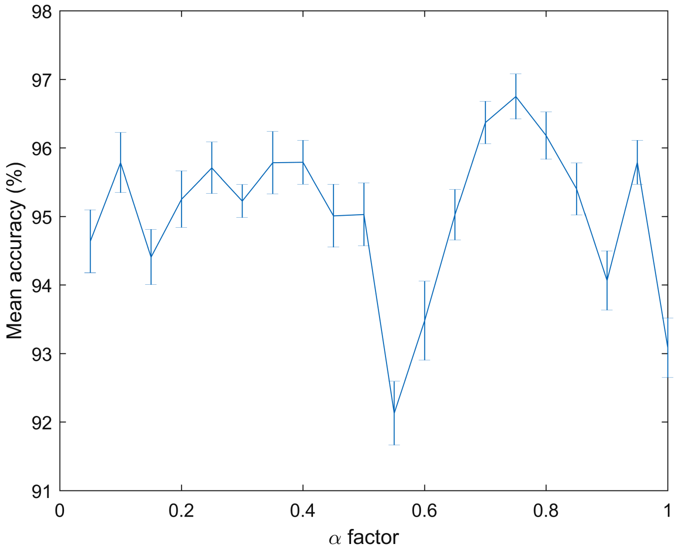


Figure 1. Graph showing the accuracy obtained when fixing the number of neurons and changing the α factor.

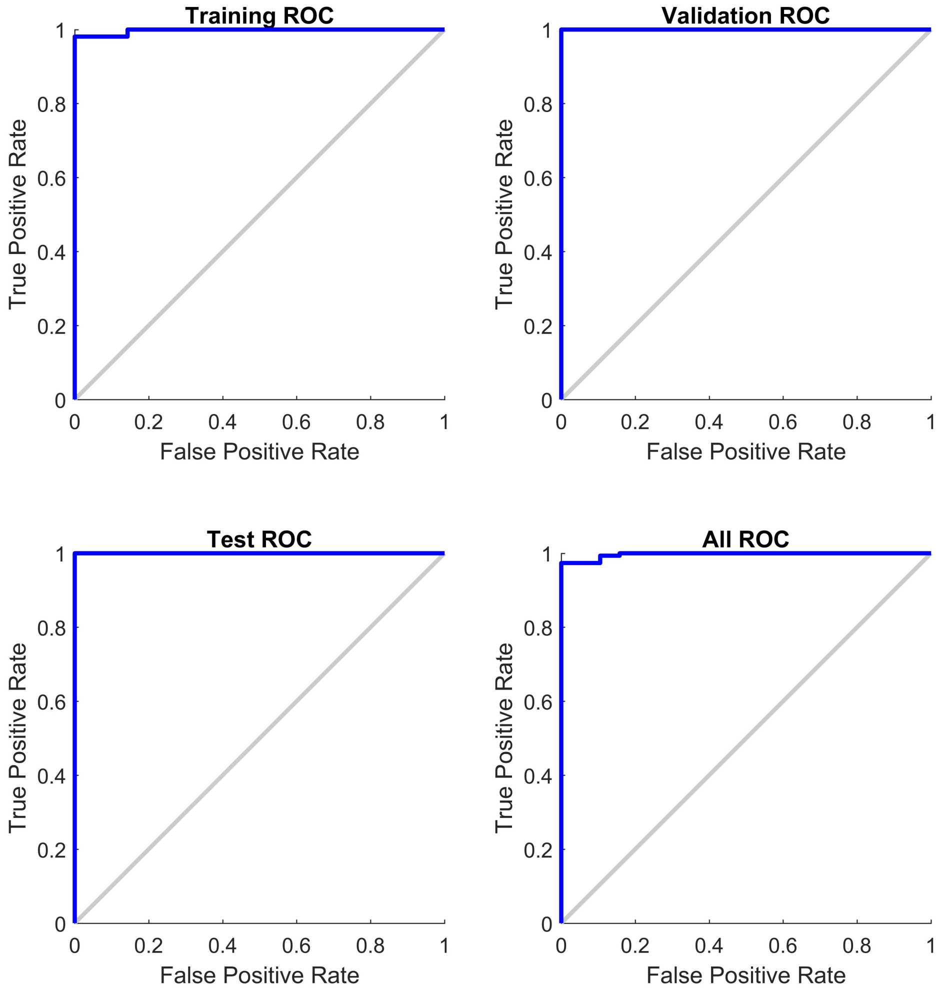


Figure 2. ROC sample curve for one of the estimations.

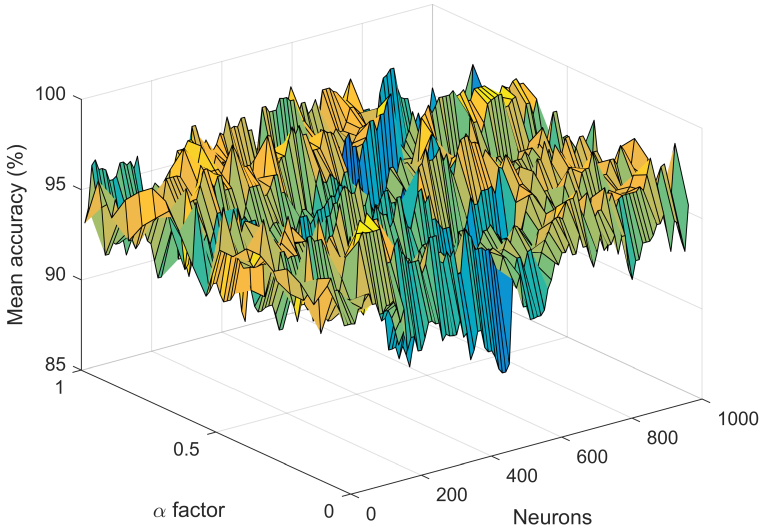


Figure 3. Accuracy obtained using Algorithm 1 (grid approach varying the number of neurons and α factor in a grid).

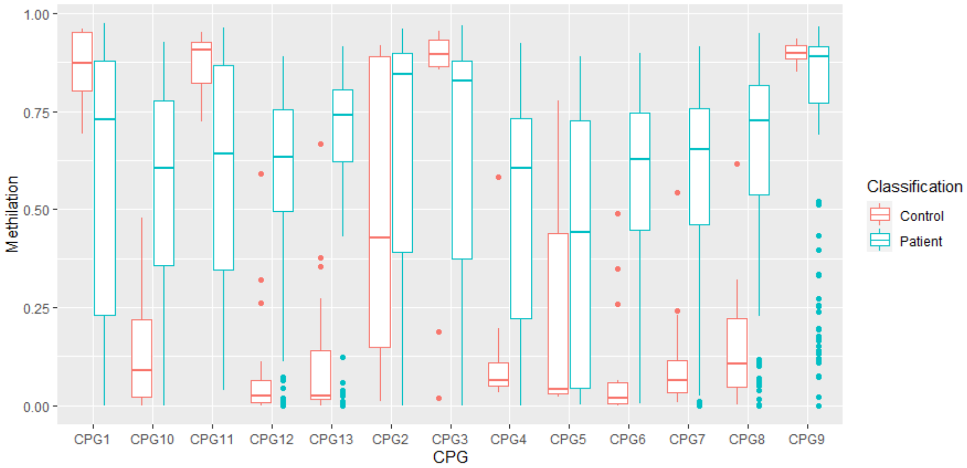


Figure 4. Mean methylation values for patients and control cases.

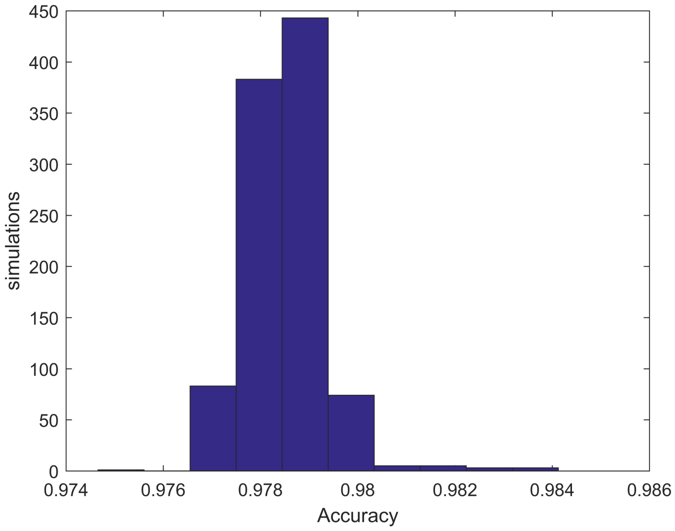


Figure 5. Histogram of the accuracy obtained in 1000 simulations.

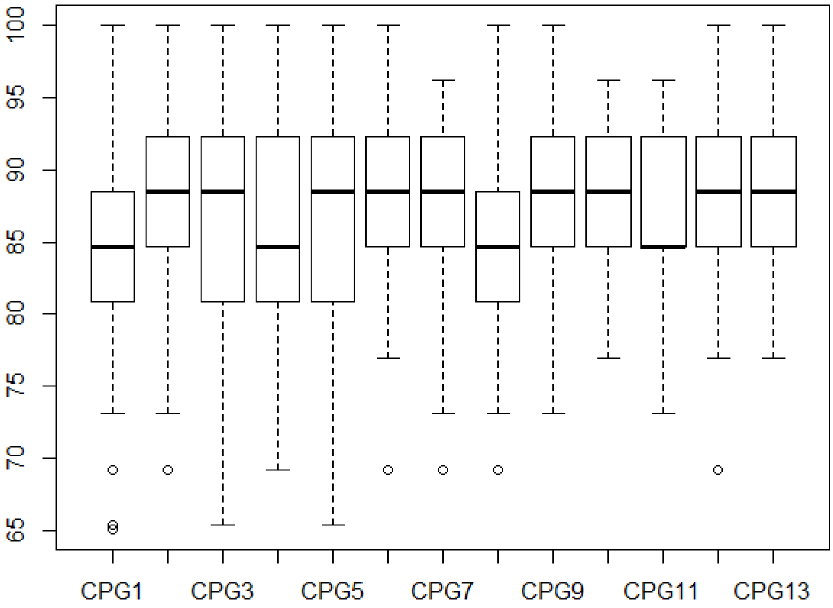


Figure 6. Classification accuracy (%) of each CpG individually.

In Table 1, the results for Algorithms 1 and 2 can be seen. One of the main differences between Algorithms 1 and 2 is that in Algorithm 2, the number of layers was also modified, and two penalty functions were used. The results of the second algorithm were slightly less precise than those in the first algorithm. The best results using Algorithm 1 were with one hidden layer, 790 neurons (with the penalty function $g(i,\kappa) = \beta_1 \cdot i + \beta_2 \cdot \kappa$), and with two hidden layers, 840 neurons (with the quadratic penalty function $g(i,\kappa) = \beta_1 \cdot i^2 + \beta_2 \cdot \kappa^2$). The base case, using all the CpGs and no optimization, is also shown for comparison purposes. The excessive number of inputs in this base case (no filtering) might cause overfitting in the model.

Table 1. Metrics comparing the results of the algorithms.

Metric	Algorithm 1	Algorithm 2 *	Algorithm 2 **	Base
Accuracy	97.69	96.92	94.62	69.23
Specificity	98.26	97.34	98.26	78.95
Sensitivity	95.02	93.33	78.67	42.86

* Algorithm 2 with linear penalty function. ** Algorithm 2 with quadratic penalty function.

4. Discussion

The proposed approach of using DNA methylation data, as inputs, and an algorithm combining ridge regression and artificial neural networks, for the task of differentiating between healthy control individuals and individuals with anal and cervical carcinomas, generated accurate results with specificity and sensitivity higher than ones obtained in other papers in the field. The algorithm selected 13 CpGs from a starting point of approximately 450,000 CpGs per patient. Technological developments have made it possible to obtain such large amounts of methylation data but at the same time have made the analysis of such data challenging. Given that there is no indication that there is a linear relationship between the level of methylation (CpGs) and the presence of anal or cervical carcinoma, the modeling approach was performed with nonlinear techniques such as artificial neural networks. One of the issues with this type of model is the risk of overfitting, particularly in this type of situation in which there is a large number of inputs per patient but a smaller number of patients. In order to reduce this type of risk, it is important to reduce the dimensionality of the data. Additionally, this reduction in the dimensionality can point to CpGs that might be important as biomarkers in the context of the disease. The selected model was tested for robustness, with the classification estimates remaining accurate for the vast majority of the simulations. No individual CpGs, of those 13 selected by the model, achieved a mean accuracy above 88.94%, which is substantially lower than the 97.69% accuracy obtained by the model. Increasing the complexity of the models, by for instance adding more layers to the neural network, did not appear to increase the accuracy of the model. This might be again related to the issue of overfitting. Similarly, adding more complex penalty functions, such as for instance a quadratic function rather than a linear function, did not improve the accuracy.

Limitations and Future Work

There are some limitations in this analysis. For instance, there were only 171 patients analyzed. While the number of patients is not too small, this type of analysis would benefit from a larger cohort of patients. As more data become available, this type of approach can be retested with larger cohorts. Given the larger number of cases of anal carcinoma compared to cervical carcinoma, it is likely that the model will be more precise when classifying anal carcinomas. While there is a clear protocol for obtaining DNA methylation data, there are will always be some small differences in the way that different laboratories collect and present the data. These experimental differences could result in differences in the DNA methylation data and hence reduce the accuracy (and other metrics). It would be very interesting to have time evolution data for the patients that have carcinomas as well as their treatments. It is conceivable that treatment of the patients could potentially be individualized according to their methylation profile, but there is currently, to the best of our knowledge, no available data to actually test this hypothesis. This could be a very interesting area of future research with direct clinical applications.

5. Conclusions

The proposed approach is able to generate an accuracy, sensitivity and specify of classification forecasts of 97.69%, 95.02% and 98.26%, respectively, illustrating that a combination of DNA methylation with nonlinear methods such as artificial neural networks might be useful in the task of identifying patients with a carcinoma. This approach could be complementary to the existing techniques such as occult blood test and pap smear. This is conceivable, but additional testing would be required to support this hypothesis, that DNA methylation changes might be present in the patient before there are clinical indications (occult blood test). This is an important research question that should be addressed in future research. Additionally, it is possible that finding different DNA methylation signatures could be used for personalized treatments. This is another area in which more research would be needed. The model achieved a substantial reduction in the number of CpGs used as input from a starting point of approximately 450,000 to only 13. This is important, as having an excessively large number of inputs could lead to overfitting issues. The combination of these 13 CpGs generated more accurate forecasts that any of them individually. The list of these 13 CpGs can be found in the Appendix A.

Author Contributions: Methodology, G.A.P. and R.C.; software, G.A.P.; validation, G.A.P. and R.C.; formal analysis, G.A.P. and R.C.; investigation, G.A.P. and R.C.; resources, G.A.P. and R.C.; data curation, G.A.P. and R.C.; writing—original draft preparation, G.A.P.; writing—review and editing, G.A.P. and R.C.; visualization, G.A.P. and R.C.; supervision, G.A.P. and R.C.; project administration, G.A.P. and R.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Spanish Ministerio de Ciencia y Tecnología (PID2021-1233320B-C21), and Universitat Jaume I (UJI-B2022-12).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data were obtained from the GEO database with accession code GSE 186859 (publicly available).

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

Appendix A

List of 13 CpGs selected by Algorithm 1.

Table A1. CpG obtained using Algorithm 1.

CpG	CpG Code (GEO)
1	cg15290312
2	cg14331362
3	cg01270299
4	cg07352438
5	cg19393008
6	cg26110710
7	cg21523564
8	cg14487131
9	cg00259849
10	cg14262681
11	cg02263377
12	cg06073449
13	cg18456523

References

1. Deshmukh, A.A.; Suk, R.; Shiels, M.S.; Sonawane, K.; Nyitray, A.G.; Liu, Y.; Gaisa, M.M.; Palefsky, J.M.; Sigel, K. Recent trends in squamous cell carcinoma of the anus incidence and mortality in the United States, 2001–2015. *JNCI J. Natl. Cancer Inst.* **1991**, *338*, 657–659. [CrossRef] [PubMed]
2. Eng, C.; Ciombor, K.K.; Cho, M.; Dorth, J.A.; Rajdev, L.N.; Horowitz, D.P.; Gollub, M.J.; Jacome, A.A.; Lockney, N.A.; Muldoon, R.L. Anal cancer: Emerging standards in a rare disease. *J. Clin. Oncol.* **2022**, *40*, 2774–2788. [CrossRef] [PubMed]
3. Monsrud, A.L.; Avadhani, V.; Mosunjac, M.B.; Flowers, L.; Krishnamurti, U. Programmed death ligand-1 expression is associated with poorer survival in anal squamous cell carcinoma. *Arch. Pathol. Lab. Med.* **2022**, *146*, 1094–1101. [CrossRef] [PubMed]
4. Saiki, Y.; Yamada, K.; Tanaka, M.; Fukunaga, M.; Irei, Y.; Suzuki, T. Prognosis of anal canal adenocarcinoma versus lower rectal adenocarcinoma in Japan: A propensity score matching study. *Surg. Today* **2022**, *52*, 420–430. [CrossRef] [PubMed]
5. Lupi, M.; Brogden, D.; Howell, A.; Tekkis, P.; Mills, S.; Kontovounisios, C. Anal Cancer in High-Risk Women: The Lost Tribe. *Cancers* **2022**, *15*, 60. [CrossRef]
6. Melbye, M.; Sprogel, P. Aetiological parallel between anal cancer and cervical cancer. *Lancet* **1991**, *338*, 657–659. [CrossRef]
7. Rabkin, C.S.; Biggar, R.J.; Melbye, M.; Curtis, R.E. Second primary cancers following anal and cervical carcinoma: Evidence of shared etiologic factors. *Am. J. Epidemiol.* **1992**, *136*, 54–58. [CrossRef]
8. Scholefield, J.H.; Talbot, I.C.; Whatrup, C.; Sonnex, C.; Palmer, J.G.; Mindel, A.; Northover, J.M.A. Anal and cervical intraepithelial neoplasia: Possible parallel. *Lancet* **1989**, *334*, 765–769. [CrossRef]
9. Palmer, J.G.; Scholffield, J.H.; Coates, P.J.; Shepherd, N.A.; Jass, J.R.; Crawford, L.V.; Northover, J.M.A. Anal cancer and human papillomaviruses. *Dis. Colon Rectum* **1989**, *32*, 1016–1022. [CrossRef]
10. Doggett, S.W.; Green, J.P.; Cantril, S.T. Efficacy of radiation therapy alone for limited squamous cell carcinoma of the anal canal. *Int. J. Radiat. Oncol. Biol. Phys.* **1988**, *15*, 1069–1072. [CrossRef]
11. Darragh, T.M.; Winkler, B. Anal cancer and cervical cancer screening: Key differences. *Cancer Cytopathol.* **2011**, *119*, 5–19. [CrossRef] [PubMed]
12. Franceschi, S.; De Vuyst, H. Human papillomavirus vaccines and anal carcinoma. *Curr. Opin. HIV AIDS* **2009**, *4*, 57–63. [CrossRef] [PubMed]
13. Škamperle, M.; Kocjan, B.J.; Maver, P.J.; Seme, K.; Poljak, M. Human papillomavirus (HPV) prevalence and HPV type distribution in cervical, vulvar, and anal cancers in central and eastern Europe. *Acta Dermatovenerol. Alpina Panon. Adriat.* **2013**, *22*, 1–5. [PubMed]
14. Ryan, D.P.; Compton, C.C.; Mayer, R.J. Carcinoma of the anal canal. *N. Engl. J. Med.* **2000**, *342*, 792–800. [CrossRef]
15. de Sanjose, S.; Bruni, L.; Alemany, L. HPV in genital cancers (at the exception of cervical cancer) and anal cancers. *La Presse Médicale* **2014**, *43*, 423–428. [CrossRef]
16. Williams, G.R.; Talbot, I.C. Anal carcinoma—A histological review. *Histopathology* **1994**, *25*, 507–516. [CrossRef]
17. Sumner, L.; Kamitani, E.; Chase, S.; Wang, Y. A systematic review and meta-analysis of mortality in anal cancer patients by HIV status. *Histopathology* **2022**, *76*, 102069. [CrossRef]
18. Naito, T.; Suzuki, M.; Fukushima, S.; Yuda, M.; Fukui, N.; Tsukamoto, S.; Fujibayashi, K.; Goto-Hirano, K.; Kuwatsuru, R. Comorbidities and co-medications among 28 089 people living with HIV: A nationwide cohort study from 2009 to 2019 in Japan. *HIV Med.* **2022**, *23*, 485–493. [CrossRef]

19. Muchengeti, M.; Bartels, L.; Olago, V.; Dhokotera, T.; Chen, W.C.; Spoerri, A.; Rohner, E.; Butikofer, L.; Ruffieux, Y.; Singh, E. Cohort profile: The South African HIV Cancer Match (SAM) Study, a national population-based cohort. *BMJ Open* **2022**, *12*, 053460. [CrossRef]
20. Varnai, A.D.; Bollmann, M.; Griefingholt, H.; Speich, N.; Schmitt, C.; Bollmann, R.; Decker, D. HPV in anal squamous cell carcinoma and anal intraepithelial neoplasia (AIN) Impact of HPV analysis of anal lesions on diagnosis and prognosis. *Int. J. Color. Dis.* **2006**, *21*, 135–142. [CrossRef]
21. van der Zee, R.P.; Richel, O.; van Noesel, C.J.M.; Novianti, P.W.; Ciocanea-Teodorescu, I.; van Splunter, A.P.; Duin, S.; van den Berk, G.E.L.; Meijer, C.; Quint, W. Host cell deoxyribonucleic acid methylation markers for the detection of high-grade anal intraepithelial neoplasia and anal cancer. *Clin. Infect. Dis.* **2019**, *68*, 1110–1117. [CrossRef] [PubMed]
22. Legarth, R.; Helleberg, M.; Kronborg, G.; Larsen, C.S.; Pedersen, G.; Pedersen, C.; Jensen, J.; Nielsen, L.N.; Gerstoft, J.; Obel, N. Anal carcinoma in HIV-infected patients in the period 1995–2009: A Danish nationwide cohort study. *Scand. J. Infect. Dis.* **2013**, *45*, 453–459. [CrossRef] [PubMed]
23. Kreuter, A.; Potthoff, A.; Brockmeyer, N.H.; Gambichler, T.; Swoboda, J.; Stucker, M.; Schmitt, M.; Pfister, H.; Wieland, U. Anal carcinoma in human immunodeficiency virus-positive men: Results of a prospective study from Germany. *Br. J. Dermatol.* **2010**, *162*, 1269–1277. [CrossRef] [PubMed]
24. Zhang, J.; Martins, C.R.; Fansler, Z.B.; Roemer, K.L.; Kincaid, E.A.; Gustafson, K.S.; Heitjan, D.F.; Clark, D.P. DNA methylation in anal intraepithelial lesions and anal squamous cell carcinoma. *Clin. Cancer Res.* **2005**, *11*, 6544–6549. [CrossRef]
25. Siegel, E.M.; Ajidahun, A.; Berglund, A.; Guerrero, W.; Eschrich, S.; Putney, R.M.; Magliocco, A.; Riggs, B.; Winter, K.; Simko, J.P. Genome-wide host methylation profiling of anal and cervical carcinoma. *PLoS ONE* **2021**, *16*, e0260857. [CrossRef]
26. Greener, J.G.; Kandathil, S.M.; Moffat, L.; Jones, D.T. A guide to machine learning for biologists. *Nat. Rev. Mol. Cell Biol.* **2022**, *23*, 40–55. [CrossRef]
27. Salau, A.O.; Jain, S. Feature extraction: A survey of the types, techniques, applications. In Proceedings of the 2019 International Conference on Signal Processing and Communication (ICSC), Noida, India, 7–9 March 2019; pp. 158–164.
28. Guarino, A.; Lettieri, N.; Malandrino, D.; Zaccagnino, R.; Capo, C. Adam or Eve? Automatic users' gender classification via gestures analysis on touch devices. *Neural Comput. Appl.* **2022**, *34*, 18473–18495. [CrossRef]
29. Rabbani, N.; Kim, G.Y.; Suarez, C.J.; Chen, J.H. Applications of machine learning in routine laboratory medicine: Current state and future directions. *Clin. Biochem.* **2021**, *103*, 1–7. [CrossRef]
30. Quazi, S. Artificial intelligence and machine learning in precision and genomic medicine. *Med. Oncol.* **2022**, *39*, 120. [CrossRef]
31. Mueller, B.; Kinoshita, T.; Peebles, A.; Graber, M.A.; Lee, S. Artificial intelligence and machine learning in emergency medicine: A narrative review. *Acute Med. Surg.* **2022**, *9*, 740. [CrossRef]
32. Cai, Z.; Poulos, R.C.; Liu, J.; Zhong, Q. GMachine learning for multi-omics data integration in cancer. *iScience* **2022**, *2022*, 103798. [CrossRef] [PubMed]
33. Capobianco, E. High-dimensional role of AI and machine learning in cancer research. *Br. J. Cancer* **2022**, *126*, 523–532. [CrossRef] [PubMed]
34. Painuli, D.; Bhardwaj, S. Recent advancement in cancer diagnosis using machine learning and deep learning techniques: A comprehensive review. *Comput. Biol. Med.* **2022**, *2022*, 105580. [CrossRef]
35. Cuocolo, R.; Caruso, M.; Perillo, T.; Ugga, L.; Petretta, M. Machine learning in oncology: A clinical appraisal. *Cancer Lett.* **2020**, *481*, 55–62. [CrossRef]
36. Forsch, S.; Klauschen, F.; Hufnagl, P.; Roth, W. Artificial intelligence in pathology. *Deutsches Ärzteblatt Int.* **2021**, *118*, 199. [CrossRef]
37. Kourou, K.; Exarchos, T.P.; Exarchos, K.P.; Karamouzis, M.V.; Fotiadis, D.I. Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* **2015**, *13*, 8–17. [CrossRef]
38. Huang, G.; Wang, C.; Fu, X. Bidirectional deep neural networks to integrate RNA and DNA data for predicting outcome for patients with hepatocellular carcinoma. *Future Oncol.* **2021**, *17*, 4481–4495. [CrossRef] [PubMed]
39. Nartowt, B.J.; Hart, G.R.; Roffman, D.A.; Llor, X.; Ali, I.; Muhammad, W.; Liang, Y.; Deng, J. Scoring colorectal cancer risk with an artificial neural network based on self-reportable personal health data. *PLoS ONE* **2019**, *14*, 0221421. [CrossRef]
40. Marchevsky, A.M. The Use of Artificial Neural Networks for the Diagnosis and Estimation of Prognosis in Cancer Patients. *Outcome Predict. Cancer* **2007**, 243–259. [CrossRef]
41. Ligor, T.; Pater, L.; Buszewski, B. Application of an artificial neural network model for selection of potential lung cancer biomarkers. *J. Breath Res.* **2015**, *9*, 027106. [CrossRef]
42. Calabrese, E.; Rudie, J.D.; Rauschecker, A.M.; Villanueva-Meyer, J.E.; Clarke, J.L.; Solomon, D.A.; Cha, S. Combining radiomics and deep convolutional neural network features from preoperative MRI for predicting clinically relevant genetic biomarkers in glioblastoma. *Neuro-Oncol. Adv.* **2022**, *4*, 60. [CrossRef] [PubMed]
43. Pergialiotis, V.; Pouliakis, A.; Parthenis, C.; Damaskou, V.; Chrelias, C.; Papantoniou, N.; Panayiotides, I. The utility of artificial neural networks and classification and regression trees for the prediction of endometrial cancer in postmenopausal women. *Public Health* **2018**, *164*, 1–6. [CrossRef] [PubMed]
44. Zhang, G.; Xue, Z.; Yan, C.; Wang, J.; Luo, H. A novel biomarker identification approach for gastric cancer using gene expression and DNA methylation dataset. *Front. Genet.* **2021**, *12*, 644378. [CrossRef] [PubMed]

45. Lin, Y.; Yao, Y.; Wang, Y.; Wang, L.; Cui, H. PD-L1 and immune infiltration of m6A RNA methylation regulators and its miRNA regulators in hepatocellular carcinoma. *BioMed Res. Int.* **2021**, *2021*, 1–16. [CrossRef]
46. Butcher, L.M.; Beck, S. Probe Lasso: A novel method to rope in differentially methylated regions with 450 K DNA methylation data. *Methods* **2015**, *72*, 21–28. [CrossRef] [PubMed]
47. Zhong, H.; Kim, S.; Zhi, D.; Cui, X. Predicting gene expression using DNA methylation in three human populations. *PeerJ* **2019**, *7*, 6757. [CrossRef]
48. Zhang, G.P.; Patuwo, B.E.; Hu, M.Y. A simulation study of artificial neural networks for nonlinear time-series forecasting. *Comput. Oper. Res.* **2001**, *28*, 381–396. [CrossRef]
49. Liu, S.; Xu, M.; Wang, J.; Lu, F.; Zhang, W.; Tian, H.; Chang, G. A multilevel artificial neural network nonlinear equalizer for millimetre-wave mobile fronthaul systems. *J. Light. Technol.* **2017**, *35*, 4406–4417. [CrossRef]
50. Cong, S.; Liang, Y. PID-like neural network nonlinear adaptive control for uncertain multivariable motion control systems. *IEEE Trans. Ind. Electron.* **2009**, *56*, 3872–3879. [CrossRef]
51. Wang, H.Y.; Chang, S.C.; Lin, W.Y.; Chen, C.H.; Chiang, S.H.; Huang, K.Y.; Chu, B.Y.; Lu, J.J.; Lee, T.Y. Machine Learning-Based Method for Obesity Risk Evaluation Using Single-Nucleotide Polymorphisms Derived from Next-Generation Sequencing. *J. Comput. Biol.* **2018**, *25*, 1347–1360. [CrossRef]
52. Tu, J.V. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J. Clin. Epidemiol.* **1996**, *49*, 1225–1231. [CrossRef] [PubMed]
53. Alfonso Perez, G.; Castillo, R. Identification of Systemic Sclerosis through Machine Learning Algorithms and Gene Expression. *Mathematics* **2022**, *10*, 4632. [CrossRef]
54. Puleston, D.J.; Buck, M.D.; Klein, G.R.I.; Kyle, R.L.; Caputa, G.; O'Sullivan, D.; Cameron, A.M.; Castoldi, A.; Musa, Y.; Kabat, A.M.; et al. Polyamines and eIF5A Hypusination Modulate Mitochondrial Respiration and Macrophage Activation. *Cell Metab.* **2019**, *30*, 352–363. [CrossRef] [PubMed]
55. Vabalas, A.; Gowen, E.; Poliakoff, E.; Casson, A.J. Machine learning algorithm validation with a limited sample size. *PLoS ONE* **2019**, *14*, e0224365. [CrossRef] [PubMed]
56. McDonald, G.C. Ridge regression. *Wiley Interdiscip. Rev. Comput. Stat.* **2009**, *1*, 93–100. [CrossRef]
57. Marquardt, D.W.; Snee, R.D. Ridge regression in practice. *Am. Stat.* **1975**, *29*, 3–20.
58. Hoerl, A.E.; Kannard, R.W.; Baldwin, K.F. Ridge regression: Some simulations. *Commun.-Stat.-Theory Methods* **1975**, *4*, 105–123. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

3.3 Paper III

Identification of Systemic Sclerosis through Machine Learning Algorithms and Gene Expression

Authors: Gerardo Alfonso Perez, Raquel Castillo

Mathematics. 2022, 10(4632).

https://doi.org/10.3390/math102446

mathematics

Article

Identification of Systemic Sclerosis through Machine Learning Algorithms and Gene Expression

Gerardo Alfonso Perez * and Raquel Castillo

Biocomp Group, Institute of Advanced Materials (INAM), Universitat Jaume I, 12071 Castelló, Spain
* Correspondence: ga284@cantab.net or al409883@uji.es

Abstract: Systemic sclerosis (SSc) is an autoimmune, chronic disease that remains not well understood. It is believed that the cause of the illness is a combination of genetic and environmental factors. The evolution of the illness also greatly varies from patient to patient. A common complication of the illness, with an associated higher mortality, is interstitial lung disease (ILD). We present in this paper an algorithm (using machine learning techniques) that it is able to identify, with a 92.2% accuracy, patients suffering from ILD-SSc using gene expression data obtained from peripheral blood. The data were obtained from public sources (GEO accession GSE181228) and contains genetic data for 134 patients at an initial stage as well as at a follow up date (12 months later) for 98 of these patients. Additionally, there are 45 control (healthy) cases. The algorithm also identified 172 genes that might be involved in the illness. These 172 genes appeared in all the 20 most accurate classification models among a total of half a million models estimated. Their frequency might suggest that they are related to the illness to some degree. The proposed algorithm, besides differentiating between control and patients, was also able to distinguish among different variants of the illness (diffuse variants). This can have a significance from a treatment point of view. The different type of variants have a different associated prognosis.

Keywords: systemic sclerosis; gene expression; machine learning

MSC: 62H20; 62H25; 62H99

check for updates

Citation: Alfonso Perez, G.; Castillo, R. Identification of Systemic Sclerosis through Machine Learning Algorithms and Gene Expression. *Mathematics* **2022**, *10*, 4632. https://doi.org/10.3390/math10244632

Academic Editor: Kang Lu

Received: 21 October 2022
Accepted: 3 December 2022
Published: 7 December 2022

1. Introduction

Systemic sclerosis (SSc), also called Scleroderma [1], is an autoimmune [2], relatively uncommon, chronic illness [3] with associated high morbidity and mortality [4,5]; similar to other autoimmune illnesses it is more common in females [6]. There is no curative treatment for SSc but there are some treatment options for commonly associated complications [7–9]. SSc can significantly impact the quality of life of the patient [10] and attack internal organs [11]. The prevalence of the illness appears to vary depending in the geographic location with, for instance, Zhong et al. [12] estimating a prevalence in the US of approximately 50 cases per 100,000, while Englert et al. finding a lower prevalence in Sidney, Australia of approximately 8.6 patients per 100,000 [13]. The illness has a higher prevalence in some ethnics groups such as African American [14] and Native American. Banabe et al. [15] concluded that females in the First Nation (Native American) in Canada have a prevalence twice as high as females in the rest of the population. The usual age of onset of the illness is between 30 to 60 years old with Hoffman-Vold et al. [16] estimating a mean onset age of 47 in a study covering the Norwegian population. The illness is characterized by excessive collagen content in tissue, fibrosis and vascular damage [17–19]. The causes of SSc are not yet well understood and it is theorized that it is likely caused by a combination of genetic predisposition [20] and environmental factors [21,22]. It is very likely that there is a genetic component with Varga and Abraham [23] estimating that the illness is more frequent in families (1.6%) than in the general population (0.026%). There are also likely some environmental triggers and while many hypothesis have been formulated, such as exposure to silica dust (miners disease) [24,25], certain chemical compounds (toluene or benzene) drugs (cocaine or carbidopa) and infections [25–27], there is, to the best of our knowledge, no irrefutable proof of the link between these factors and SSc, which suggests some complex interaction between genetic and environmental factors. SSc is also associated with the increased likelihood of some malignancies [28].

There are different variants such as Limited cutaneous systemic sclerosis (also referred as CREST) and Diffuse systemic sclerosis [29,30]. Roadnan et al. [31] compared the skin collagen content of 117 individuals with SSc (107 of the diffuse variant and 40 with the CREST variant) and compared it with 58 control (healthy) individuals finding a significant thickening of the skin, associated with higher collagen deposits. It should be noted that there is still some disagreement in the existing literature in the classification of SSc variants [32]. Interstitial lung disease is a relatively common complication of SSc that significantly worsens the prognosis [33].

While there is no curative treatment for the illness, over the years, multiple treatment options for the related complications, such as some treatment options for renal crisis (using ACE inhibitors) [34], have been developed, improving survival rates [35]. The evolution of the illness varies significantly from patient to patient [6]. As previously mentioned, some variants of SSc such as the diffuse variants [36] have a worse prognosis [37]. In this paper, we focus on Interstitial lung disease systemic sclerosis (SSc-ILD) with and without diffuse cutaneous involvement. According to figures from the US FDA, approximately half of the patients with Scleroderma have ILD-SSc. Some researchers, such as Boussone and Mouthon [38] have estimated a higher percentage. According to these authors, approximately 75% of SSc patients develop some level of ILD. They do, however, mention that only a small fraction of these patients evolve into critical respiratory insufficiency. Goh et al. [39] mentioned that in some cases it might be challenging to obtain a firm diagnosis on SSc-ILD by using the classical approach of pulmonary function tests (PFTs) and high resolution computed tomography (HRCT) [40]. SSc-ILD typically present fibrosis in the lower section of the lungs. In recent years, there has been a substantial amount of research targeting a reduction in mortality on ILD-SSc [41,42]. In an illness as heterogeneous as ILD-SSc, it seems important to develop biomarkers for its detection, ideally at early stage, as well as for distinguishing different variants such the presence of diffuse cutaneous involvement. Most of the existing literature uses the clinical presentation of the patient [4] and/or imaging rather than a genetic big data approach for the identification of the illness. We have followed a gene expression approach. This is supported by indications of a genetic component in the illness [43–45]. We present a new algorithm for the selection of the genes considered. In an interesting article, Jamin et al. [46] use neural networks to the same classification task but using electronic health records (clinical factors). Our proposed approach is complementary to this type of analysis, as it uses a different set of information. Another complementary approach is the one used by Akay et al. [47], in which skin images are used as an input for a machine learning algorithm. These approaches use clinical manifestations and images of skin lesions. A genetic approach has the potential advantage of not requiring clear clinical manifestations such as skin lesions.

2. Aims

The main objectives of this paper are to be able to distinguish between control and SSc patients using gene expression data analyzed with machine learning techniques as well as to differentiate between different variants of the illness using the same approach.

3. Materials and Methods

Assuming that there are n genes analyzed per patient and m patients. The information for each patient can be stored in the form of a column vector x_i.

$$X_i = \left\{ \begin{array}{c} X_i^1 \\ X_i^2 \\ X_i^3 \\ . \\ . \\ X_i^n \end{array} \right\} \tag{1}$$

with x_i^1 representing the expression of the first gene for patient i, the information for all the patients can be expressed in a matrix form, as follows.

$$X = \begin{pmatrix} X_1^1 & X_2^1 & \dots & X_m^1 \\ X_1^2 & X_2^2 & \dots & Xm^2 \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ X_1^n & X_2^n & \dots & X_m^n \end{pmatrix} \tag{2}$$

There is also an associated variable $Y_i = \{0, 1\}$ describing the status of the patient with $\{0\}$ indicating a control (healthy) individual and $\{1\}$ indicating a patient with the illness. This can be represented with a row vector (including all patients).

$$Y = \{y_1, y_2, \dots, y_m\} \tag{3}$$

3.1. Algorithm

1. The first step entails dividing the data into the control and patient subsets.

$$X_c = \begin{pmatrix} X_{1,c}^1 & X_{2,c}^1 & \dots & X_{l,c}^1 \\ X_{1,c}^2 & X_{2,c}^2 & \dots & X_{l,c}^2 \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ X_{1,c}^n & X_{2,c}^n & \dots & X_{l,c}^n \end{pmatrix} \tag{4}$$

$$Y_c = \{y_1, y_2, \dots, y_l\} \tag{5}$$

$$X_p = \begin{pmatrix} X_{l+1,p}^1 & X_{l+2,p}^1 & \dots & X_{m,p}^1 \\ X_{l+1,p}^2 & X_{l+2,p}^2 & \dots & X_{m,p}^2 \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ X_{l+1,p}^n & X_{l+2,p}^n & \dots & X_{m,p}^n \end{pmatrix} \tag{6}$$

$$Y_p = \{y_{l+1}, y_{l+2}, \dots, y_m\} \tag{7}$$

2. Estimating the mean values for each gene in each subset

$$X_c^{me} = \left\{ \begin{matrix} M_c^1 \\ M_c^2 \\ . \\ . \\ . \\ M_c^n \end{matrix} \right\} \tag{8}$$

$$X_p^{me} = \left\{ \begin{matrix} M_p^1 \\ M_p^2 \\ . \\ . \\ . \\ M_p^n \end{matrix} \right\} \tag{9}$$

3. Compare the expression value for each gene on both sets

$$C^j = \frac{M_p^j}{M_c^j} \tag{10}$$

4. If $c^j < c_{th}^j$ (with c_{th}^j a predefined threshold) then eliminate the gene from both subsets. Hence:

$$X_{c*} = \begin{pmatrix} X_{1,c}^1 & X_{2,c}^1 & \dots & X_{l,c}^1 \\ X_{1,c}^2 & X_{2,c}^2 & \dots & X_{l,c}^2 \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ X_{1,c}^{n*} & X_{2,c}^{n*} & \dots & X_{l,c}^{n*} \end{pmatrix} \tag{11}$$

$$Y_c = \{y_1, y_2, \dots, y_l\} \tag{12}$$

$$X_p = \begin{pmatrix} X_{l+1,p}^1 & X_{l+2,p}^1 & \dots & X_{m,p}^1 \\ X_{l+1,p}^2 & X_{l+2,p}^2 & \dots & X_{m,p}^2 \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ X_{l+1,p}^{n*} & X_{l+2,p}^{n*} & \dots & X_{m,p}^{n*} \end{pmatrix} \tag{13}$$

$$Y_p = \{y_{l+1}, y_{l+2}, \dots, y_m\} \tag{14}$$

with $n* < n$. This process results in a reduction in the number of genes taken into consideration. The data can now be consolidated into a X^* matrix and a Y^* vector containing both control and patients.

$$X^* = \begin{pmatrix} X_1^1 & X_2^1 & \dots & X_m^1 \\ X_1^2 & X_2^2 & \dots & X_m^2 \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ X_1^{n*} & X_2^{n*} & \dots & X_m^{n*} \end{pmatrix} \tag{15}$$

$$Y^* = \{y_1, y_2, \ldots, y_m\} \tag{16}$$

5. Divide the data into a testing and a training datasets with both containing control and patients.

$$X_{Tr} = \begin{pmatrix} X^1_{Tr,1} & X^1_{Tr,2} & \cdots & X^1_{Tr,s} \\ X^2_{Tr,1} & X^2_{Tr,2} & \cdots & X^2_{Tr,s} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ X^{n*}_{Tr,1} & X^{n*}_{Tr,2} & \cdots & X^{n*}_{Tr,s} \end{pmatrix} \tag{17}$$

$$Y_{Tr} = \{y_1, y_2, \ldots, y_s\} \tag{18}$$

$$X_{Ts} = \begin{pmatrix} X^1_{Ts,s+1} & X^1_{Ts,s+2} & \cdots & X^1_{Ts,m} \\ X^2_{Ts,s+1} & X^2_{Ts,s+2} & \cdots & X^2_{Ts,m} \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ X^{n*}_{Ts,s+1} & X^{n*}_{Ts,s+2} & \cdots & X^{n*}_{Ts,m} \end{pmatrix} \tag{19}$$

$$Y_{Ts} = \{y_{s+1}, y_{s+2}, \ldots, y_m\} \tag{20}$$

6. Choose a classification technique (F), such as an artificial neural network.
7. Train the classification technique (F) with the training data $(F(X_{tr}, Y_{Tr}))$.
8. Estimate the classification forecast (CF) using the trained algorithm.

$$CF = \{CF_1, CF_2, \ldots, CF_s\} \tag{21}$$

9. Compare the classification forecasts (CF) with the the actual values Y_{Tr}.
10. If $C_i = Y_i$ then $Acc_i = 1$ otherwise $Acc_i = 0$. Estimate mean accuracy.

$$Acc^m = \frac{\sum Acc_i}{s} \tag{22}$$

Similarly estimate the sensitivity (S^m).

11. This is the first iteration

$$Se(1) = S^m \tag{23}$$

12. Then, define an integer $\kappa \in (1, a_n)$ with $a_n < n*$.
13. Eliminate κ genes randomly chosen from the previous group of $n*$ genes.
14. Repeat steps 7 to 11, estimating the new sensitivity S^m_t. If $S^m_t > Se(1)$ then the new configuration (group of genes) is accepted, else $Se(2) = Se(1)$ and revert to the previous configuration.
15. Repeat until the maximum number of iterations (i_{max}) is reached.
16. Repeat entire process j_{max} times.
17. Select the configuration with the highest sensitivity.

To the best of our knowledge, this is a new algorithm for the identification of relevant genes in the context of SSc. One of the advantages of this algorithm is that it does not require previous knowledge regarding which genes are more relevant in the context of the illness, as they are automatically selected by the algorithm and can potentially select complex combinations of genes.

3.2. Data

Peripheral blood gene expression data was obtained from the publically available database GEO (accession code GSE181228) [48]. The data is composed of 45 healthy control cases, as well as patients with systemic sclerosis-related interstitial lung disease (SSc-ILD), see Figure 1. A total of 134 patients were analyzed at an initial stage (baseline), see Table 1. There was also a follow up test, 12 months later, including 98 patients. The two drugs used in this trial were mycophenolate mofetil (MMF), administered to 65 patients, and cyclophosphamide (CYC) administered to 69 patients. The total number of samples was 277. Our objective is not to replicate this paper [48] but to find biomarkers for the identification of the illness regardless of the actual medication taken.

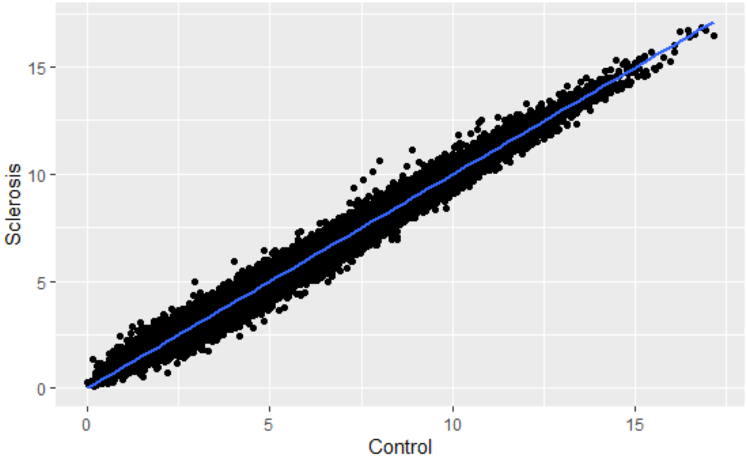


Figure 1. Gene expression in SSc-ILD vs. control patients.

Table 1. Patients characteristics at baseline.

Category	Value
Age	52.4
Male	36
Female	98
White	93
African American	29
Asian	9
Native American	3

The range in age of the patients (at baseline) was from 28 to 79 and there was a large percentage of female (73.1%), consistent with a higher prevalence among the female population of the disease. The majority of the cases 93 (69.4%) were of white race with smaller number of samples of African American (21.6%), Asian (6.7%) and Native American (2.2%). Some of the patients, see Table 2, presented diffuse cutaneous involvement, which has been mentioned as an indicator for the evolution of the illness.

Table 2. Patient with diffuse cutaneous involvement (dc) [1].

	dc	non-dc
Baseline	79	55
12 months	59	38

[1] One of the samples was not identified as either dc or non-dc.

3.3. Classification Algorithm

There are several potential classification algorithms [49] that could be used in the context of this paper. We used artificial neural networks (ANN) [50–52]. This is a well-known and robust technique applied in many different fields. ANNs have been successfully used in the context of SSc identification [53] using as inputs hand photographs of the patients. Similarly, Chassagnon et al. [54] and Chandrasekaran et al. [55] also used neural networks for the assessment of interstitial lung disease in systemic sclerosis using CT images. ANNs are a versatile tool that does not require previous knowledge of the system that it is attempting to model. The ANN used had one hidden layer [56] with 100 neurons. As standard practice, the data were divided into a training and a testing dataset [57,58]. The training dataset contained approximately 66% of the samples. The rest of the samples were included in the testing data set. Only data in the training dataset were used during the training phase of the algorithm. The algorithm in this paper was designed to be flexible, hence other classification techniques, such as support vector machines [59,60], could be potentially used. The required computational time is a factor to be taken into account. Training all the half a million models used in this paper required approximately 197 h (roughly 8.2 days). All the calculations were carried out in Matlab (models' optimization and accuracy estimations were carried out in an automated way) using five Core i5-8265U computers.

4. Results

As described in the methodology, the initial steps of the algorithm included an initial filtering in which the mean values of the gene expression for the control and patient cohorts were estimated. Only genes with a 25% difference in gene expression (absolute value), compared to the base case (control), were included in the analysis.This 25% level was chosen in order to conduct an initial filtering in the data while at the same time had not been too restrictive as the algorithm will further filter the genes. The algorithm then further reduced the number of genes included. As mentioned, a Mote Carlo approach was followed, setting the algorithm to 1000 iterations and repeating the process 500 times, generating half a million models in the process (see Figure 2). The best model resulted in a list of 1157 genes with a average sensitivity, specificity, accuracy and ROC of 74.8%, 95.3%, 92.2% and 86.3%, respectively. As an example, an ROC curve is shown in Figure 3 for a given iteration. There were no improvements when controlling for age, gender or ethnicity. The precision obtained using the algorithm was higher than the base case precision using all genes (see Table 3). The way that the models are constructed, the sensitivity is guaranteed not to decrease from iteration to iteration, but the same cannot be said for the specificity or the overall accuracy of the model (see Figure 4). The list of these 1157 genes can be found in the supplementary files. It was also tested whether the model, using the same genes, is able to differentiate between the diffuse and non-diffuse variants, obtaining a sensitivity of 72.4% (out-of-sample). As in the previous case, the precision obtained using the algorithm was higher than the precision using the base case (all genes), as shown in Table 3.

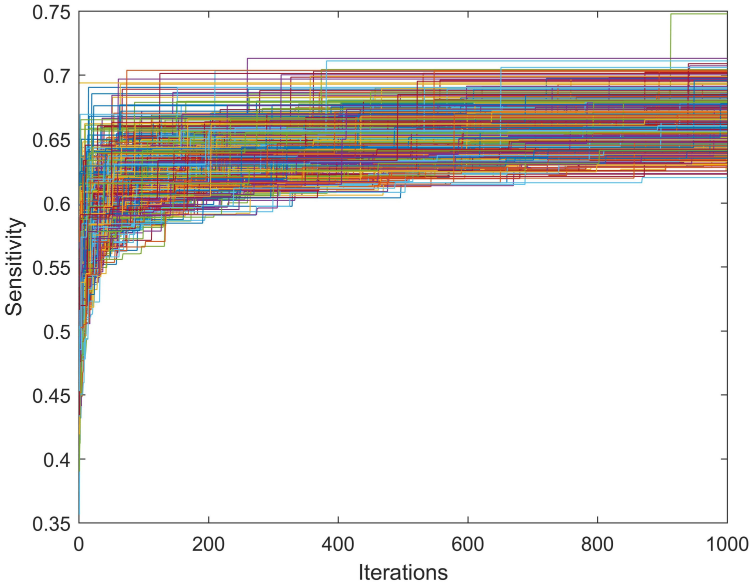


Figure 2. Sensitivity results of the models.

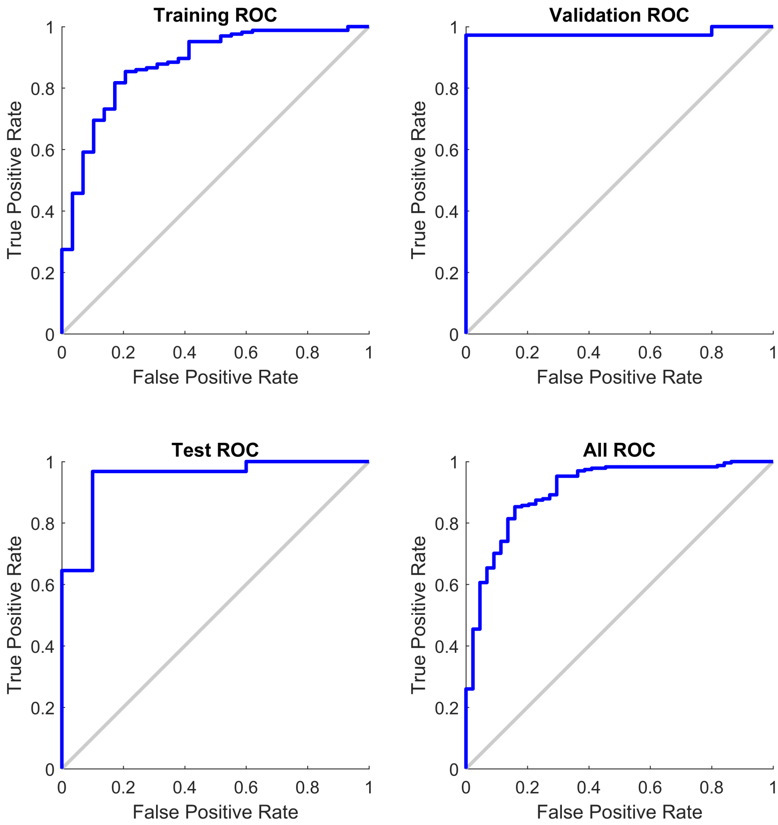


Figure 3. ROC sample for one iteration.

Table 3. Average precision of the model distinguishing SSc and control patients as well as SSc variants (diffuse vs. non-diffuse).

Metric	SSc (Model)	SSc (Base)	Variant (Model)	Variant (Base)
Avg. Sensitivity	0.7478	0.5146	0.7241	0.5152
Avg. Specificity	0.9533	0.8664	0.7000	0.5833
Avg. Accuracy	0.9217	0.8060	0.7101	0.5507
Avg. ROC	0.8632	0.6907	0.6962	0.5549

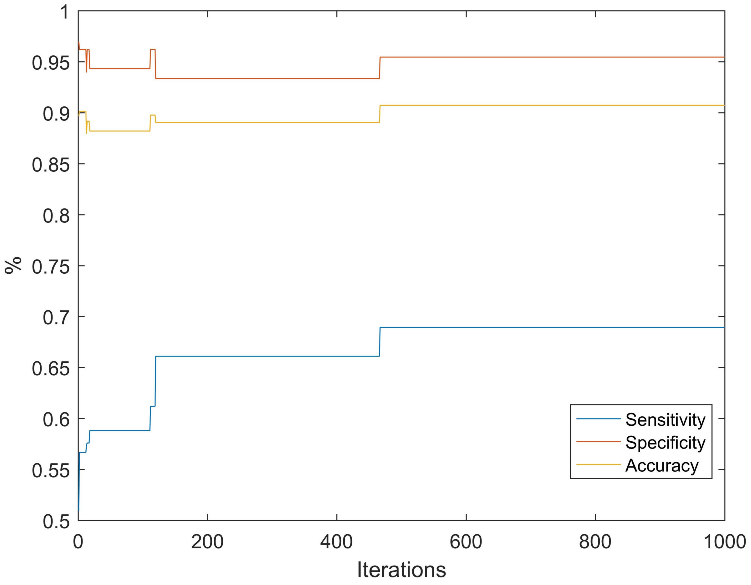


Figure 4. Sensitivity, specificity and accuracy of a sample model.

In Figure 5, the average gene expression is shown for the control and SSc patients. The genes are ordered from the highest to lowest gene expression according to the control data. It can be observed that the SSc data fluctuates more compared to the control data.

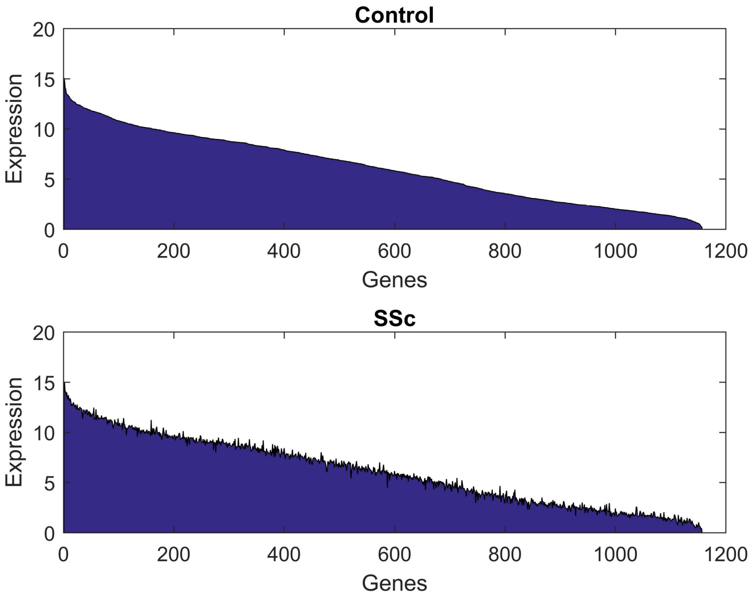


Figure 5. Mean gene expression for controls and patients.

The asymptotic behavior was also tested, increasing the number of iterations to relatively large amounts, such as 50,000 (see Figure 6). There was no indication that substantially increasing the number of iterations necessarily translate into better forecasting precision with the sensitivity reaching a plateau relatively fast. Due to the scale, it is hard to appreciate but in Figure 6 it is shown how the model quickly reaches this plateau. It is also interesting to analyze which genes tend to appear more frequently in the best models. Out of the half a million models calculated, the 20 most accurate were selected and the genes compared. A total of 172 genes appeared in all of these 20 models. The list of these 172 genes can be found in the supplementary material. It is reasonable to assume that the genes that appear more frequently in the most accurate models might, at least potentially, be related to the disease.

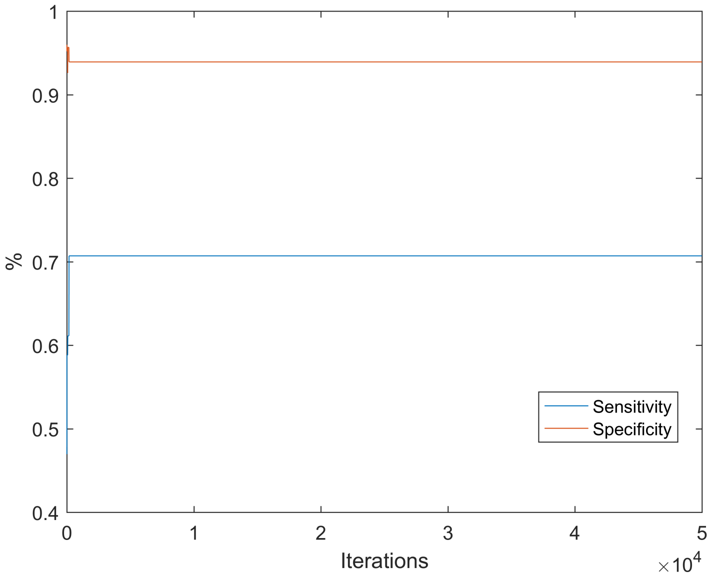


Figure 6. Sample of asymptotic analysis (50,000 iterations).

5. Discussion

Systemic sclerosis is a chronic and potentially life threatening illness which is not yet fully understood. The illness has different variants, such as the diffuse form, with different levels of severity in the prognosis. SSc is believed to be caused by a combination of genetic predisposition and environmental factors. While there is currently no curative therapy, there have been many advances on the treatments of related complications of the illness. Some of these complications are potentially life threatening. One common and severe complication of SSc is interstitial lung disease (ILD). In this paper, we present an algorithm that uses machine learning techniques, applied to gene expression data, to be able to distinguish between control (healthy) patients and patients suffering from interstitial lung disease systemic sclerosis (ILD-SSc). This algorithm selects the genes (and their expression levels) to be included as inputs into machine learning models for the detection of the illness. The precision of this approach is higher than the one obtained using the genes expression for all the available genes. Having biomarkers that are able to identify the illness might be important from an early detection point of view. The accuracy of the presented model was relatively high, at 92%, with a sensitivity of approximately 75%. Our approach is complementary of some of the existing research in this field that use clinical manifestation of the illness. An example of such an approach would be [53] that uses hand photographs and a neural networks classification algorithm or [54] that also uses a neural networks approach but in the case applied to CT images. A potential advantage of using the genetic expression information is that there is no need for the illness to have clear clinical manifestations, such as skin lesions. Milanese et al. [61] achieved an accuracy of 84% using CT texture analysis. Another interesting alternative for the identification and classification of SSc is presented in Filippini et al. [62], in which the authors use hand thermal images and neural networks for diagnosis, achieving an overall accuracy of 84%. Another imaging base paper is Nitkunanantharajah et al. [63], in which the authors use nailfold capillaries imaging, obtaining a high sensitivity of 78.3%.

The approach followed in the algorithm is also allowed for the identification of 172 genes that might potentially have some relevance in the context of ILD-SSc. These 172 genes appeared in all the 20 most accurate models (out of half a million models estimated). The assumption is that given the frequency with which these genes appear in the most accurate models, they might be related to the illness. The proposed algorithm was also able to distinguish between the variants of the illness (diffuse). While the precision was lower that in the previous case (distinguishing between control and patients), it was reasonably high with a sensitivity of approximately 72%. This is reasonable, taking into consideration that the illness is likely not only caused by genetic factors but from a combination of genetic factors and environmental exposures.

Supplementary Materials: The following are available at https://www.mdpi.com/article/10.3390/math10244632/s1.

Author Contributions: Conceptualization, G.A.P. and R.C.; methodology, G.A.P. and R.C.; software, G.A.P.; validation, G.A.P. and R.C.; formal analysis, G.A.P. and R.C.; investigation, G.A.P. and R.C.; resources, G.A.P. and R.C.; data curation, G.A.P. and R.C.; writing—original draft preparation, G.A.P.; writing—review and editing, G.A.P. and R.C.; visualization, G.A.P. and R.C.; supervision, G.A.P. and R.C.; project administration, G.A.P. and R.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Publicly available data. Accession code GSE181228 (GEO).

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

References

1. Sapadin, A.N.; Fleischmajer, R. Treatment of scleroderma. *Arch. Dermatol.* **2002**, *138*, 99–105. [CrossRef] [PubMed]
2. Pattanaik, D.; Brown, M.; Postlethwaite, B.C.; Postlethwaite, A.E. Pathogenesis of systemic sclerosis. *Front. Immunol.* **2015**, *6*, 272. [CrossRef] [PubMed]
3. Domsic, R.; Fasanella, K.; Bielefeldt, K. Gastrointestinal manifestations of systemic sclerosis. *Dig. Dis. Sci.* **2008**, *53*, 1163–1174. [CrossRef]
4. Denton, C.P.; Khanna, D. Systemic sclerosis. *Lancet* **2017**, *390*, 1685–1699. [CrossRef]
5. Yen, E.Y.; Singh, D.R.; Singh, R.R. Trends in systemic sclerosis mortality over Forty-Eight years, 1968–2015: A US Population–Based study. *Arthritis Care Res.* **2021**, *73*, 1502–1510. [CrossRef]
6. Allanore, Y.; Simms, R.; Distler, O.; Trojanowska, M.; Pope, J.; Denton, C.P.; Varga, J. Systemic sclerosis. *Nat. Rev. Dis. Prim.* **2015**, *1*, 15002. [CrossRef] [PubMed]
7. Moore, S.C.; Hermes, E.R. Systemic sclerosis. *Treat. Complicat. Assoc. Syst. Scler.* **2008**, *65*, 315–321.
8. Godard, D. The needs of patients with systemic sclerosis—What are the difficulties encountered? *Autoimmun. Rev.* **2011**, *10*, 291–294. [CrossRef] [PubMed]
9. Cheng, H.; Yu, Z.; Yan, C.; Yang, H.; Gao, C.; Wen, H. Long-term efficacy and low adverse events of methylprednisolone pulses combined to low-dose glucocorticoids for systemic sclerosis: A retrospective clinical study of 10 years' follow-up. *J. Inflamm. Res.* **2022**, *15*, 4421–4433. [CrossRef] [PubMed]
10. Almeida, C.; Almeida, I.; Vasconcelos, C. Autoimmunity reviews. *Autoimmun. Rev.* **2015**, *14*, 1087–1096. [CrossRef] [PubMed]
11. Green, E.W.; Kahl, L.; Jou, J.H. Systemic sclerosis and the liver. *Clin. Liver Dis.* **2021**, *18*, 76–80. [CrossRef] [PubMed]
12. Zhong, L.; Pope, M.; Shen, Y.; Hernandez, J.J.; Wu, L. Prevalence and incidence of systemic sclerosis: A systematic review and meta-analysis. *Int. J. Rheum. Dis.* **2019**, *22*, 2096–2107. [CrossRef]
13. Englert, H.; Small-McMahon, J.; Davis, K.; O'Connor, H.J.; Chambers, P.; Brooks, P. Systemic sclerosis prevalence and mortality in Sydney 1974-88. *Aust. N. Z. J. Med.* **1999**, *29*, 42–50. [CrossRef] [PubMed]
14. Mayes, M.D.; Lacey, J.V.; Beebe-Dimmer, J.; Gillespie, B.W.; Cooper, B.; Brooks, P.; Laing, T.J.; Schottenfeld, D. Prevalence, incidence, survival, and disease characteristics of systemic sclerosis in a large US population. *Arthritis Rheum. Off. J. Am. Coll. Rheumatol.* **2003**, *48*, 2246–2255. [CrossRef] [PubMed]
15. Barnabe, C.; Joseph, L.; Belisle, P.; Labrecque, J.; Edworthy, S.; Barr, S.G.; Fritzler, M.; Fritzler, M.; Svenson, L.W.; Hemmelgarn, B.; et al. Prevalence of systemic lupus erythematosus and systemic sclerosis in the First Nations population of Alberta, Canada. *Arthritis Care Res.* **2012**, *64*, 138–143. [CrossRef]
16. Hoffmann-Vold, A.; Midtvedt, O.; Molberg, O.; Garen, T.; Gran, J.T. Prevalence of systemic sclerosis in south-east Norway. *Rheumatology* **2012**, *51*, 1600–1605. [CrossRef] [PubMed]
17. Gu, Y.S.; Kong, J.; Cheema, G.S.; Keen, C.L.; Wick, G.; Gershwin, M.E. The immunobiology of systemic sclerosis. *Semin. Arthritis Rheum.* **2015**, *38*, 132–160. [CrossRef]
18. Ngian, G.; Sahhar, J.; Proudman, S.M.; Stevens, W.; Wicks, I.P.; Van Doornum, S. Prevalence of coronary heart disease and cardiovascular risk factors in a national cross-sectional cohort study of systemic sclerosis. *Ann. Rheum. Dis.* **2012**, *71*, 1980–1983. [CrossRef] [PubMed]
19. Hughes, M.; Zanatta, E.; Sandler, R.D.; Avouac, J.; Allanore, Y. Improvement with time of vascular outcomes in systemic sclerosis: A systematic review and meta-analysis study. *Rheumatology* **2022**, *61*, 2755–2769. [CrossRef] [PubMed]
20. Ingegnoli, F.; Ughi, N.; Mihai, C. Update on the epidemiology, risk factors, and disease outcomes of systemic sclerosis. *Best Pract. Res. Clin. Rheumatol.* **2018**, *32*, 223–240. [CrossRef]
21. Marie, I. Systemic sclerosis and exposure to heavy metals. *Autoimmun. Rev.* **2019**, *18*, 62–72. [CrossRef] [PubMed]
22. Ota, Y.; Kuwana, M. Updates on genetics in systemic sclerosis. *Inflamm. Regen.* **2021**, *41*, 17. [CrossRef]
23. Varga, J.; Abraham, D. Systemic sclerosis: A prototypic multisystem fibrotic disorder Systemic sclerosis. *J. Clin. Investig.* **2007**, *117*, 557–567. [CrossRef] [PubMed]
24. Cowie, R.L. Silica-dust-exposed mine workers with scleroderma (systemic sclerosis). *Chest* **1987**, *92*, 260–262. [CrossRef]
25. Mora, G.F. High serum levels of silica nanoparticles in systemic sclerosis patients with occupational exposure: Possible pathogenetic role in disease phenotypes. *Semin. Arthritis Rheum.* **2009**, *48*, 475–481.
26. Ouchene, L.; Muntyanu, A.; Lavoue, J.; Baron, M.; Litvinov, I.V.; Netchiporouk, E. Toward Understanding of Environmental Risk Factors in Systemic Sclerosis. *J. Cutan. Med. Surg.* **2021**, *25*, 188–204. [CrossRef] [PubMed]
27. Andreussi, R.; Silva, L.; Carrico, H.; Luppino-Asad, A.P.; Andrade, D.C.; Sampaio-Barros, P.D. systemic sclerosis induced by the use of cocaine: Is there an association? *Rheumatol. Int.* **2019**, *39*, 387–393. [CrossRef]
28. Dolcino, M.; Pelosi, A.; Fiore, P.F.; Patuzzo, G.; Tinazzi, E.; Lunardi, C.; Puccetti, A. Gene Profiling in Patients with Systemic Sclerosis Reveals the Presence of Oncogenic Gene Signatures. *Front. Immunol.* **2018**, *9*, 449. [CrossRef] [PubMed]
29. Bertsch, C. CREST syndrome: A variant of systemic sclerosis *Orthop. Nurs.* **1995**, *14*, 53–60. [CrossRef]
30. Velayos, E.E.; Masi, A.T.; Stevens, M.B.; Shulman, L.E. The 'CREST' syndrome: Comparison with systemic sclerosis (scleroderma). *Arch. Intern. Med.* **1979**, *11*, 1240–1244. [CrossRef]
31. Rodnan, G.P.; Lipinski, E.; Luksick, J. Skin thickness and collagen content in progressive systemic sclerosis and localized scleroderma. *Arthritis Rheum. Off. J. Am. Coll. Rheumatol.* **1979**, *2*, 130–140. [CrossRef]

32. Bobeica, C.; Niculet, E.; Craescu, M.; Parapiru, E.; Musat, C.L.; Dinu, C.; Chiscop, I.; Nechita, L.; Debita, M.; Stefanescu, V. CREST Syndrome in Systemic Sclerosis Patients–Is Dystrophic Calcinosis a Key Element to a Positive Diagnosis? *J. Inflamm. Res.* **2022**, *15*, 3387–3394. [CrossRef]
33. Schoenfeld, S.R.; Castelino, F.V. Interstitial lung disease in scleroderma. *Rheum. Dis. Clin. N. Am.* **2015**, *41*, 237–248. [CrossRef]
34. Woodworth, T.G.; Suliman, Y.A.; Li, W.; Furst, D.E.; Clements, P. Scleroderma renal crisis and renal involvement in systemic sclerosis. *Nat. Rev. Nephrol.* **2016**, *12*, 678–691. [CrossRef]
35. Steen, V.D.; Medsger, T.A. Changes in causes of death in systemic sclerosis. *Ann. Rheum. Dis.* **2007**, *66*, 1972–2002. [CrossRef] [PubMed]
36. Steen, V.D.; Medsger, T.A. Severe organ involvement in systemic sclerosis with diffuse scleroderma. *Arthritis Rheum. Off. J. Am. Coll. Rheumatol.* **2000**, *43*, 2437–2444. [CrossRef]
37. Al-Dhaher, F.F.; Pope, J.E.; Ouimet, J.M. Determinants of Morbidity and Mortality of Systemic Sclerosis in Canada. *Semin. Arthritis Rheum.* **2010**, *39*, 269–277. [CrossRef] [PubMed]
38. Bussone, G.; Mouthon, L. Interstitial lung disease in systemic sclerosis. *Autoimmun. Rev.* **2011**, *10*, 248–255. [CrossRef]
39. Goh, N.S.L.; Desai, S.R.; Veeraraghavan, S.; Hansell, D.M.; Copley, S.J.; Maher, T.M.; Corte, T.J.; Sander, C.R.; Ratoff, J.; Devaraj, A. Interstitial lung disease in systemic sclerosis. *Am. J. Respir. Crit. Care Med.* **2008**, *177*, 1248–1254. [CrossRef]
40. Lynch, D.A.; Godwin, J.D.; Safrin, S.; Starko, K.M.; Hormel, P.; Brown, K.K.; Raghu, G.; King, T.E.; Bradford, W.Z.; Schwartz, D.A. High-resolution computed tomography in idiopathic pulmonary fibrosis: Diagnosis and prognosis. *Am. J. Respir. Crit. Care Med.* **2005**, *172*, 488–493. [CrossRef]
41. Hoffmann-Vold, A.; Maher, T.M.; Philpot, E.E.; Ashrafzadeh, A.; Barake, R.; Barsotti, S.; Bruni, C.; Carducci, P.; Carreira, P.E.; Castellvi, I. The identification and management of interstitial lung disease in systemic sclerosis: Evidence-based European consensus statement. *Lancet Rheumatol.* **2020**, *2*, 71–83. [CrossRef]
42. Giacomelli, R.; Liakouli, V.; Berardicurti, O.; Ruscitti, P.; Di Benedetto, P.; Carubbi, F.; Guggino, G.; Di Bartolomeo, S.; Ciccia, F.; Triolo, G. Interstitial lung disease in systemic sclerosis: Current and future treatment. *Lancet Rheumatol.* **2017**, *37*, 853–863. [CrossRef] [PubMed]
43. Luo, Y.; Wang, Y.; Wang, Q.; Xiao, R.; Lu, Q. Systemic sclerosis: Genetics and epigenetics. *J. Autoimmun.* **2013**, *41*, 161–167. [CrossRef] [PubMed]
44. Romano, E.; Manetti, M.; Guiducci, S.; Ceccarelli, C.; Allanore, Y.; Matucci-Cerinic, M. The genetics of systemic sclerosis: An update. *Clin. Exp.-Rheumatol.-Incl Suppl.* **2011**, *29*, S75.
45. Murdaca, G.; Contatore, M.; Gulli, R.; Mandich, P.; Puppo, F. Genetic factors and systemic sclerosis. *Autoimmun. Rev.* **2016**, *15*, 427–432. [CrossRef] [PubMed]
46. Jamian, L.; Wheless, L.; Crofford, L.J.; Barnado, A. Rule-based and machine learning algorithms identify patients with systemic sclerosis accurately in the electronic health record. *Arthritis Res. Ther.* **2019**, *21*, 305. [CrossRef]
47. Akay, M.; Du, Y.; Sershen, C.L.; Wu, M.; Chen, T.Y.; Assassi, S.; Mohan, C.; Akay, Y.M. Deep learning classification of systemic sclerosis skin using the MobileNetV2 model. *IEEE Open J. Eng. Med. Biol.* **2021**, *2*, 104–110. [CrossRef] [PubMed]
48. Assassi, S.; Volkmann, E.R.; Zheng, W.J.; Wang, X.; Wilhalme, H.; Lyons, M.A.; Roth, M.D.; Tashkin, D.P. Peripheral blood gene expression profiling shows predictive significance for response to mycophenolate in systemic sclerosis-related interstitial lung disease. *Ann. Rheum. Dis.* **2022**, *81*, 854–860. [CrossRef] [PubMed]
49. Sen, P.C.; Hajra, M.; Ghosh, M. *Emerging Technology in Modelling and Graphics*; Springer: Singapore, 2020.
50. Li, L.; Liu, X.; Yang, F.; Xu, W.; Wang, J.; Shu, R. A review of artificial neural network based chemometrics applied in laser-induced breakdown spectroscopy analysis. *Spectrochim. Acta Part B At. Spectrosc.* **2021**, *180*, 106183. [CrossRef]
51. Jawad, J.; Hawari, A.H.; Zaidi, S.J. Artificial neural network modeling of wastewater treatment and desalination using membrane processes: A review. *Chem. Eng. J.* **2021**, *419*, 129540. [CrossRef]
52. Jena, P.R.; Majhi, R.; Kalli, R.; Managi, S.; Majhi, B. Impact of COVID-19 on GDP of major economies: Application of the artificial neural network forecaster. *Econ. Anal. Policy* **2021**, *69*, 324–339. [CrossRef]
53. Norimatsu, Y.; Yoshizaki, A.; Kabeya, Y.; Fukasawa, T.; Omatsu, J.; Fukayama, M.; Kuzumi, A.; Ebata, S.; Yoshizaki-Ogawa, A.; Asano, Y.; et al. Expert-Level Distinction of Systemic Sclerosis from Hand Photographs Using Deep Convolutional Neural Networks. *J. Investig. Dermatol.* **2021**, *141*, 2536–2539. [CrossRef] [PubMed]
54. Chassagnon, G.; Vakalopoulou, M.; Regent, A.; Zacharaki, E.I.; Aviram, G.; Martin, C.; Marini, R.; Bus, N.; Jerjir, N.; Mekinian, A. Deep learning–based approach for automated assessment of interstitial lung disease in systemic sclerosis on CT images. *Radiol. Artif. Intell.* **2020**, *2*, e190006. [CrossRef] [PubMed]
55. Chandrasekaran, A.C.; Fu, Z.; Kraniski, R.; Wilson, F.P.; Teaw, S.; Cheng, M.; Wang, A.; Ren, S.; Omar, I.M.; Hinchcliff, M.E. Computer vision applied to dual-energy computed tomography images for precise calcinosis cutis quantification in patients with systemic sclerosis. *Arthritis Res. Ther.* **2021**, *23*, 6. [CrossRef] [PubMed]
56. Karsoliya, S. Approximating number of hidden layer neurons in multiple hidden layer BPNN architecture. *Int. J. Eng. Trends Technol.* **2012**, *3*, 714–717.
57. Deng, Y.; Zhou, X.; Shen, J.; Xiao, G.; Hong, H.; Lin, H.; Wu, F.; Liao, B. New methods based on back propagation (BP) and radial basis function (RBF) artificial neural networks (ANNs) for predicting the occurrence of haloketones in tap water. *Sci. Total Environ.* **2021**, *772*, 145534. [CrossRef]

58. Rahman, A.; Chandren, M.R.; Albashish, D.; Rahman, M.; Usman, O.L. Artificial neural network with Taguchi method for robust classification model to improve classification accuracy of breast cancer. *PeerJ Comput. Sci.* **2021**, *7*, e344. [CrossRef]
59. Cervantes, J.; Garcia-Lamont, F.; Rodríguez-Mazahua, L.; Lopez, A. A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing* **2020**, *408*, 189–215. [CrossRef]
60. Pisner, D.A.; Schnyer, D.M. Support vector machine. In *Machine Learning*; Elsevier: Amsterdam, The Netherlands, 2020.
61. Milanese, G.; Mannil, M.; Martini, K.; Maurer, B.; Alkadhi, H.; Frauenfelder, T. Quantitative CT texture analysis for diagnosing systemic sclerosis: Effect of iterative reconstructions and radiation doses. *Medicine* **2019**, *98*, e16423. [CrossRef]
62. Filippini, C.; Cardone, D.; Perpetuini, D.; Chiarelli, A.M.; Gualdi, G.; Amerio, P.; Merla, A. Convolutional neural networks for differential diagnosis of raynaud's phenomenon based on hands thermal patterns. *Appl. Sci.* **2021**, *11*, 3614. [CrossRef]
63. Nitkunanantharajah, S.; Haedicke, K.; Moore, T.B.; Manning, J.B.; Dinsdale, G.; Berks, M.; Taylor, C.; Dickinson, M.R.; Justel, D.; Ntziachristos, V. Three-dimensional optoacoustic imaging of nailfold capillaries in systemic sclerosis and its potential for disease differentiation using deep learning. *Sci. Rep.* **2020**, *10*, 16444. [CrossRef] [PubMed]

3.4 Paper IV

Gene Identification in Inflammatory Bowel Disease via a Machine Learning Approach

Authors: Gerardo Alfonso Perez, Raquel Castillo

Medicina. 2023, 59(1218).

doi.org/10.3390/medicina59071218

medicina

MDPI

Article

Gene Identification in Inflammatory Bowel Disease via a Machine Learning Approach

Gerardo Alfonso Perez * **and Raquel Castillo**

Biocomp Group, Institute of Advanced Materials (INAM), Universitat Jaume I, 12071 Castello, Spain
* Correspondence: ga284@cantab.net

Abstract: Inflammatory bowel disease (IBD) is an illness with increasing prevalence, particularly in emerging countries, which can have a substantial impact on the quality of life of the patient. The illness is rather heterogeneous with different evolution among patients. A machine learning approach is followed in this paper to identify potential genes that are related to IBD. This is done by following a Monte Carlo simulation approach. In total, 23 different machine learning techniques were tested (in addition to a base level obtained using artificial neural networks). The best model identified 74 genes selected by the algorithm as being potentially involved in IBD. IBD seems to be a polygenic illness, in which environmental factors might play an important role. Following a machine learning approach, it was possible to obtain a classification accuracy of 84.2% differentiating between patients with IBD and control cases in a large cohort of 2490 total cases. The sensitivity and specificity of the model were 82.6% and 84.4%, respectively. It was also possible to distinguish between the two main types of IBD: (1) Crohn's disease and (2) ulcerative colitis.

Keywords: inflammatory bowel disease; Crohn's disease; ulcerative colitis

check for updates

Citation: Alfonso Perez, G.; Castillo, R. Gene Identification in Inflammatory Bowel Disease via a Machine Learning Approach. *Medicina* **2023**, *59*, 1218. https://doi.org/10.3390/medicina59071218

Academic Editors: Wisit Cheungpasitporn, Charat Thongprayoon, Pattharawin Pattharanitima and Wisit Kaewput

Received: 1 June 2023
Revised: 24 June 2023
Accepted: 26 June 2023
Published: 28 June 2023

1. Introduction

In this paper, the genetic expression signature of inflammatory bowel disease is analyzed using machine learning techniques. Inflammatory bowel disease (IBD) is a chronic [1] inflammatory disease, whose cause remains unclear. Patients can show an array of different symptoms. According to the Mayo Clinic, some of the most common symptoms associated with inflammatory bowel disease include pain, diarrhea, fatigue, cramps, blood present in stools and weight loss. Extraintestinal symptoms appear in approximately 24% of patients [2]. Patients can also have very different evolution and responses to treatments.

Another interesting characteristic of this illness, so far without a good explanation, is that it tends to have a higher incidence and prevalence in urban areas [3] compared to rural areas, perhaps suggesting a link to lifestyles. The incidence of IBD has been increasing [4]. Inflammatory bowel disease is becoming an increasingly important health problem [5]. Developing and newly industrialized countries are seeing a particularly rapid increase in the incidence of the illness [6]. The reasons behind this increase remain unclear. It might be related to changes in dietary habits or exposure to pollutants, but there are currently, to the best of our knowledge, no definitive data to prove it. It is also likely that the illness is being detected earlier in those countries as their healthcare infrastructure develops. Nevertheless, environmental factors appear to play a role in the illness. IBD increases the chances of developing other illnesses, such as colorectal cancer [7] and osteoporosis [8]. More than 7% of patients with IBD develop osteoporosis [8]. Additionally, IBD can have a very significant impact on the quality of life of the patient and can make normal activities, such as working, challenging in some severe cases.

One of the main theories of the cause of IBD is that it is an abnormal immune response in genetically predisposed individuals, triggered by some external factor such as a virus or bacteria [9,10]. Cytokines appear to play an important role in IBD [11]. Lifestyle factors, such as stress, smoking and diet [12], have also been identified in the literature as having a role in the illness [13]. The illness results in a defective regulation of the mucosa [14]. Tamboli et al. [15] specifically mentioned intestinal bacteria as a major factor in the initial stages of the disease. Chang [16] concluded that the two causative agents are (1) abnormal immune response in the gastrointestinal mucosa and (2) alterations in the gut microbiome [17]. The two major forms of IBD are ulcerative colitis (UC) and Crohn's disease (CD) [18]. A visual representation of UC and CD is shown in Figure 1.

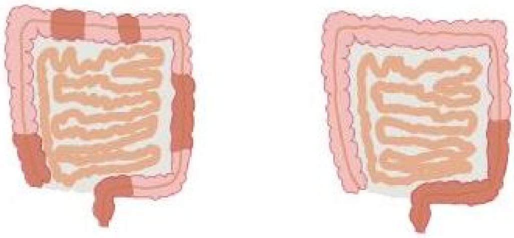


Figure 1. Visual representation of Crohn's disease (**left**) and ulcerative colitis (**right**). It can be seen some of the usual areas involved in UC and CD. It should be noted that there is substantial variation among patients.

IBD appears to have a genetic component. Loddo and Romano [19] mentioned that approximately 15% of the patients with Crohn's disease have a family member with the same condition. They also mentioned a 50% concordance in monozygotic twins. Bernard and Ramnik [20] concluded that genes help regulate the complex interaction between microbial and environmental factors. Another indications of a genetic component in the disease is that some ethnic groups, such as Ashkenazim, have higher incidence and prevalence [21]. Some authors, such as McGovern et al. [22], highlighted the issue that a large amount of the existing literature focuses on individuals of European ancestry. This is especially important in an illness such as IBD, in which ethnicity seems to play an important role not only in terms of prevalence but also in terms of early onset, reaction to the treatment and severity of the illness. A schematic representation of the interaction between genetic predisposition and environmental factors is shown in Figure 2. The underlying mechanics of this interaction between genetic predisposition and environmental factors remain not well understood.

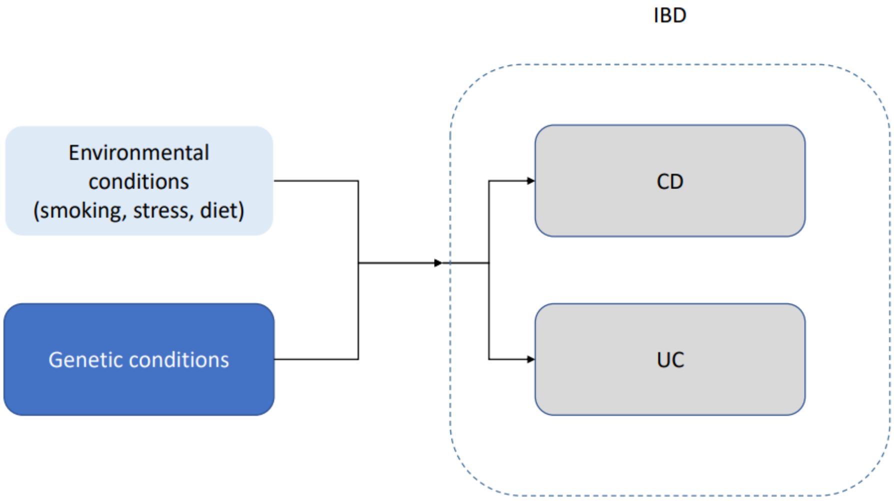


Figure 2. Schematic representation of the interaction between genetic predisposition and environmental factors in ulcerative colitis (UC) and Crohn's disease (CD). IBD, in both of its main forms, is likely caused by a combination of underlying genetic conditions and environmental conditions.

There have been many developments in the genetics of IBD, but despite the identification of some genes, the underlying process remains not well understood. The evidence points to a process in which multiple genes are involved (polygenic) [23,24]. Cho and Abraham [25] cited the well-known Nod2 (CARD15) polymorphism association with Crohn's disease. This gene is located in chromosome 16 and has been mentioned by multiple authors [26]. Katuka et al. [27] mentioned that in Japan, the NUDT15 polymorphism is routinely tested before administering thiopurine to inflammatory bowel disease patients. Mathew and Lewis [28] studied genes in chromosome 5q31n 6p21 and 19p. Achkar and Duerr [29] identified IL23R and ATG16L1 as being involved in CD. These two genes are frequently mentioned in the existing literature [30]. Stoll et al. [31] identified DLG5, while Cleynen et al. [32] identified 163 susceptibility loci for IBD. Ahmad et al. mentioned that CD and UC are related diseases that share some but not all the susceptibility genes [33]. Inflammatory bowel disease is a chronic disease that typically requires lifelong medication [34]. Given the heterogeneity in the illness, it is not surprising that there are multiple treatment options with different levels of expected success.

Machine learning techniques are increasingly popular in medicine with applications in many different types of illness [35–37]. There has been some interesting research applying machine learning techniques in the context of inflammatory bowel disease [38–40]. This has been in part due to the large amount of data generated experimentally [41] and the need to come up with appropriate techniques to analyze such a large quantity of data. For instance, Wei et al. [42] used GWAS data to carry out a risk assessment of patients with ulcerative colitis or Crohn's disease. Isakov et al. [43] identified 67 genes using machine learning techniques related to IBD. Coelho et al. [44] also used machine learning techniques, but their analysis covers pediatric patients, who have some characteristics different from the usual adult case. The same group of authors published another interesting paper [38] using three different machine learning techniques and endoscopic data, achieving an accuracy of 71.0%, 76.9% and 82.7% respectively. The work of Smolander et al. [45] is another interesting paper analyzing gene expression, using machine learning techniques in the context of complex disorders. Some authors, such as Stankvic et al. [46], mentioned that despite an increase in the use of machine learning techniques in IBD, the understanding of the illness remains incomplete.

One of the main objectives of this article is trying to identify genes that are relevant in the context of inflammatory bowel disease using machine learning techniques. The genes are chosen by selecting those genes with a gene expression level that is empirically useful to distinguish between control individuals and patients with IBD. The details of this process will be explained in the next section, but it is based on using different machine learning techniques (classification purposes) in combination with Monte Carlo simulations for the selection of genes. Another objective of this article is to be able to identity appropriate genes differentiating between Crohn's disease and ulcerative colitis using a similar approach than when distinguishing between healthy and IBD patients.

2. Materials and Methods

The dataset was retrieved from the Gene Expression Omnibus. The identification number is GSE 193677 [47]. The data include 2490 total cases. Of these 2490 cases, 461 are controls cases, while 2029 are individuals with adult inflammatory bowel disease (IBD). Of those 2029, a slight majority of 1157 have Crohn's disease while 872 have ulcerative colitis. The average age of the patient is 44.9 years, with a range from 19 to 82 years old. A histogram showing the age distribution is shown in Figure 3. There are 1174 female and 1316 male cases. Tissue biopsies were obtained in the right colon, left colon, transverse, rectum, Ileum, sigmoid and cecum. The number of cases for each of this regions is summarized below in Table 1. The data consist of gene expression profiling by high throughput sequencing obtained using the Illumina HiSeq 2500 (Illumina, Inc. San Diego, CA, USA). There are 56,632 expression profiling data per patient.

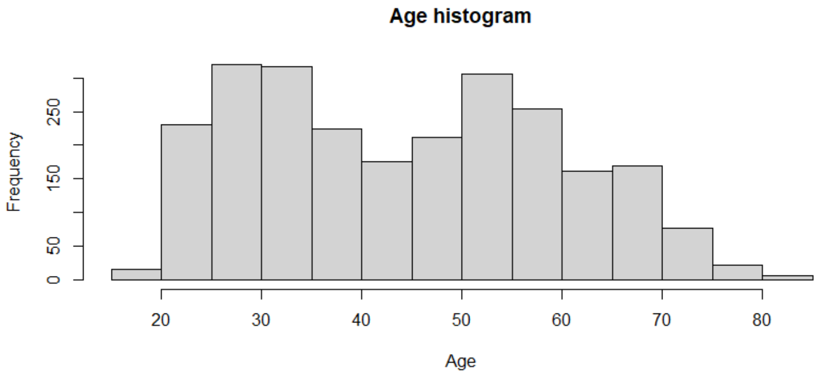


Figure 3. Histogram describing the age of the patients. The range is from 19 to 82 years old.

Table 1. Biopsies (tissue areas).

Area	Cases
Rectum	904
Left colon	180
Right colon	252
Ileum	672
Transverse	90
Sigmoid	163
Cecum	229

The data were divided into two subgroups, a training dataset and a testing dataset. Ψ_{Tr} denotes the training dataset and Ψ_{Ts} the testing dataset. The training and testing datasets contain approximately 80% and 20% of all the cases, respectively. Each column represents a patient. The division into a training and a testing dataset was carried out in a randomized way to try to avoid introducing biases in the analysis. The first row in each dataset contains a numerical classifier identifying the subject as a control or patient (UD or CD) as shown in Equation (1):

$$\forall j \in [1,n],\ \Phi_j = \{Control = 0, UC = CD = 1\} \tag{1}$$

with n being the total number of cases. An example, for clarity purposes, can be seen in Equation (2):

$$\Phi = \{\Phi_1, \Phi_2, ..., \Phi_n\} = \{1, 0, ..., 1\} \tag{2}$$

The following two rows contain the age (a), see Equations (3) and (4), and the gender (S), see Equations (5) and (6), of each individual, respectively:

$$\forall j \in [1,n], a_j = \{x_j\}\ x \in \mathbb{R} \tag{3}$$

$$a = \{age\} = \{a_1, a_2, ..., a_n\} = \{47, 52, ..., 61\} \tag{4}$$

$$\forall j \in [1,n], S_j = \{Female = 0, Male = 1\} \tag{5}$$

$$S = \{gender\} = \{S_1, S_2, ..., S_m\} = \{0, 1, ..., 1\} \tag{6}$$

In a similar way, the following row contains the region for the biopsy. All the other rows contain gene expression data (see Equations (7) and (8)):

$$\forall j \in [1,n], \forall k \in [1,m], g_{kj} = \{Z_{kj}\}\ Z \in \mathbb{R} \tag{7}$$

$$G_k = \{g_{kj}\} = \{18, 241, ..., 132\} \tag{8}$$

where k is the index for each row. An example, for visualization purposes, of the data can be seen in Equation (9):

$$\Psi_{Tr} = \begin{pmatrix} 0 & 1 & 2 & 0 & \cdots \\ 60 & 45 & 35 & 55 & \cdots \\ 0 & 0 & 1 & 1 & \cdots \\ 0 & 0 & 0 & 1 & \cdots \\ 80 & 30 & 55 & 40 & \cdots \\ \vdots & \vdots & & \vdots & \end{pmatrix} \tag{9}$$

As a first step, the correlation $C_0(c,d)$ between the categorical data representing the classification group (control or IBD) and each row is calculated (Equations (10)):

$$\forall\, k \in [1,m], C_0 = C_0(\Phi, G_k) \tag{10}$$

Therefore, C_0 is a vector with m components. From this mapping, the highest $q\%$ $(0 \leq q \leq 100)$ is selected among these m values. Hence, there is a reduction in the dimension of the vector (Equation (11)):

$$C_0(dim = m) \Rightarrow C_0^*(dim = m < k) \tag{11}$$

This step is performed in an attempt to include the factors that are potentially able to generate an accurate model while filtering out potential noise (not all genes are involved in inflammatory bowel disease). In other words, it is an attempt to filter out noise from genes than have no biological impact on the disease but that can lead the model to find spurious relationships given the large amount of data. The above-mentioned step is carried out only with the training dataset (containing approximately 80% of the cases). After this step, when the genes have already been selected, then all the other genes will be excluded from both the training and the testing dataset. In this way, it is possible to carry out a filtering of the initial gene list. A selection of 23 machine learning techniques was selected; see Table 2. Ten times cross validation was carried out (training dataset).

Table 2. Selected machine learning algorithms.

Algorithm	Algorithm
Complex Tree	Fine KNN
Medium Tree	Medium KNN
Simple Tree	Coarse KNN
Linear Discriminant	Cosine KNN
Quadratic Discriminant	Cubic KNN
Logistic Regression	Weighted KNN
Linear SVM	Boosted Trees
Quadratic SVM	Bagged Trees
Cubic SVM	Subspace Discriminant
Fine Gaussian SVM	Subsspace KNN
Medium Gaussian SVM	RUSBossted Trees
Coarse Gaussian SVM	

The artificial neural network (ANN) is a well-known machine learning algorithm. Given its versatility and wide use, this technique is used to determine a baseline classification accuracy, against which the other techniques are compared. In the ANN approach, it is necessary to carry out hyperparameter optimization. One of the key parameters to optimize is the number of layers in the ANN. This is achieved by carrying out simulations from 1 to 1000 layers and the related accuracy estimated. Unless explicitly mentioned, the accuracy (and other measures of the goodness of the fit) is that of the testing dataset (not used during the training phase). In this way, for each configuration γ ($\gamma = \{1, ..., 1000\}$), an accuracy A_{nn} measure is estimated (A^{γ}_{nn}). Then, the best model $(\bar{A}_{nn})$ is selected as

$$\bar{A}_{nn}(\gamma) = sup(A^{\gamma}_{c}) \tag{12}$$

This is the baseline model. For each machine learning techniques, the model is trained with the training dataset, and then an accuracy estimate is obtained, and the best model $\bar{A}(\lambda)$ is selected (Equation (13)). The training and model selection (gene selection) is entirely performed with the training dataset. After the model is selected (including the genes), the accuracy and other metrics are expressed in terms of the testing dataset (not used for training or model selection):

$$\bar{A}(\lambda) = sup(A^{\lambda}) \tag{13}$$

Then this is compared to the base level, selecting the final best model $\bar{A}_{max}$ as follows:

$$A_{max} = max\{\bar{A}_{nn}(\gamma), \bar{A}(\lambda)\} \tag{14}$$

This analysis is initially carried out for all the gene expression data available after selecting the top $q = 1\%$. In this case, the initial number of gene expression data per patient entails 566 rows of information. Then a Monte Carlo approach is followed, in which the number of rows is randomly reduced in each iteration by a random number β. This random number β is changed in each iteration and is strictly less than the total number of rows in the previous iteration. An example is summarized in Table 3. The rationale behind using a Mote Carlo simulation approach is that it is not feasible to estimate all the possible combinations of 566 genes, and hence some type of combinatorial approach needs to be used. This is a frequent situation in polygenic illness, such as IBD, in which a potentially large number of genes might be involved in the disease.

Table 3. Example of iterative algorithm testing different configurations of gene expressions.

Iteration	Initial N. Genes	β
0	566	30
1	536	125
2	411	58
3	353	9
4	344	215
⋮	⋮	⋮

This process is repeated p times (p = 100), and the ten most accurate models are selected.

In the second section, a similar approach is followed but the mapping shown in Equation (1) has to be changed, as the objective is now to distinguish between ulcerative colitis and Crohn's disease cases (the two major types of IBD). The mapping in this case is as follows (Equation (15)):

$$\forall j \in [1, u],\ \Phi_j = \{UC = 0, CD = 1\} \tag{15}$$

An alternative approach to the one presented is using a linear approach, such as, for instance, lasso regression [48,49]. Lasso regression offers the advantage that it makes some of the coefficients equal to zero, in practice reducing the number of inputs to the model. Using lasso regression, it is possible to reduce the number of genes selected for the classification model. In fact, lasso has become a frequently used feature selection algorithm [50,51].

3. Results

As previously described, the first step involves estimating a base level for the accuracy using artificial neural networks with simulations using 1 to 100 hidden layers. Each layer consists of 30 neurons. As it can be seen in Figure 4, increasing the number of layers does not necessarily translate into higher accuracy. The highest accuracy (testing dataset) obtained is 80.35% with a configuration including 920 hidden layers. The only other simulation reaching an accuracy above 80.00% is an ANN with 330 layers, reaching 80.10%. All the other simulations achieve a mean accuracy below 80.00%. No model has an accuracy below 70%. These results are obtained for a configuration of 74 rows (gene expression) which, as will be shown later, is the configuration that obtains the highest accuracy for the machine learning algorithm tested. As previously mentioned, the reported accuracy is the accuracy of the testing dataset, which is not used during the training phase.

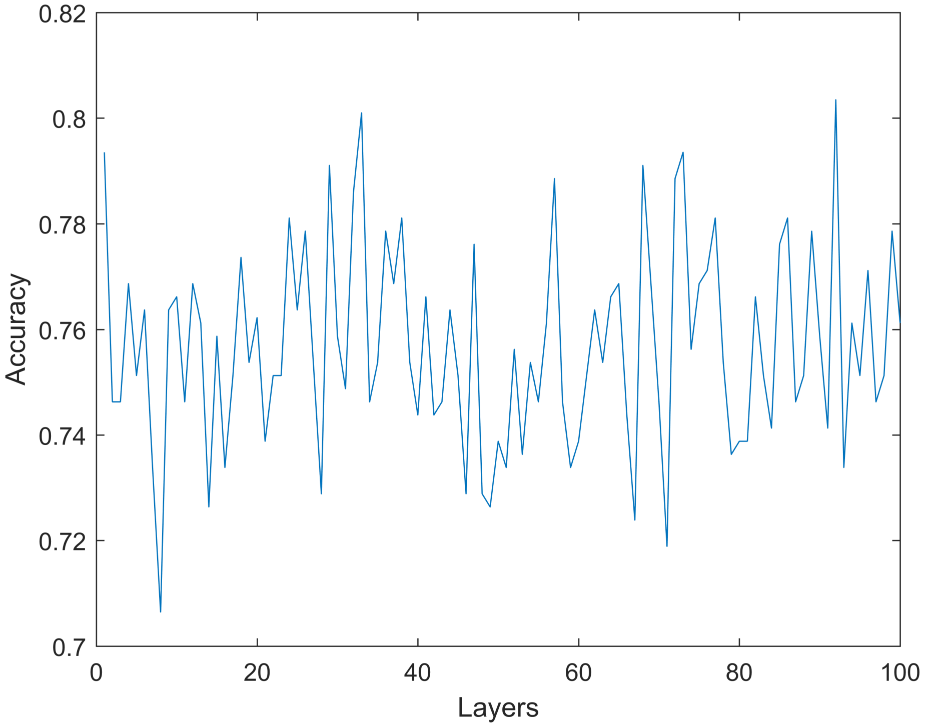


Figure 4. Accuracy of the neural network model for a range of number of artificial neurons. No model has an accuracy below 70% or higher than 80.35%.

Different machine learning algorithms are tested (as described in the Materials and Methods section). As an example, in Table 4, the accuracy results for one of the simulations are shown (140 gene expressions). In this specific case, the highest accuracy obtained is 81.5%. This accuracy is obtained by five different algorithms (Linear SVM, Fine Gaussian SVM, Medium Gaussian SVM, Coarse Gaussian SVM and Coarse KNN).

The results from the 10 most accurate simulations can be seen in Table 5. Of the ten most accurate results, nine use the bagged trees algorithm. The only other algorithm in the top ten most accurate models is the Subspace KNN. The highest accuracy is obtained for a model with 74 gene expression data, obtaining an accuracy, sensitivity and specificity of 84.2%, 82.6% and 84.4%, respectively. The list with these 74 genes can be found in Table 6.

Table 4. As an example, in this table, sample training with all 23 algorithms is shown. In this case, the model uses 140 gene expression data and the highest accuracy is 81.5%. This accuracy is actually reached by several algorithms (Linear SVM, Fine Gaussian SVM, Medium Gaussian SVM, Coarse Gaussian SVM and Coarse KNN).

Algorithm	Accuracy
Complex Tree	0.701
Medium Tree	0.783
Simple Tree	0.804
Linear Discriminant	0.645
Quadratic Discriminant	0.711
Logistic Regression	0.807
Linear SVM	0.815
Quadratic SVM	0.788
Cubic SVM	0.756
Fine Gaussian SVM	0.815
Medium Gaussian SVM	0.815
Coarse Gaussian SVM	0.815
Fine KNN	0.719
Medium KNN	0.770
Coarse KNN	0.815
Cosine KNN	0.768
Cubic KNN	0.764
Weighted KNN	0.773
Boosted Trees	0.805
Bagged Trees	0.804
Subspace Discriminant	0.812
Subsspace KNN	0.748
RUSBossted Trees	0.606

Table 5. Top ten models obtained according to the accuracy metric.

N. Genes	Algorithm	Accuracy	Sensitivity	Specificity
74	Bagged Trees	0.842	0.826	0.844
38	Subspace KNN	0.842	0.755	0.859
18	Bagged Trees	0.839	0.787	0.847
139	Bagged Trees	0.836	0.755	0.850
220	Bagged Trees	0.834	0.758	0.847
266	Bagged Trees	0.833	0.740	0.850
26	Bagged Trees	0.833	0.821	0.834
16	Bagged Trees	0.833	0.879	0.828
17	Bagged Trees	0.831	0.738	0.848
104	Bagged Trees	0.830	0.750	0.843

The results, when differentiating UC and CD cases, are not as accurate as when differentiating between control cases and IBD cases. This is in line with the expectations, as we are differentiating between two types of the same illness. These results are shown in Table 7. The most accurate result is obtained when using 562 gene expression data and the bagged trees algorithm. The accuracy, sensitivity and specificity are 73.4%, 79.0% and 71.2%, respectively. The list with these 562 genes can be found in the Supplementary Material.

Table 6. List of 74 genes selected by the algorithm.

B2M	RPS3	CHP1	SLC35A3
MALAT1	MAN2B1	ETNK1	PDIA3
EEF1A1	NDRG1	SLC1A2	DDX3X
MUC2	AHCYL2	GHITM	WDR1
FABP6	RPS14	MGAT4A	KLF5
KRT20	MYO1D	CLDN7	TSC22D1
CA1	A2M	COPZ2	RPL35A
FLNB	ADH1C	APOC3	SCP2
PHGR1	DDX17	SAT1	MATR3
IGKV1-5	FOS	ACE	CD46
CKB	RPL7	CD2AP	HNRNPH1
FABP1	SLC44A1	PAPSS2	PRKDC
FABP2	FN1	PDCD4	RPL37
CLDN4	RPL18	HPGD	LUM
TSPAN3	TDP2	UGT2A3	HSPA9
CDHR2	RPS12	UQCRC1	KIAA1109
CLTC	SPINT2	ST6GALNAC6	MIM24
COL1A2	RPL10A	ARF1	
ENO1	NCOA4	PRKACB	

Table 7. Top ten models obtained according to the accuracy metric distinguishing UC and CD patients.

N. Genes	Algorithm	Accuracy	Sensitivity	Specificity
562	Bagged Trees	0.734	0.790	0.712
66	Bagged Trees	0.728	0.679	0.767
24	Bagged Trees	0.718	0.665	0.742
37	Bagged Trees	0.718	0.821	0.687
564	Bagged Trees	0.712	0.909	0.671
132	Bagged Trees	0.704	0.929	0.676
49	Bagged Trees	0.704	0.679	0.719
15	Bagged Trees	0.700	0.713	0.697
550	Bagged Trees	0.694	0.871	0.659
277	Bagged Trees	0.673	0.616	0.717

As previously mentioned, an alternative approach to the one proposed is using lasso regression as a tool for the selection of inputs. The lasso approach selects 470 genes with the goodness-of-fit metric shown in Table 8. The accuracy and specificity results obtained in this approach are similar to those obtained in the proposed approach in the previous section. However, the sensitivity results from the lasso approach seem to be lower.

Table 8. Top ten models obtained using the lasso approach (470 genes) according to the accuracy metric distinguishing between control and UC and CD patients.

Algorithm	Accuracy	Sensitivity	Specificity
Medium KNN	0.817	0.667	0.817
Bagged Trees	0.817	0.667	0.817
Weighted KNN	0.815	0.500	0.816
Cubic KNN	0.807	0.143	0.815
Simple Tree	0.804	0.231	0.816
Subspace Dis.	0.804	0.405	0.829
Linear Dis.	0.802	0.433	0.842
Cosine KNN	0.802	0.300	0.819
Medium Tree	0.797	0.200	0.817
Subspace KNN	0.786	0.313	0.826

The lasso approach is also used to distinguish between UC and CD patients. In this case, the lasso approach selects 430 genes. The table with the goodness-of-fit results in this approach is shown below (Table 9). The results using the lasso approach to distinguish between UC and DC patients are not as accurate as in the previous section. In both cases, using lasso or the proposed approach, differentiating between UC and DC patients appears to be more challenging than differentiating between control health individuals and patients with UC/CD. The lasso approach does not appear to increase the goodness of fit of the classification forecasts compared to the approached followed in the previous section.

Table 9. Top ten models obtained using the lasso approach (430 genes) according to the accuracy metric distinguishing between UC and CD patients.

Algorithm	Accuracy	Sensitivity	Specificity
Subspace Dis.	0.584	0.611	0.523
Logistic Reg.	0.572	0.617	0.503
Medium KNN	0.568	0.580	0.493
Cubic KNN	0.562	0.578	0.474
Weighted KNN	0.560	0.584	0.478
Simple Tree	0.558	0.569	0.412
Bagged Trees	0.558	0.583	0.474
Boosted Trees	0.556	0.580	0.467
Cosine KNN	0.550	0.574	0.448
Fine KNN	0.538	0.595	0.463

4. Discussion

Machine learning techniques are used to identify a set of 74 genes, which can be used, with an average accuracy of 84.2%, to distinguish between control (healthy individuals) and patients with inflammatory bowel disease. The specificity and sensitivity of this model are also relatively high at 82.6% and 84.4%, respectively. The selection of these 74 genes is carried out following a Monte Carlo simulation approach. Given that some of the symptoms of inflammatory bowel disease are common in other illnesses, it might be interesting to have another objective diagnostic tool. It is also interesting to observe that among multiple machine learning techniques used in the cohort of patients analyzed, the bagged trees approach seems to consistently achieve a high level of accuracy, particularly when compared to other, arguably more sophisticated machine learning techniques, such as artificial neural networks. The analysis controls for age, gender and region of the biopsy. The proportion of female and male cases is balanced, with 1174 female patients and 1316 male patients. The average age in the cohort is 44.9 years, covering a wide age range (from 19 to 82 years old). The results of the artificial neural networks include an optimization of the hyperparameters with simulations ranging from 1 to 1000 hidden layers. It is also observed that simply increasing the number of layers in an artificial neural network does not necessarily translate into better accuracy. It is also possible to distinguish between the two main types of IBD—Crohn's disease and ulcerative colitis—but in this case with a lower level of accuracy. The accuracy, using this approach is 73.4%. The accuracy, sensitivity and specificity reported are those of the testing dataset. As normal practice, the data are divided into training and testing datasets in an attempt to increase the reproducibility of the analysis. Approximately 20% of the total cases are included in the testing dataset. The relatively large number of genes obtained in the bets model is in line with the prevalent view in the existing literature that the illness is polygenic.

There is a high degree of heterogeneity in inflammatory bowel disease, leading to varied severity and evolution of the illness. The existing literature, see, for instance, Yamamot et al. [52] or Ahmad et al. [33], points towards a polygenic illness with a complex interaction with environmental factors. Our results are consistent with this polygenic description. In this context, it is important to generate algorithms that are able to differentiate among control and patients as well as between different types of inflammatory bowel disease, namely Crohn's disease and ulcerative colitis. A promising area of future research is to apply this type of approach in order to target treatments in a more personalized way. It seems reasonable that there could be genetic differences among patients that can have a substantial impact on the outcome of the suggested treatments. This is particularly important in the context of inflammatory bowel disease, given the heterogeneity of the responses to treatments by different patients.

Some of the genes identified by the proposed algorithm are cited in the existing literature on intestinal-related illnesses. B2M was mentioned by Krzystek-Korpacka et al. [53] in the context of bowel inflammation. There are other papers, such as that of Bednarz-Misa et al. [54], discussing B2M in the context of bowel inflammation and cancer. Another gene identified by the algorithm is MALAT1, which is also mentioned in the existing literature. Li et al. [55] suggested that MALAT1 maintains intestinal mucosal homeostasis in Crohn's disease. The authors concluded that the downregulation of MALAT1 contributes to the pathogenesis of CD. EEF1A1 was identified in a dog study as being involved in inflammatory bowel disease and cancer by Sahoo et al. [56]. The role of MUC2 in protecting the integrity of the mucosa was mentioned by Huang et al. [57]. The authors mentioned that it is possible to induce colitis in mice by suppressing the MUC2 gene. Heimel et al. [58] found high levels of expression of FABP2 and FABP6 when analyzing alterations in intestinal fatty acid metabolism in IBD. CA1 was mentioned by Xie et al. [59] as playing a role in IBD. PHGR1 was identified by Camilleri et al. [60] as potentially increasing the risk of diverticular disease of the colon. FABP1 was identified as a biomarker for Crohn's disease by Dooley et al. [61]. COL1A2 was mentioned by Prados et al. [62] in murine models of IBD. ENO1 was mentioned by Shkoda et al. [63] for its role in IBD pathobiology. Another gene selected by the algorithm and mentioned in the literature as being related to IBD is NDRG1 [64]. Song et al. [65] showed that ADH1C is downregulated in UC. FN1 was suggested by Al-Numan [66] to be related to the early onset of IBD. SPINT2 plays a role in epithelial adhesion [17]. CLDN7 is associated with colitis according to several authors [67,68]. Darsigny et al. [69] found a link between APOC3 and chronic inflammation in mice resembling IBD. KLF5 was identified by Dong et al. [70] as one of the genes downregulated in IBD. Gorenjak et al. [71] linked HSPA9 with IBD.

One of the challenges, and possible limitations, of this type of analysis is the fact that it is impossible to estimate all possible combinations of genes, and hence it is necessary to use some sort of combinatorial approach, such as the Monte Carlo model used to select the genes. There is also no indication that gene expression and IBD are related by an underlying linear model. Given this assumption, using machine learning techniques, which are adept to modeling nonlinear systems, seems like a reasonable approach. Another factor to take into account is that, while the cohort of cases is not small, including 2490 cases, it can always be larger.

5. Conclusions

Following a machine learning approach, it was possible to identify a list of genes that appear to be related to inflammatory bowel disease. Given the complexity of this illness, which appears to be caused by a combination of polygenic factors as well as environmental factors, which could, in principle, interact in a non-linear way, the illness was analyzed using non-linear models, such as machine learning techniques. This approach was able to distinguish, using a small number of genes, between patients with IBD and control (healthy) patients as well as patients with the two major forms of IBD, which are Crohn's disease and ulcerative colitis. In other words, the machine learning algorithms are able to classify different types of gene expression signatures associated with IBD. It might be possible in the future, when more data become available, to be able to distinguish between different genetic signatures of the illness that might potentially help develop more personalized treatments. This is important for an illness as heterogeneous as IBD, for which patients follow different evolutions and might present different clinical manifestations.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/medicina59071218/s1, Supplementary Material: The list of 562 genes.

Author Contributions: Methodology, G.A.P. and R.C.; software, G.A.P; validation, G.A.P. and R.C.; formal analysis, G.A.P. and R.C.; investigation, G.A.P. and R.C.; resources, G.A.P. and R.C.; data curation, G.A.P. and R.C.; writing—original draft preparation, G.A.P.; writing—review and editing, G.A.P. and R.C.; visualization, G.A.P. and R.C.; supervision, G.A.P. and R.C.; project administration, G.A.P. and R.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Spanish Ministerio de Ciencia, Innovación y Universidades (PGC2018-094852-B-C21), and Universitat Jaume I (UJI-B2019-43).

Informed Consent Statement: Not applicable.

Data Availability Statement: The data are accessible at: https://www.ncbi.nlm.nih.gov/geo/ (accessed on 1 June 2023).

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

Abbreviations

The following abbreviations are used in this manuscript:

IBD	Inflammatory Bowel Disease (IBD)
UC	Ulcerative Colitis
CD	Crohn’s Disease
GEO	Gene Expression Omnibus
KNN	k-Nearest Neighbors
SVM	Support Vector Machine
ANN	Artificial Neural Network

References

1. Frolkis, A.; Dieleman, L.A.; Barkema, H.W.; Panaccione, R.; Ghosh, S.; Fedorak, R.N.; Madsen, K.; Kaplan, G.G. Environment and the inflammatory bowel diseases. *Can. J. Gastroenterol.* **2013**, *27*, 18–24. [CrossRef] [PubMed]
2. Rogler, G.; Singh, A.; Kavanaugh, A.; Rubin, D.T. Extraintestinal manifestations of inflammatory bowel disease: Current concepts, treatment, and implications for disease management. *Gastroenterol.* **2021**, *161*, 1118–1132. [CrossRef] [PubMed]
3. Seyedian, S.S.; Nokhostin, F.; Malamir, M.D. A review of the diagnosis, prevention, and treatment methods of inflammatory bowel disease. *J. Med. Life* **2019**, *12*, 113. [CrossRef] [PubMed]
4. Zhang, Y.; Li, Y. Inflammatory bowel disease: Pathogenesis. *World J. Gastroenterol.* **2014**, *20*, 91. [CrossRef] [PubMed]
5. Pithadia, A.B.; Jain, S. Treatment of inflammatory bowel disease (IBD). *Pharmacol. Rep.* **2011**, *63*, 629–642. [CrossRef]
6. Kaplan, G.G.; Windsor, J.W. The four epidemiological stages in the global evolution of inflammatory bowel disease. *Nat. Rev. Gastroenterol. Hepatol.* **2021**, *18*, 56–66. [CrossRef]
7. Xie, J.; Itzkowitz, S.H. Cancer in inflammatory bowel disease. *World J. Gastroenterol.* **2008**, *14*, 378. [CrossRef]
8. Lewandowski, K.; Kaniewska, M.; Więcek, M.; Szwarc, P.; Panufnik, P.; Tulewicz-Marti, E.; Walicka, M.; Franek, E.; Rydzewska, G. Risk factors for osteoporosis among patients with inflammatory bowel disease—Do we already know everything? *Nutrients* **2023**, *15*, 1151. [CrossRef]
9. Xavier, R.J.; Podolsky, D.K. Unravelling the pathogenesis of inflammatory bowel disease. *Nature* **2007**, *448*, 427–434. [CrossRef]
10. Liu, T.-C.; Stappenbeck, T.S. Genetics and pathogenesis of inflammatory bowel disease. *Annu. Rev. Pathol. Mech. Dis.* **2016**, *11*, 127–148. [CrossRef]
11. Neurath, M.F. Cytokines in inflammatory bowel disease. *Nat. Rev. Immunol.* **2014**, *14*, 329–342. [CrossRef] [PubMed]
12. Kuang, R.; O’Keefe, S.J.D.; Ramos del Aguila de Rivers, C.; Koutroumpakis, F.; Binion, D.G. Is salt at fault? Dietary salt consumption and inflammatory bowel disease. *Inflammatory Bowel Diseases* **2023**, *29*, 140–150. [CrossRef] [PubMed]
13. Baumgart, D.C.; Carding, S.R. Inflammatory bowel disease: Cause and immunobiology. *Lancet* **2007**, *369*, 1627–1640. [CrossRef] [PubMed]
14. Shanahan, F. Inflammatory bowel disease: Immunodiagnostics, immunotherapeutics, and ecotherapeutics. *Gastroenterology* **2001**, *120*, 622–635. [CrossRef] [PubMed]
15. Tamboli, C.P.; Neut, C.; Desreumaux, P.; Colombel, J.F. Dysbiosis in inflammatory bowel disease. *Gut* **2004**, *53*, 1–4. [CrossRef] [PubMed]
16. Chang, J.T. Pathophysiology of inflammatory bowel diseases. *New Engl. J. Med.* **2020**, *383*, 2652–2664. [CrossRef] [PubMed]
17. Graham, D.B.; Xavier, R.J. Pathway paradigms revealed from the genetics of inflammatory bowel disease. *Nature* **2020**, *578*, 527–539. [CrossRef]

18. Strober, W.; Fuss, I.; Mannon, P. The fundamental basis of inflammatory bowel disease. *J. Clin. Investig.* **2007**, *117*, 514–521. [CrossRef]
19. Loddo, I.; Romano, C. Inflammatory bowel disease: Genetics, epigenetics, and pathogenesis. *Front. Immunol.* **2015**, *6*, 551. [CrossRef]
20. Khor, B.; Gardet, A.; Xavier, R.J. Genetics and pathogenesis of inflammatory bowel disease. *Nature* **2011**, *474*, 307–317. [CrossRef]
21. Duerr, R.H. Update on the genetics of inflammatory bowel disease. *J. Clin. Gastroenterol.* **2003**, *37*, 358–367. [CrossRef] [PubMed]
22. McGovern, D.P.B.; Kugathasan, S.; Cho, J.H. Genetics of inflammatory bowel diseases. *Gastroenterology* **2015**, *149*, 1163–1176. [CrossRef] [PubMed]
23. Eguchi, R.; Karim, M.B.; Hu, P.; Sato, T.; Ono, N.; Kanaya, S.; Altaf-Ul-Amin, M. An integrative network-based approach to identify novel disease genes and pathways: A case study in the context of inflammatory bowel disease. *BMC Bioinform.* **2018**, *19*, 1–12. [CrossRef] [PubMed]
24. Lees, C.W.; Satsangi, J. Genetics of inflammatory bowel disease: Implications for disease pathogenesis and natural history. *Expert Rev. Gastroenterol. Hepatol.* **2009**, *3*, 513–534. [CrossRef]
25. Cho, J.H.; Abraham, C. Inflammatory bowel disease genetics: Nod2. *Annu. Rev. Med.* **2007**, *58*, 401–416. [CrossRef]
26. Bonen, D.K.; Cho, J.H. The genetics of inflammatory bowel disease. *Gastroenterology* **2003**, *124*, 521–536. [CrossRef]
27. Kakuta, Y.; Naito, T.; Kinouchi, Y.; Masamune, A. Current Status and Future Prospects of Inflammatory Bowel Disease Genetics. *Digestion* **2023**, *104*, 7–15. [CrossRef]
28. Mathew, C.G.; Lewis, C.M. Genetics of inflammatory bowel disease: Progress and prospects. *Digestion* **2004**, *13*, 161–168. [CrossRef]
29. Achkar, J.; Duerr, R. The expanding universe of inflammatory bowel disease genetics. *Curr. Opin. Gastroenterol.* **2008**, *24*, 429–434. [CrossRef]
30. Van Limbergen, J.; Russell, R.K.; Nimmo, E.R.; Satsangi, J. The genetics of inflammatory bowel disease. *Off. J. Am. Coll. Gastroenterol.* **2007**, *102*, 2820–2831. [CrossRef]
31. Stoll, M.; Corneliussen, B.; Costello, C.M.; Waetzig, G.H.; Mellgard, B.; Koch, W.A.; Rosenstiel, P.; Albrecht, M.; Croucher, P.J.P.; Seegert, D. Genetic variation in DLG5 is associated with inflammatory bowel disease. *Nat. Genet.* **2004**, *36*, 476–480. [CrossRef]
32. Cleynen, I.; Boucher, G.; Jostins, L.; Schumm, L.P.; Zeissig, S.; Ahmad, T.; Andersen, V.; Andrews, J.M.; Annese, V.; Brand, S. Inherited determinants of Crohn's disease and ulcerative colitis phenotypes: A genetic association study. *Lancet* **2016**, *387*, 156–167. [CrossRef] [PubMed]
33. Ahmad, T.; Satsangi, J.; McGovern, D.; Bunce, M.; Jewell, D.P. The genetics of inflammatory bowel disease. *Aliment. Pharmacol. Ther.* **2001**, *15*, 731–748. [CrossRef] [PubMed]
34. Hazel, K.; O'Connor, A. Emerging treatments for inflammatory bowel disease. *Ther. Adv. Chronic Dis.* **2020**, *11*, 1–12. [CrossRef] [PubMed]
35. Dhillon, A.; Singh, A.; Bhalla, V.K. A systematic review on biomarker identification for cancer diagnosis and prognosis in multi-omics: From computational needs to machine learning and deep learning. *Arch. Comput. Methods Eng.* **2023**, *30*, 917–949. [CrossRef]
36. Bhatt, M.; Shende, P. Advancement in Machine Learning: A Strategic Lookout from Cancer Identification to Treatment. *Arch. Comput. Methods Eng.* **2023**, *2023*, 1–16. [CrossRef]
37. Alfonso Perez, G.; Castillo, R. Identification of Systemic Sclerosis through Machine Learning Algorithms and Gene Expression. *Mathematics* **2022**, *10*, 4632. [CrossRef]
38. Mossotto, E.; Ashton, J.J.; Coelho, T.; Beattie, R.M.; MacArthur, B.D.; Ennis, S.J.S.R. Classification of paediatric inflammatory bowel disease using machine learning. *Sci. Rep.* **2017**, *7*, 1–10. [CrossRef]
39. Kohli, A.; Holzwanger, E.A.; Levy, A.N. Emerging use of artificial intelligence in inflammatory bowel disease. *World J. Gastroenterol.* **2020**, *26*, 6923. [CrossRef]
40. Gardiner, L.; Carrieri, A.P.; Bingham, K.; Macluskie, G.; Bunton, D.; McNeil, M.; Pyzer-Knapp, E.O. Combining explainable machine learning, demographic and multi-omic data to inform precision medicine strategies for inflammatory bowel disease. *PLoS ONE* **2022**, *17*, e0263248. [CrossRef]
41. Gubatan, J.; Levitte, S.; Patel, A.; Balabanis, T.; Wei, M.T.; Sinha, S.R. Artificial intelligence applications in inflammatory bowel disease: Emerging technologies and future directions. *World J. Gastroenterol.* **2021**, *27*, 1920. [CrossRef] [PubMed]
42. Wei, Z.; Wang, W.; Bradfield, J.; Li, J.; Cardinale, C.; Frackelton, E.; Kim, C.; Mentch, F.; Van Steen, K.; Visscher, P.M. Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. *Am. J. Hum. Genet.* **2013**, *92*, 1008–1012. [CrossRef] [PubMed]
43. 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] [PubMed]
44. Coelho, T.; Mossotto, E.; Gao, Y.; Haggarty, R.; Ashton, J.J.;Batra, A.; Stafford, I.S.; Beattie, R.M.; Williams, A.P.; Ennis, S. Immunological profiling of paediatric inflammatory bowel disease using unsupervised machine learning. *J. Pediatr. Gastroenterol. Nutr.* **2020**, *70*, 833–840. [CrossRef] [PubMed]
45. Smolander, J.; Dehmer, M.; Emmert-Streib, F. Comparing deep belief networks with support vector machines for classifying gene expression data from complex disorders. *FEBS Open Bio.* **2019**, *9*, 1232–1248. [CrossRef]

46. Stankovic, B.; Kotur, N.; Nikcevic, G.; Gasic, V.; Zukic, B.; Pavlovic, S. Machine learning modeling from omics data as prospective tool for improvement of inflammatory bowel disease diagnosis and clinical classifications. *Genes* **2017**, *12*, 1438. [CrossRef]
47. Argmann, C.; Hou, R.; Ungaro, R.C.; Irizar, H.; Al-Taie, Z.; Huang, R.; Kosoy, R.; Venkat, S.; Song, W.; Di'Narzo, A.F. Biopsy and blood-based biomarker of inflammation in IBD. *Gut* **2022**, *2022*, 1271–1287. [CrossRef]
48. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser.* **1996**, *58*, 267–288. [CrossRef]
49. Roth, V. The generalized LASSO. *IEEE Trans. Neural Netw.* **2004**, *15*, 16–28. [CrossRef]
50. Muthukrishnan, R.; Rohini, R. LASSO: A feature selection technique in predictive modeling for machine learning. In Proceedings of the 2016 IEEE International Conference on Advances in Computer Applications (ICACA), Coimbatore, India, 24 October 2016; pp. 18–20.
51. Wu, F.; Yuan, Y.; Zhuang, Y.Heterogeneous feature selection by group lasso with logistic regression. In Proceedings of the 18th ACM international conference on Multimedia, San Jose, CA, USA, 3–5 June 2010; pp. 983–986.
52. Yamamoto-Furusho, J.K. Genetic factors associated with the development of inflammatory bowel disease. *J. Gastroenterol.* **2007**, *13*, 5594. [CrossRef]
53. Krzystek-Korpacka, M.; Diakowska, D.; Bania, J.; Gamian, A. Expression stability of common housekeeping genes is differently affected by bowel inflammation and cancer: Implications for finding suitable normalizers for inflammatory bowel disease studies. *Inflamm. Bowel Dis.* **2014**, *20*, 1147–1156. [CrossRef] [PubMed]
54. Bednarz-Misa, I.; Neubauer, K.; Zacharska, E.; Kapturkiewicz, B.; Krzystek-Korpacka, M. Whole blood ACTB, B2M and GAPDH expression reflects activity of inflammatory bowel disease, advancement of colorectal cancer, and correlates with circulating inflammatory and angiogenic factors: Relevance for real-time quantitative PCR. *Adv. Clin. Exp. Med.* **2020**, *29*. [CrossRef] [PubMed]
55. Li, Y.; Zhu, L.; Chen, P.; Wang, Y.; Yang, G.; Zhou, G.; Li, L.; Feng, R.; Qiu, Y.; Han, J. MALAT1 maintains the intestinal mucosal homeostasis in Crohn's disease via the miR-146b-5p-CLDN11/NUMB pathway. *J. Crohn'S Colitis* **2021**, *15*, 1542–1557. [CrossRef] [PubMed]
56. Sahoo, D.K.; Borcherding, D.C.; Chandra, L.; Jergens, A.E.; Atherly, T.; Bourgois-Mochel, A.; Ellinwood, N.M.; Snella, E.; Severin, A.J.; Martin, M. Differential transcriptomic profiles following stimulation with lipopolysaccharide in intestinal organoids from dogs with inflammatory bowel disease and intestinal mast cell tumor. *Cancers* **2021**, *14*, 3525. [CrossRef] [PubMed]
57. Huang, E.Y.; Inoue, T.; Leone, V.A.; Dalal, S.; Touw, K.; Wang, Y.; Musch, M.W.; Theriault, B.; Higuchi, K.; Donovan, S. Using corticosteroids to reshape the gut microbiome: Implications for inflammatory bowel diseases. *Inflamm. Bowel Dis.* **2015**, *21*, 963–972. [CrossRef] [PubMed]
58. Heimerl, S.; Moehle, C.; Zahn, A.; Boettcher, A.; Stremmel, W.; Langmann, T.; Schmitz, G. Alterations in intestinal fatty acid metabolism in inflammatory bowel disease. *Biochim. Biophys. Acta Mol. Basis Dis.* **2006**, *1762*, 341–350. [CrossRef]
59. Xie, D.; Zhang, Y.; Qu, H. Crucial genes of inflammatory bowel diseases explored by gene expression profiling analysis. *Scand. J. Gastroenterol.* **2018**, *53*, 685–691. [CrossRef]
60. Camilleri, M.; Sandler, R.S.; Peery, A.F. Etiopathogenetic mechanisms in diverticular disease of the colon. *Cell. Mol. Gastroenterol. Hepatol.* **2020**, *9*, 15–32. [CrossRef]
61. Dooley, T.P.; Curto, E.V.; Reddy, S.P.; Davis, R.L.; Lambert, G.W.; Wilborn, T.W.; Elson, C.O. Regulation of gene expression in inflammatory bowel disease and correlation with IBD drugs. Screening by DNA microarrays. *Inflamm. Bowel Dis.* **2004**, *10*, 1–14. [CrossRef]
62. Prados, M.E.; García-Martín, A.; Unciti-Broceta, J.D.; Palomares, B.; Collado, J.A.; Minassi, A.; Calzado, M.A.; Appendino, G.; Muñoz, E. Betulinic acid hydroxamate prevents colonic inflammation and fibrosis in murine models of inflammatory bowel disease. *Acta Pharmacol. Sin.* **2021**, *42*, 1124–1138. [CrossRef]
63. Shkoda, A.; Werner, T.; Daniel, H.; Gunckel, M.; Rogler, G.; Haller, D. Differential protein expression profile in the intestinal epithelium from patients with inflammatory bowel disease. *J. Proteome Res.* **2007**, *6*, 1114–1125. [CrossRef] [PubMed]
64. Knyazev, E.; Maltseva, D.; Raygorodskaya, M.; Shkurnikov, M. HIF-dependent NFATC1 activation upregulates ITGA5 and PLAUR in intestinal epithelium in inflammatory bowel disease. *Front. Genet.* **2021**, *2277*, 1–12. [CrossRef] [PubMed]
65. Song, F.; Zhang, Y.; Pan, Z.; Hu, X.; Zhang, Q.; Huang, F.; Ye, X.; Huang, P. The role of alcohol dehydrogenase 1C in regulating inflammatory responses in ulcerative colitis. *Front. Genet.* **2021**, *192*, 114691. [CrossRef] [PubMed]
66. Al-Numan, H.H.; Jan, R.M.; Al-Saud, N.B.S.; Rashidi, O.M.; Alrayes, N.M.; Alsufyani, H.A.; Mujalli, A.; Shaik, N.A.; Mosli, M.H.; Elango, R. Exome sequencing identifies the extremely rare ITGAV and FN1 variants in early onset inflammatory bowel disease patients. *Front. Pediatr.* **2022**, *10*, 859074. [CrossRef]
67. Ding, Y.; Wang, K.; Xu, C.; Hao, M.; Li, H.; Ding, L. Intestinal Claudin-7 deficiency impacts the intestinal microbiota in mice with colitis. *BMC Gastroenterol.* **2022**, *22*, 24. [CrossRef]
68. Wang, K.; Ding, Y.; Xu, C.; Hao, M.; Li, H.; Ding, L. Cldn-7 deficiency promotes experimental colitis and associated carcinogenesis by regulating intestinal epithelial integrity. *Oncoimmunology* **2021**, *10*, 1923910. [CrossRef]
69. Darsigny, M.; Babeu, J.P.; Dupuis, A.A.; Furth, E.; Seidman, E.G.; Lévy, E.; Verdu, E.F.; Gendron, F.P.; Boudreau, F. Loss of hepatocyte-nuclear-factor-4α affects colonic ion transport and causes chronic inflammation resembling inflammatory bowel disease in mice. *PLoS ONE* **2009**, *4*, *7609*. [CrossRef]

70. Dong, J.T.; Chen, C. Essential role of KLF5 transcription factor in cell proliferation and differentiation and its implications for human diseases. *Cell. Mol. Life Sci.* **2009**, *66*, 2691–2706. [CrossRef]
71. Gorenjak, M.; Jezernik, G.; Krušič, M.; Skok, P.; Potočnik, U. Identification of Novel Loci Involved in Adalimumab Response in Crohn's Disease Patients Using Integration of Genome Profiling and Isoform-Level Immune-Cell Deconvoluted Transcriptome Profiling of Colon Tissue. *Pharmaceutics* **2022**, *14*, 1893. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

CHAPTER 4

DISCUSSION

4.1 Protein classification using machine learning techniques

Mapping categorical variables into numerical variables is a common practice in many machine learning classification tasks, and it is frequently carried out in an arbitrary matter. In this dissertation, we proposed four different assumptions related to this topic in the context of protein classification: (1) translation, (2) permutation, (3) constant, and (4) eigenvalues. Assumptions 1–3 are related to the concept of equivalent mappings in which changes to the mapping should, in principle, not alter the results of a classification analysis (for instance, adding a constant to all the input parameters). Assumption 4 relates to a less strict requirement in which the mappings are not in principle strictly equivalent, but they are comparable. An example is the eigenvalue mapping approach in which the information about the order of the amino acids (present in the initial mapping) is not contained in this new mapping. The results for Assumptions 1–3 showed that, in the majority of the cases, no statistically significant difference exists between the mappings when we compared their mean accuracy. The case of Assumption 4 is different, and we see that using the eigenvalue approach generates similar or more accurate classifications than the base case model. All these numerical simulations were carried out for 23 different classification algorithms, including KNN, Tress, and SVMs. As previously mentioned, the eigenvalue approach (related to Assumption 4) generated accurate estimations for most algorithms. One noticeable exception was SVM, which, in many cases, failed to generate a classification estimation and was, therefore, excluded from the analysis. For the majority of the other algorithms, the eigenvalue approach generated results that were not statistically significantly different from the base case or that had higher mean accuracy than the base case. The best model obtained a mean classification accuracy of 83.25%. While direct comparisons are challenging, this result is 14.15% better than the lower-bound result obtained by Cai et al. [213] but lower than the upper bound. This is consistent with the idea of focusing the analysis on the stability of results rather than only focusing on increasing accuracy. This result is also substantially higher than the lower bound achieved by Karchin et al. [214], in which the authors focused on a specific subset of proteins

An optimization analysis algorithm was also presented for the automated selection of the number of neurons in a classification model using only the frequency of the occurrence of amino acid in the amino acid chain as input (no order information), as well as the length of the chain. The model included a quadratic penalty function to try to decrease the chance of overfitting. This approach generated an accuracy of 85.02% percent. This result is even closer to the upper bound (and substantially higher than the lower bound) of Cai et al. [213]

even after accounting for the penalty function introduced to avoid an overly complex model, which potentially could impact the generalization capabilities of the model, i.e., the accuracy of the classification when faced with new data. Furthermore, this approach does not require the use of techniques such as NLP, which could be beneficial from an implementation point of view, as there is a large number of machine learning applications that can be easily and accurately applied to numerical values, and there is no indication that an NLP approach will generate more accurate results. It should be noted that this accuracy is not directly comparable with the accuracy obtained in the previous sections, as there was no additional algorithm optimization. The focus was on the comparability of the models, and hence it did not appear appropriate to add additional optimization techniques that differ in the different algorithms. For instance, an optimization process based on finding an appropriate number of neurons, as shown in the optimization section, cannot be performed for other classification techniques such as KNN, SVM, or Trees, as they do not use artificial neurons. This type of big data analysis is challenging and can be computationally expensive, depending on the type of machine learning applied and/or the optimization algorithm followed.

4.2 Non-linear identification of carcinoma

The proposed approach of using DNA methylation data, as inputs, and an algorithm combining ridge regression and artificial neural networks, for the task of differentiating between healthy control individuals and individuals with anal and cervical carcinomas, generated accurate results with specificity and sensitivity higher than ones obtained in other papers in the field. The algorithm selected 13 CpGs from a starting point of approximately 450,000 CpGs per patient. Technological developments have made it possible to obtain such large amounts of methylation data but at the same time have made the analysis of such data challenging. Given that there is no indication that there is a linear relationship between the level of methylation (CpGs) and the presence of anal or cervical carcinoma, the modeling approach was performed with nonlinear techniques such as artificial neural networks. One of the issues with this type of model is the risk of overfitting, particularly in this type of situation in which there is a large number of inputs per patient but a smaller number of patients. In order to reduce this type of risk, it is important to reduce the dimensionality of the data.

Additionally, this reduction in the dimensionality can point to CpGs that might be important as biomarkers in the context of the disease. The selected model was tested for robustness, with the classification estimates remaining accurate for the vast majority of the simulations. No individual CpGs, of those 13 selected by the model, achieved a mean accuracy above 88.94%, which is substantially lower than the 97.69% accuracy obtained by the model. Increasing the complexity of the models, by for instance adding more layers to the neural network, did not appear to increase the accuracy of the model. This might be again related to the issue of overfitting. Similarly, adding more complex penalty functions, such as for instance a quadratic function rather than a linear function, did not improve the accuracy.

4.3 Identification of systemic sclerosis

Systemic sclerosis is a chronic and potentially life threatening illness which is not yet fully understood. The illness has different variants, such as the diffuse form, with different levels of severity in the prognosis. SSc is believed to be caused by a combination of genetic predisposition and environmental factors. While there is currently no curative therapy, there have been many advances on the treatments of related complications of the illness. Some of these complications are potentially life threatening. One common and severe complication of SSc is interstitial lung disease (ILD). In this dissertation, we present an algorithm that uses machine learning techniques, applied to gene expression data, to be able to distinguish between control (healthy) patients and patients suffering from interstitial lung disease systemic sclerosis (ILD-SSc). This algorithm selects the genes (and their expression levels) to be included as inputs into machine learning models for the detection of the illness. The precision of this approach is higher than the one obtained using the genes expression for all the available genes. Having biomarkers that are able to identify the illness might be important from an early detection point of view. The accuracy of the presented model was relatively high, at 92%, with a sensitivity of approximately 75%. Our approach is complementary of some of the existing research in this field that use clinical manifestation of the illness. A potential advantage of using the genetic expression information is that there is no need for the illness to have clear clinical manifestations, such as skin lesions. The approach followed in the algorithm also allowed for the identification of 172 genes that might potentially have some relevance in the context of ILD-SSc. These 172 genes appeared in all the 20 most accurate models (out of half a million models estimated). The assumption is that given the frequency with which these genes appear in the most accurate models, they might be related to the illness. The proposed algorithm was also able to distinguish between the variants of the illness (diffuse). While the precision was lower than in the previous case (distinguishing between control and patients), it was reasonably high with a sensitivity of approximately 72%. This is reasonable, taking into consideration that the illness is likely not only caused by genetic factors but from a combination of genetic factors and environmental exposures.

4.4 Gene Identification on inflammatory bowel disease

Machine learning techniques are used to identify a set of 74 genes, which can be used, with an average accuracy of 84.2%, to distinguish between control (healthy individuals) and patients with inflammatory bowel disease. The specificity and sensitivity of this model are also relatively high at 82.6% and 84.4%, respectively. The selection of these 74 genes is carried out following a Monte Carlo simulation approach. Given that some of the symptoms of inflammatory bowel disease are common in other illnesses, it might be interesting to have another objective diagnostic tool. It is also interesting to observe that among multiple machine learning techniques used in the cohort of patients analyzed, the bagged trees approach seems to consistently achieve a high level of accuracy, particularly when compared to other, arguably more sophisticated machine learning techniques, such as artificial neural networks. The analysis controls for age, gender and region of the biopsy. The proportion of female and male cases is balanced, with 1174 female patients and 1316 male patients. The average age in the cohort is 44.9 years, covering a wide age range (from 19 to 82 years old). The results of the artificial neural networks include an optimization of the hyperparameters with simulations ranging from 1 to 1000 hidden layers. It is also observed that simply increasing the number of layers in an artificial neural network does not necessarily translate into better accuracy. It is also possible to distinguish between the two main types of IBD—Crohn's disease and ulcerative colitis—but in this case with a lower level of accuracy. The accuracy, using this approach is 73.4%. The accuracy, sensitivity and specificity reported are those of the testing dataset. As normal practice, the data are divided into training and testing datasets in an attempt to increase the reproducibility of the analysis. Approximately 20% of the total cases are included in the testing dataset. The relatively large number of genes obtained in the bets model is in line with the prevalent view in the existing literature that the illness is polygenic. There is a high degree of heterogeneity in inflammatory bowel disease, leading to varied severity and evolution of the illness. The existing literature, see, for instance, Yamamoto et al. [215], points towards a polygenic illness with a complex interaction with environmental factors. Our results are consistent with this polygenic description. In this context, it is important to generate algorithms that are able to differentiate among control and patients as well as between different types of inflammatory bowel disease, namely Crohn's disease and ulcerative colitis.

Chapter 5

Conclusions

In paper I the four proposed assumptions in the context of categorical variable mapping in protein classification problems: (1) translation, (2) permutation, (3) constant, and (4) eigenvalues were tested against empirical data. The results suggest that these four assumptions are valid. The first three assumptions are of a more fundamental nature i.e., there is no chemical or biological reasons for them not to be satisfied. The fourth assumption was also tested, with the results suggesting that an eigenvalue approach can be used in the context of protein classification generating accurate results

In paper II the proposed approach is able to generate an accuracy, sensitivity and specify of classification forecasts of 97.69%, 95.02% and 98.26%, respectively, illustrating that a combination of DNA methylation with nonlinear methods such as artificial neural networks might be useful in the task of identifying patients with a carcinoma. This approach could be complementary to the existing techniques such as occult blood test and pap smear. This is conceivable, but additional testing would be required to support this hypothesis, that DNA methylation changes might be present in the patient before there are clinical indications (occult blood test). This is an important research question that should be addressed in future research. Additionally, it is possible that finding different DNA methylation signatures could be used for personalized treatments. This is another area in which more research would be needed. The model achieved a substantial reduction in the number of CpGs used as input from a starting point of approximately 450,000 to only 13. This is important, as having an excessively large number of inputs could lead to overfitting issues. The combination of these 13 CpGs generated more accurate forecasts that any of them individually.

In paper III it was shown that gene expression data can be successfully analyzed with machine learning techniques in order to differentiate healthy patients and patients with interstitial lung disease systemic sclerosis (ILD-SSc). The same approach was also successfully used to differentiate between variants of the illness. This type of approach might be used in the future to provide more Personalized treatments for patients. It was also possible to identify a list of genes that were suggested by the algorithm as related to ILD-SSc.

In paper IV, following a machine learning approach, it was possible to identify a list of genes that appear to be related to inflammatory bowel disease. Given the complexity of this illness, which appears to be caused by a combination of polygenic factors as well as environmental factors, which could, in principle, interact in a non-linear way, the illness was analyzed using non-linear models, such as machine learning techniques. This approach was able to distinguish, using a small number of genes, between patients with IBD and control (healthy) patients as well as patients with the two major forms of IBD, which are

Crohn's disease and ulcerative colitis. In other words, the machine learning algorithms are able to classify different types of gene expression signatures associated with IBD. It might be possible in the future, when more data become available, to be able to distinguish between different genetic signatures of the illness that might potentially help develop more personalized treatments. This is important for an illness as heterogeneous as IBD, for which patients follow different evolutions and might present different clinical manifestations.

CHAPTER 6

FUTURE WORK

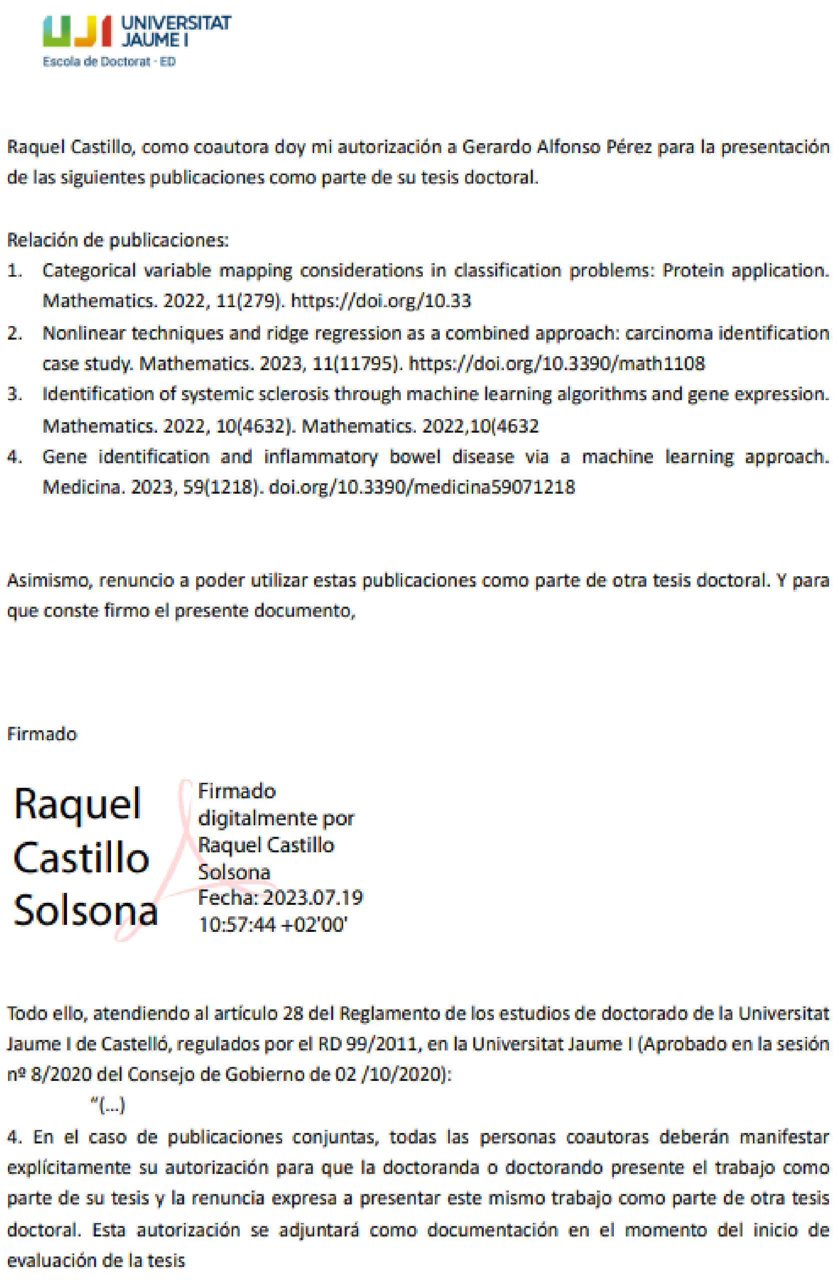
There are some interesting areas of future work. For instance, in the context of protein modeling, it would be interesting to use genetic algorithms or particle algorithms as potential optimization strategies. There is a wide range of options to optimize this type of analysis. There is, however, the risk of overfitting the model, and some measures should be taken to minimize that risk, such as using a penalty function, as we used in this article, to penalize the accuracy of overly complex models. Arguably, an overly complex model is more likely to result in an overfitting issue than a simpler model.

Another interesting area of future research is combining different types of genetic information such as gene expression levels and DNA methylation levels to create more accurate fingerprints of individuals differentiating between healthy and patients suffering from some of the illnesses analyzed in this dissertation. This type of analysis will require both availability of data (gene expression and methylation) as well as having the appropriate tools and algorithms to analyze such information.

Chapter 7

Co-author's consent

Co-author's consent for "tesis por compendio de publicaciones".



UNIVERSITAT
JAUME I
Escola de Doctorat · ED

Raquel Castillo, como coautora doy mi autorización a Gerardo Alfonso Pérez para la presentación de las siguientes publicaciones como parte de su tesis doctoral.

Relación de publicaciones:

1. Categorical variable mapping considerations in classification problems: Protein application. Mathematics. 2022, 11(279). https://doi.org/10.33
2. Nonlinear techniques and ridge regression as a combined approach: carcinoma identification case study. Mathematics. 2023, 11(11795). https://doi.org/10.3390/math1108
3. Identification of systemic sclerosis through machine learning algorithms and gene expression. Mathematics. 2022, 10(4632). Mathematics. 2022,10(4632
4. Gene identification and inflammatory bowel disease via a machine learning approach. Medicina. 2023, 59(1218). doi.org/10.3390/medicina59071218

Asimismo, renuncio a poder utilizar estas publicaciones como parte de otra tesis doctoral. Y para que conste firmo el presente documento,

Firmado

Raquel
Castillo
Solsona

Firmado
digitalmente por
Raquel Castillo
Solsona
Fecha: 2023.07.19
10:57:44 +02'00'

Todo ello, atendiendo al artículo 28 del Reglamento de los estudios de doctorado de la Universitat Jaume I de Castelló, regulados por el RD 99/2011, en la Universitat Jaume I (Aprobado en la sesión nº 8/2020 del Consejo de Gobierno de 02 /10/2020):

"(...)

4. En el caso de publicaciones conjuntas, todas las personas coautoras deberán manifestar explícitamente su autorización para que la doctoranda o doctorando presente el trabajo como parte de su tesis y la renuncia expresa a presentar este mismo trabajo como parte de otra tesis doctoral. Esta autorización se adjuntará como documentación en el momento del inicio de evaluación de la tesis

Bibliography

[1] Igor V Tetko, Ola Engkvist, Uwe Koch, Jean-Louis Reymond, and Hongming Chen.
Bigchem: challenges and opportunities for big data analysis in chemistry.
Molecular informatics, 35(11-12):615–621, 2016.

[2] Scott J Lusher, Ross McGuire, René C van Schaik, C David Nicholson, and Jacob de Vlieg.
Data-driven medicinal chemistry in the era of big data.
Drug discovery today, 19(7):859–868, 2014.

[3] Moises Álvarez-Moreno, Coen de Graaf, Nuria Lopez, Feliu Maseras, Josep M Poblet, and Carles Bo.
Managing the computational chemistry big data problem: the iochem-bd platform.
Journal of chemical information and modeling, 55(1):95–

103, 2015.

[4] Kara Dolinski and Olga G Troyanskaya.
Implications of big data for cell biology.
Molecular biology of the cell, 26(14):2575–2578, 2015.

[5] Sabina Leonelli.
What difference does quantity make? on the epistemology of big data in biology.
Big data & society, 1(1):2053951714534395, 2014.

[6] Md Altaf-Ul-Amin, Farit Mochamad Afendi, Samuel Kuria Kiboi, Shigehiko Kanaya, et al.
Systems biology in the context of big data and networks.
BioMed research international, 2014, 2014.

[7] Jianqing Fan, Fang Han, and Han Liu.
Challenges of big data analysis.
National science review, 1(2):293–314, 2014.

[8] Puneet Singh Duggal and Sanchita Paul.
Big data analysis: challenges and solutions.
In *International conference on cloud, big data and trust*, volume 15, pages 269–276, 2013.

[9] Vivien Marx.
The big challenges of big data.
Nature, 498(7453):255–260, 2013.

[10] Daniel E O'Leary.
Artificial intelligence and big data.
IEEE intelligent systems, 28(2):96–99, 2013.

[11] Hao Zhu.
Big data and artificial intelligence modeling for drug discovery.
Annual review of pharmacology and toxicology, 60:573–589, 2020.

[12] Omar Y Al-Jarrah, Paul D Yoo, Sami Muhaidat, George K Karagiannidis, and Kamal Taha.
Efficient machine learning for big data: A review.
Big Data Research, 2(3):87–93, 2015.

[13] José Jiménez, Stefan Doerr, Gerard Martínez-Rosell, Alexander S Rose, and Gianni De Fabritiis.
Deepsite: protein-binding site predictor using 3d-convolutional neural networks.

Bioinformatics, 33(19):3036–3042, 2017.

[14] Guillaume Pagès, Benoit Charmettant, and Sergei Grudinin.
Protein model quality assessment using 3d oriented convolutional neural networks.
Bioinformatics, 35(18):3313–3319, 2019.

[15] Xiao Wang, Genki Terashi, Charles W Christoffer, Mengmeng Zhu, and Daisuke Kihara.
Protein docking model evaluation by 3d deep convolutional neural networks.
Bioinformatics, 36(7):2113–2118, 2020.

[16] Matthew Ragoza, Joshua Hochuli, Elisa Idrobo, Jocelyn Sunseri, and David Ryan Koes.
Protein–ligand scoring with convolutional neural networks.
Journal of chemical information and modeling, 57(4):942–957, 2017.

[17] John A Keith, Valentin Vassilev-Galindo, Bingqing

Cheng, Stefan Chmiela, Michael Gastegger, Klaus-Robert Muuller, and Alexandre Tkatchenko.
Combining machine learning and computational chemistry for predictive insights into chemical systems.
Chemical reviews, 121(16):9816–9872, 2021.

[18] Nongnuch Artrith, Keith T Butler, François-Xavier Coudert, Seungwu Han, Olexandr Isayev, Anubhav Jain, and Aron Walsh.
Best practices in machine learning for chemistry.
Nature chemistry, 13(6):505–508, 2021.

[19] Joel R Bock and David A Gough.
Predicting protein–protein interactions from primary structure.
Bioinformatics, 17(5):455–460, 2001.

[20] Alexander Radovic, Mike Williams, David Rousseau, Michael Kagan, Daniele Bonacorsi, Alexander Himmel, Adam Aurisano, Kazuhiro Terao, and Taritree Wongjirad.

Machine learning at the energy and intensity frontiers of particle physics.
Nature, 560(7716):41–48, 2018.

[21] George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang.
Physics-informed machine learning.
Nature Reviews Physics, 3(6):422–440, 2021.

[22] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann.
Software engineering for machine learning: A case study.
In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, pages 291–300. IEEE, 2019.

[23] Cheolsoo Park, Clive Cheong Took, and Joon-Kyung Seong.
Machine learning in biomedical engineering.
Biomedical Engineering Letters, 8:1–3, 2018.

[24] Du Zhang and Jeffrey JP Tsai.

Machine learning and software engineering.
Software Quality Journal, 11:87–119, 2003.

[25] Di Qi and Andrew J Majda.
Using machine learning to predict extreme events in complex systems.
Proceedings of the National Academy of Sciences, 117(1):52–59, 2020.

[26] David A Wood.
A transparent open-box learning network provides insight to complex systems and a performance benchmark for more-opaque machine learning algorithms.
Advances in Geo-Energy Research, 2(2):148–162, 2018.

[27] Jian Qin, Fu Hu, Ying Liu, Paul Witherell, Charlie CL Wang, David W Rosen, Timothy W Simpson, Yan Lu, and Qian Tang.
Research and application of machine learning for additive manufacturing.
Additive Manufacturing, 52:102691, 2022.

[28] Keiron O'Shea and Ryan Nash.

An introduction to convolutional neural networks.
arXiv preprint arXiv:1511.08458, 2015.

[29] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun.
Shufflenet: An extremely efficient convolutional neural network for mobile devices.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018.

[30] Sebastian Ventura, Manuel Silva, Dolores Perez-Bendito, and Cesar Hervas.
Artificial neural networks for estimation of kinetic analytical parameters.
Analytical Chemistry, 67(9):1521–1525, 1995.

[31] Yuting Xu, Deeptak Verma, Robert P Sheridan, Andy Liaw, Junshui Ma, Nicholas M Marshall, John McIntosh, Edward C Sherer, Vladimir Svetnik, and Jennifer M Johnston.
Deep dive into machine learning models for protein engineering.

Journal of chemical information and modeling, 60(6):2773–2790, 2020.

[32] Ayodeji Olalekan Salau and Shruti Jain.
Adaptive diagnostic machine learning technique for classification of cell decisions for akt protein.
Informatics in Medicine Unlocked, 23:100511, 2021.

[33] Yogesh Singh, Pradeep Kumar Bhatia, and Omprakash Sangwan.
A review of studies on machine learning techniques.
International Journal of Computer Science and Security, 1(1):70–84, 2007.

[34] Mario Bkassiny, Yang Li, and Sudharman K Jayaweera.
A survey on machine-learning techniques in cognitive radios.
IEEE Communications Surveys & Tutorials, 15(3):1136–1159, 2012.

[35] Mohammad Tanveer, T Rajani, Reshma Rastogi, Yuan-Hai Shao, and MA Ganaie.
Comprehensive review on twin support vector machines.

Annals of Operations Research, pages 1–46, 2022.

[36] Zhi Hong Kok, Abdul Rashid Mohamed Shariff, Meftah Salem M Alfatni, and Siti Khairunniza-Bejo.
Support vector machine in precision agriculture: a review.
Computers and Electronics in Agriculture, 191:106546, 2021.

[37] Isaac Triguero, Diego García-Gil, Jesús Maillo, Julián Luengo, Salvador García, and Francisco Herrera.
Transforming big data into smart data: An insight on the use of the k-nearest neighbors algorithm to obtain quality data.
Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(2):e1289, 2019.

[38] Jennifer McDowall and Sarah Hunter.
Interpro protein classification.
Bioinformatics for Comparative Proteomics, pages 37–47, 2011.

[39] Loris Nanni, Alessandra Lumini, Sheryl Brahnam, et al.

An empirical study of different approaches for protein classification.
The Scientific World Journal, 2014, 2014.

[40] Sotiris Diplaris, Grigorios Tsoumakas, Pericles A Mitkas, and Ioannis Vlahavas.
Protein classification with multiple algorithms.
In *Advances in Informatics: 10th Panhellenic Conference on Informatics, PCI 2005, Volas, Greece, November 11-13, 2005. Proceedings 10*, pages 448–456. Springer, 2005.

[41] Seref Sagiroglu and Duygu Sinanc.
Big data: A review.
In *2013 international conference on collaboration technologies and systems (CTS)*, pages 42–47. IEEE, 2013.

[42] Sam Madden.
From databases to big data.
IEEE Internet Computing, 16(3):4–6, 2012.

[43] Protein Data Bank.
Protein data bank.

Nature New Biol, 233:223, 1971.

[44] Cynthia Rudin.
Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.
Nature machine intelligence, 1(5):206–215, 2019.

[45] Amy McGovern, Ryan Lagerquist, David John Gagne, G Eli Jergensen, Kimberly L Elmore, Cameron R Homeyer, and Travis Smith.
Making the black box more transparent: Understanding the physical implications of machine learning.
Bulletin of the American Meteorological Society, 100(11):2175–2199, 2019.

[46] Zhi-Hua Zhou.
Learnware: on the future of machine learning.
Frontiers Comput. Sci., 10(4):589–590, 2016.

[47] Alberto Blanco-Justicia and Josep Domingo-Ferrer.
Machine learning explainability through comprehensible decision trees.

In *Machine Learning and Knowledge Extraction: Third IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2019, Canterbury, UK, August 26–29, 2019, Proceedings 3*, pages 15–26. Springer, 2019.

[48] Xu Chen, Yiqiu Gao, Yunlong Wang, and Guoqing Pan. Mussel-inspired peptide mimicking: An emerging strategy for surface bioengineering of medical implants. *Smart Materials in Medicine*, 2:26–37, 2021.

[49] Mehdi Kazemzadeh-Narbat, Hao Cheng, Rosa Chabok, Mario Moisés Alvarez, Cesar De La Fuente-Nunez, K Scott Phillips, and Ali Khademhosseini. Strategies for antimicrobial peptide coatings on medical devices: A review and regulatory science perspective. *Critical reviews in biotechnology*, 41(1):94–120, 2021.

[50] Vasso Apostolopoulos, Joanna Bojarska, Tsun-Thai Chai, Sherif Elnagdy, Krzysztof Kaczmarek, John Matsoukas, Roger New, Keykavous Parang, Octavio Paredes Lopez, Hamideh Parhiz, et al.

A global review on short peptides: Frontiers and perspectives.
Molecules, 26(2):430, 2021.

[51] Phasit Charoenkwan, Wararat Chiangjong, Md M Hasan, Chanin Nantasenamat, and Watshara Shoombuatong.
Review and comparative analysis of machine learning-based predictors for predicting and analyzing anti-angiogenic peptides.
Current Medicinal Chemistry, 29(5):849–864, 2022.

[52] Christopher D Fjell, Håvard Jenssen, Kai Hilpert, Warren A Cheung, Nelly Panté, Robert EW Hancock, and Artem Cherkasov.
Identification of novel antibacterial peptides by chemoinformatics and machine learning.
Journal of medicinal chemistry, 52(7):2006–2015, 2009.

[53] Patricio Cerda, Gaël Varoquaux, and Balázs Kégl.
Similarity encoding for learning with dirty categorical variables.
Machine Learning, 107(8-10):1477–1494, 2018.

[54] Patricio Cerda and Gaël Varoquaux.
Encoding high-cardinality string categorical variables.
IEEE Transactions on Knowledge and Data Engineering, 34(3):1164–1176, 2020.

[55] Paolo Sonego, Mircea Pacurar, Somdutta Dhir, Attila Kertesz-Farkas, András Kocsor, Zoltán Gáspári, Jack AM Leunissen, and Sándor Pongor.
A protein classification benchmark collection for machine learning.
Nucleic acids research, 35(suppl_1):D232–D236, 2007.

[56] Alberto Prieto, Beatriz Prieto, Eva Martinez Ortigosa, Eduardo Ros, Francisco Pelayo, Julio Ortega, and Ignacio Rojas.
Neural networks: An overview of early research, current frameworks and new challenges.
Neurocomputing, 214:242–268, 2016.

[57] Ananthan Nambiar, Simon Liu, Maeve Heflin, John Malcolm Forsyth, Sergei Maslov, Mark Hopkins, and Anna Ritz.

Transformer neural networks for protein family and interaction prediction tasks.

Journal of Computational Biology, 30(1):95–111, 2023.

[58] HaiXia Long, Mi Wang, and HaiYan Fu.

Deep convolutional neural networks for predicting hydroxyproline in proteins.

Current Bioinformatics, 12(3):233–238, 2017.

[59] Roi Livni, Shai Shalev-Shwartz, and Ohad Shamir.

On the computational efficiency of training neural networks.

Advances in neural information processing systems, 27, 2014.

[60] Vladimir M Krasnopolsky, Dmitry V Chalikov, and Hendrik L Tolman.

A neural network technique to improve computational efficiency of numerical oceanic models.

Ocean Modelling, 4(3-4):363–383, 2002.

[61] Emmanuel Bengio, Pierre-Luc Bacon, Joelle Pineau, and Doina Precup.

Conditional computation in neural networks for faster models.
arXiv preprint arXiv:1511.06297, 2015.

[62] Xin Yao.
A review of evolutionary artificial neural networks.
International journal of intelligent systems, 8(4):539–567, 1993.

[63] Sonali B Maind, Priyanka Wankar, et al.
Research paper on basic of artificial neural network.
International Journal on Recent and Innovation Trends in Computing and Communication, 2(1):96–100, 2014.

[64] Md Monirul Islam and Kazuyuki Murase.
A new algorithm to design compact two-hidden-layer artificial neural networks.
Neural Networks, 14(9):1265–1278, 2001.

[65] Yoshiyasu Takefuji.
Neural network parallel computing, volume 164.
Springer Science & Business Media, 1992.

[66] Simon Ascher, William Sloan, Ian Watson, and Siming You.
A comprehensive artificial neural network model for gasification process prediction.
Applied Energy, 320:119289, 2022.

[67] D Hashim, P Boffetta, C La Vecchia, M4 Rota, P Bertuccio, M Malvezzi, and E Negri.
The global decrease in cancer mortality: trends and disparities.
Annals of Oncology, 27(5):926–933, 2016.

[68] Clive Osmond and MJ Gardner.
Age, period and cohort models applied to cancer mortality rates.
Statistics in Medicine, 1(3):245–259, 1982.

[69] C Bosetti, P Bertuccio, M Malvezzi, F Levi, L Chatenoud, E Negri, and C La Vecchia.
Cancer mortality in europe, 2005–2009, and an overview of trends since 1980.
Annals of oncology, 24(10):2657–2671, 2013.

[70] Cathy Eng, Kristen K Ciombor, May Cho, Jennifer A Dorth, Lakshmi N Rajdev, David P Horowitz, Marc J Gollub, Alexandre A Jácome, Natalie A Lockney, Roberta L Muldoon, et al.
Anal cancer: Emerging standards in a rare disease.
Journal of Clinical Oncology, 40(24):2774–2788, 2022.

[71] Ashish A Deshmukh, Ryan Suk, Meredith S Shiels, Kalyani Sonawane, Alan G Nyitray, Yuxin Liu, Michael M Gaisa, Joel M Palefsky, and Keith Sigel.
Recent trends in squamous cell carcinoma of the anus incidence and mortality in the united states, 2001–2015.
JNCI: Journal of the National Cancer Institute, 112(8):829–838, 2020.

[72] Wanxu Huang, Hua Li, Qingsong Yu, Wei Xiao, and Dan Ohtan Wang.
Lncrna-mediated dna methylation: an emerging mechanism in cancer and beyond.
Journal of Experimental & Clinical Cancer Research,

41(1):100, 2022.

[73] Antonios Papanicolau-Sengos and Kenneth Aldape. Dna methylation profiling: an emerging paradigm for cancer diagnosis. *Annual Review of Pathology: Mechanisms of Disease*, 17:295–321, 2022.

[74] Ashley L Monsrud, Vaidehi Avadhani, Marina B Mosunjac, Lisa Flowers, and Uma Krishnamurti. Programmed death ligand-1 expression is associated with poorer survival in anal squamous cell carcinoma. *Archives of Pathology & Laboratory Medicine*, 146(9):1094–1101, 2022.

[75] Jun Zhang, Ciro R Martins, Zoya B Fansler, Kristina L Roemer, Erik A Kincaid, Karen S Gustafson, Daniel F Heitjan, and Douglas P Clark. Dna methylation in anal intraepithelial lesions and anal squamous cell carcinoma. *Clinical cancer research*, 11(18):6544–6549, 2005.

[76] Erin M Siegel, Abidemi Ajidahun, Anders Berglund,

Whitney Guerrero, Steven Eschrich, Ryan M Putney, Anthony Magliocco, Bridget Riggs, Kathryn Winter, Jeff P Simko, et al.

Genome-wide host methylation profiling of anal and cervical carcinoma.

PloS one, 16(12):e0260857, 2021.

[77] Rakesh Singal and Gordon D Ginder.

Dna methylation.

Blood, The Journal of the American Society of Hematology, 93(12):4059–4070, 1999.

[78] Lisa D Moore, Thuc Le, and Guoping Fan.

Dna methylation and its basic function.

Neuropsychopharmacology, 38(1):23–38, 2013.

[79] Zachary D Smith and Alexander Meissner.

Dna methylation: roles in mammalian development.

Nature Reviews Genetics, 14(3):204–220, 2013.

[80] Geneviève P Delcuve, Mojgan Rastegar, and James R Davie.

Epigenetic control.

Journal of cellular physiology, 219(2):243–250, 2009.

[81] Bob Weinhold.
Epigenetics: the science of change, 2006.

[82] Shelley L Berger, Tony Kouzarides, Ramin Shiekhattar, and Ali Shilatifard.
An operational definition of epigenetics.
Genes & development, 23(7):781–783, 2009.

[83] Marc Jung and Gerd P Pfeifer.
Aging and dna methylation.
BMC biology, 13(1):1–8, 2015.

[84] Bruce Richardson.
Impact of aging on dna methylation.
Ageing research reviews, 2(3):245–261, 2003.

[85] Evan Calabrese, Jeffrey D Rudie, Andreas M Rauschecker, Javier E Villanueva-Meyer, Jennifer L Clarke, David A Solomon, and Soonmee Cha.

Combining radiomics and deep convolutional neural network features from preoperative mri for predicting clinically relevant genetic biomarkers in glioblastoma.
Neuro-Oncology Advances, 4(1):vdac060, 2022.

[86] Alberto Mario Marchevsky.
The use of artificial neural networks for the diagnosis and estimation of prognosis in cancer patients.
In *Outcome prediction in cancer*, pages 243–259. Elsevier, 2007.

[87] Allen N Sapadin and Raul Fleischmajer.
Treatment of scleroderma.
Archives of dermatology, 138(1):99–105, 2002.

[88] Christopher P Denton and Dinesh Khanna.
Systemic sclerosis.
The Lancet, 390(10103):1685–1699, 2017.

[89] Eric Y Yen, Devanshu R Singh, and Ram R Singh.
Trends in systemic sclerosis mortality over forty-eight years, 1968–2015: a us population–based study.
Arthritis care & research, 73(10):1502–1510, 2021.

[90] Yannick Allanore, Robert Simms, Oliver Distler, Maria Trojanowska, Janet Pope, Christopher P Denton, and John Varga.
Systemic sclerosis.
Nature reviews Disease primers, 1(1):1–21, 2015.

[91] Lixian Zhong, Melinda Pope, Ye Shen, Jose J Hernandez, and Lin Wu.
Prevalence and incidence of systemic sclerosis: a systematic review and meta-analysis.
International journal of rheumatic diseases, 22(12):2096–2107, 2019.

[92] Maureen D Mayes, James V Lacey Jr, Jennifer Beebe-Dimmer, Brenda W Gillespie, Brenda Cooper, Timothy J Laing, and David Schottenfeld.
Prevalence, incidence, survival, and disease characteristics of systemic sclerosis in a large us population.
Arthritis & Rheumatism: Official Journal of the American College of Rheumatology, 48(8):2246–2255, 2003.

[93] Anna-Maria Hoffmann-Vold, Øyvind Midtvedt, Øyvind

Molberg, Torhild Garen, and Jan Tore Gran.
Prevalence of systemic sclerosis in south-east norway.
Rheumatology, 51(9):1600–1605, 2012.

[94] Y Stephanie Gu, James Kong, Gurtej S Cheema, Carl L Keen, Georg Wick, and M Eric Gershwin.
The immunobiology of systemic sclerosis.
In *Seminars in arthritis and rheumatism*, volume 38, pages 132–160. Elsevier, 2008.

[95] Gene-Siew Ngian, Joanne Sahhar, Susanna M Proudman, Wendy Stevens, Ian P Wicks, and Sharon Van Doornum.
Prevalence of coronary heart disease and cardiovascular risk factors in a national cross-sectional cohort study of systemic sclerosis.
Annals of the rheumatic diseases, 71(12):1980–1983, 2012.

[96] Michael Hughes, Elisabetta Zanatta, Robert D Sandler, Jérôme Avouac, and Yannick Allanore.

Improvement with time of vascular outcomes in systemic sclerosis: a systematic review and meta-analysis study.
Rheumatology, 61(7):2755–2769, 2022.

[97] Francesca Ingegnoli, Nicola Ughi, and Carina Mihai.
Update on the epidemiology, risk factors, and disease outcomes of systemic sclerosis.
Best Practice & Research Clinical Rheumatology, 32(2):223–240, 2018.

[98] Isabelle Marie.
Systemic sclerosis and exposure to heavy metals.
Autoimmunity Reviews, 18(1):62–72, 2019.

[99] Robert L Cowie.
Silica-dust-exposed mine workers with scleroderma (systemic sclerosis).
Chest, 92(2):260–262, 1987.

[100] Gregory R Owens, Gregory J Fino, David L Herbert, Virginia D Steen, Thomas A Medsger Jr, Bernard E Pennock, Joseph J Cottrell, Gerald P Rodnan, and Robert M Rogers.

Pulmonary function in progressive systemic sclerosis: comparison of crest syndrome variant with diffuse scleroderma.
Chest, 84(5):546–550, 1983.

[101] ANITA ÅKESSON and FA Wollheim.
Organ manifestations in 100 patients with progressive systemic sclerosis: a comparison between the crest syndrome and diffuse scleroderma.
Rheumatology, 28(4):281–286, 1989.

[102] Coleen Bertsch.
Crest syndrome: a variant of systemic sclerosis.
Orthopedic Nursing, 14(2):53–60, 1995.

[103] Edward E Velayos, Alfonse T Masi, Mary Betty Stevens, and Lawrence E Shulman.
The'crest'syndrome: comparison with systemic sclerosis (scleroderma).
Archives of internal medicine, 139(11):1240–1244, 1979.

[104] Alexandra Frolkis, Levinus A Dieleman, Herman W Barkema, Remo Panaccione, Subrata Ghosh,

Richard N Fedorak, Karen Madsen, Gilaad G Kaplan, Alberta IBD Consortium, et al.
Environment and the inflammatory bowel diseases.
Canadian Journal of Gastroenterology and Hepatology, 27:e18–e24, 2013.

[105] Yoichi Kakuta, Takeo Naito, Yoshitaka Kinouchi, and Atsushi Masamune.
Current status and future prospects of inflammatory bowel disease genetics.
Digestion, 104(1):7–15, 2023.

[106] Jean-Paul Achkar and Richard Duerr.
The expanding universe of inflammatory bowel disease genetics.
Current opinion in gastroenterology, 24(4):429–434, 2008.

[107] Miquel A Gassull and Eduard Cabré.
Nutrition in inflammatory bowel disease.
Current Opinion in Clinical Nutrition & Metabolic Care, 4(6):561–569, 2001.

[108] Daniel J Mulder, Angela J Noble, Christopher J Justinich, and Jacalyn M Duffin.
A tale of two diseases: the history of inflammatory bowel disease.
Journal of Crohn's and Colitis, 8(5):341–348, 2014.

[109] Eric A Franzosa, Alexandra Sirota-Madi, Julian Avila-Pacheco, Nadine Fornelos, Henry J Haiser, Stefan Reinker, Tommi Vatanen, A Brantley Hall, Himel Mallick, Lauren J McIver, et al.
Gut microbiome structure and metabolic activity in inflammatory bowel disease.
Nature microbiology, 4(2):293–305, 2019.

[110] Natalie A Molodecky and Gilaad G Kaplan.
Environmental risk factors for inflammatory bowel disease.
Gastroenterology & hepatology, 6(5):339, 2010.

[111] Anand B Pithadia and Sunita Jain.
Treatment of inflammatory bowel disease (ibd).
Pharmacological Reports, 63(3):629–642, 2011.

[112] Henry J Binder.
Mechanisms of diarrhea in inflammatory bowel diseases.
Annals of the New York Academy of Sciences, 1165(1):285–293, 2009.

[113] Seyed Saeid Seyedian, Forogh Nokhostin, and Mehrdad Dargahi Malamir.
A review of the diagnosis, prevention, and treatment methods of inflammatory bowel disease.
Journal of medicine and life, 12(2):113, 2019.

[114] Yi-Zhen Zhang and Yong-Yu Li.
Inflammatory bowel disease: pathogenesis.
World journal of gastroenterology: WJG, 20(1):91, 2014.

[115] Ta-Chiang Liu and Thaddeus S Stappenbeck.
Genetics and pathogenesis of inflammatory bowel disease.
Annual Review of Pathology: Mechanisms of Disease, 11:127–148, 2016.

[116] Dermot PB McGovern, Subra Kugathasan, and Judy H Cho.

Genetics of inflammatory bowel diseases.

Gastroenterology, 149(5):1163–1176, 2015.

[117] Ryohei Eguchi, Mohammand Bozlul Karim, Pingzhao Hu, Tetsuo Sato, Naoaki Ono, Shigehiko Kanaya, and Md Altaf-Ul-Amin.

An integrative network-based approach to identify novel disease genes and pathways: a case study in the context of inflammatory bowel disease.

BMC bioinformatics, 19(1):1–12, 2018.

[118] T Ahmad, J Satsangi, D McGovern, M Bunce, and DP Jewell.

The genetics of inflammatory bowel disease.

Alimentary pharmacology & therapeutics, 15(6):731–748, 2001.

[119] Rebecca Kuang, Stephen JD O'Keefe, Claudia Ramos del Aguila de Rivers, Filippos Koutroumpakis, and David G Binion.

Is salt at fault? dietary salt consumption and inflammatory bowel disease.

Inflammatory Bowel Diseases, 29(1):140–150, 2023.

[120] Laurent Peyrin-Biroulet, Marc Ferrante, Fernando Magro, Simon Campbell, Denis Franchimont, Herma Fidder, Hans Strid, Sandro Ardizzone, Gigi Veereman-Wauters, Jean-Baptiste Chevaux, et al.
Results from the 2nd scientific workshop of the ecco (i): Impact of mucosal healing on the course of inflammatory bowel disease.
Journal of Crohn's and Colitis, 5(5):477–483, 2011.

[121] Alexander Swidsinski, Axel Ladhoff, Annelie Pernthaler, Sonja Swidsinski, Vera Loening-Baucke, Marianne Ortner, Jutta Weber, Uwe Hoffmann, Stefan Schreiber, Manfred Dietel, et al.
Mucosal flora in inflammatory bowel disease.
Gastroenterology, 122(1):44–54, 2002.

[122] Richard S Blumberg, Lawrence J Saubermann, and Warren Strober.
Animal models of mucosal inflammation and their relation to human inflammatory bowel disease.

Current opinion in immunology, 11(6):648–656, 1999.

[123] Robert C Langan, Patricia B Gotsch, Michael A Krafczyk, and David D Skillinge.
Ulcerative colitis: diagnosis and treatment.
American family physician, 76(9):1323–1330, 2007.

[124] Derek P Jewell, Lloyd R Sutherland, John WD McDonald, and Brian G Feagan.
Ulcerative colitis.
Evidence-based Gastroenterology and Hepatology, pages 232–247, 2010.

[125] Mahesh Gajendran, Priyadarshini Loganathan, Guillermo Jimenez, Anthony P Catinella, Nathaniel Ng, Chandraprakash Umapathy, Nathalie Ziade, and Jana G Hashash.
A comprehensive review and update on ulcerative colitis.
Disease-a-month, 65(12):100851, 2019.

[126] Warren Strober, Ivan Fuss, Peter Mannon, et al.
The fundamental basis of inflammatory bowel disease.

The Journal of clinical investigation, 117(3):514–521, 2007.

[127] Daniel C Baumgart and William J Sandborn.
Crohn's disease.
The Lancet, 380(9853):1590–1605, 2012.

[128] Fergus Shanahan.
Crohn's disease.
The Lancet, 359(9300):62–69, 2002.

[129] Bernard Khor, Agnes Gardet, and Ramnik J Xavier.
Genetics and pathogenesis of inflammatory bowel disease.
Nature, 474(7351):307–317, 2011.

[130] Richard H Duerr.
Update on the genetics of inflammatory bowel disease.
Journal of clinical gastroenterology, 37(5):358–367, 2003.

[131] Pádraig Cunningham, Matthieu Cord, and Sarah Jane Delany.
Supervised learning.

In *Machine learning techniques for multimedia: case studies on organization and retrieval*, pages 21–49. Springer, 2008.

[132] Rich Caruana and Alexandru Niculescu-Mizil.
An empirical comparison of supervised learning algorithms.
In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168, 2006.

[133] Bing Liu and Bing Liu.
Supervised learning.
Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, pages 63–132, 2011.

[134] Xiaojin Jerry Zhu.
Semi-supervised learning literature survey.
2005.

[135] Tracy Coelho, Enrico Mossotto, Yifang Gao, Rachel Haggarty, James J Ashton, Akshay Batra, Imogen S Stafford, Robert M Beattie, Anthony P Williams, and Sarah Ennis.

Immunological profiling of paediatric inflammatory bowel disease using unsupervised machine learning.
Journal of Pediatric Gastroenterology and Nutrition, 70(6):833–840, 2020.

[136] Horace B Barlow.
Unsupervised learning.
Neural computation, 1(3):295–311, 1989.

[137] Zoubin Ghahramani.
Unsupervised learning.
In *Summer school on machine learning*, pages 72–112. Springer, 2003.

[138] Peter Dayan, Maneesh Sahani, and Grégoire Deback.
Unsupervised learning.
The MIT encyclopedia of the cognitive sciences, pages 857–859, 1999.

[139] G Peter Zhang and Min Qi.
Neural network forecasting for seasonal and trend time series.

European journal of operational research, 160(2):501–514, 2005.

[140] Mehdi Khashei and Mehdi Bijari.
An artificial neural network (p, d, q) model for timeseries forecasting.
Expert Systems with applications, 37(1):479–489, 2010.

[141] Guoqiang Zhang, B Eddy Patuwo, and Michael Y Hu.
Forecasting with artificial neural networks:: The state of the art.
International journal of forecasting, 14(1):35–62, 1998.

[142] Stephan Dreiseitl and Lucila Ohno-Machado.
Logistic regression and artificial neural network classification models: a methodology review.
Journal of biomedical informatics, 35(5-6):352–359, 2002.

[143] Taskin Kavzoglu.
Increasing the accuracy of neural network classification using refined training data.
Environmental Modelling & Software, 24(7):850–858, 2009.

[144] K-L Du.

Clustering: A neural network approach.

Neural networks, 23(1):89–107, 2010.

[145] Paul Mangiameli, Shaw K Chen, and David West.

A comparison of som neural network and hierarchical clustering methods.

European Journal of Operational Research, 93(2):402–417, 1996.

[146] Shirin Shadmand and Behbood Mashoufi.

A new personalized ecg signal classification algorithm using block-based neural network and particle swarm optimization.

Biomedical Signal Processing and Control, 25:12–23, 2016.

[147] JP Kelwade and SS Salankar.

Radial basis function neural network for prediction of cardiac arrhythmias based on heart rate time series.

In *2016 IEEE first international conference on control,* *measurement and instrumentation (CMI)*, pages 454–458. IEEE, 2016.

[148] Sami Saalasti.
Neural networks for heart rate time series analysis.
Number 33. Jyväskylän yliopisto, 2003.

[149] Kin C Luk, James E Ball, and Ashish Sharma.
An application of artificial neural networks for rainfall forecasting.
Mathematical and Computer modelling, 33(6-7):683–693, 2001.

[150] Mark N French, Witold F Krajewski, and Robert R Cuykendall.
Rainfall forecasting in space and time using a neural network.
Journal of hydrology, 137(1-4):1–31, 1992.

[151] Ivan Nunes Da Silva, Danilo Hernane Spatti, Rogerio Andrade Flauzino, Luisa Helena Bartocci Liboni, Silas Franco dos Reis Alves, Ivan Nunes da Silva, Danilo Hernane Spatti, Rogerio Andrade Flauzino,

Luisa Helena Bartocci Liboni, and Silas Franco dos Reis Alves.
Artificial neural network architectures and training processes.
Springer, 2017.

[152] Masoud Yaghini, Mohammad M Khoshraftar, and Mehdi Fallahi.
A hybrid algorithm for artificial neural network training.
Engineering Applications of Artificial Intelligence, 26(1):293–301, 2013.

[153] Morteza Pakdaman, Ali Ahmadian, Sohrab Effati, Soheil Salahshour, and Dumitru Baleanu.
Solving differential equations of fractional order using an optimization technique based on training artificial neural network.
Applied Mathematics and Computation, 293:81–95, 2017.

[154] Almir Badnjević, Lejla Gurbeta, Mario Cifrek, and Damir Marjanovic.
Classification of asthma using artificial neural network.

In *2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 387–390. IEEE, 2016.

[155] A Yasar, I Saritas, MA Sahman, and AC Cinar.
Classification of parkinson disease data with artificial neural networks.
In *IOP conference series: materials science and engineering*, volume 675, page 012031. IOP Publishing, 2019.

[156] Zhenhua Yu, Ayesha Sohail, Taher A Nofal, and João Manuel RS Tavares.
Explainability of neural network clustering in interpreting the covid-19 emergency data.
Fractals, 30(05):2240122, 2022.

[157] Wei-Te Huang, Hao-Hsiu Hung, Yi-Wei Kao, Shi-Chen Ou, Yu-Chuan Lin, Wei-Zen Cheng, Zi-Rong Yen, Jian Li, Mingchih Chen, Ben-Chang Shia, et al.
Application of neural network and cluster analyses to differentiate tcm patterns in patients with breast cancer.
Frontiers in Pharmacology, 11:670, 2020.

[158] Derek Greene, Pádraig Cunningham, and Rudolf Mayer.
Unsupervised learning and clustering.
Machine learning techniques for multimedia: Case studies on organization and retrieval, pages 51–90, 2008.

[159] Ramadass Sathya, Annamma Abraham, et al.
Comparison of supervised and unsupervised learning algorithms for pattern classification.
International Journal of Advanced Research in Artificial Intelligence, 2(2):34–38, 2013.

[160] Michael W Berry, Azlinah Mohamed, and Bee Wah Yap.
Supervised and unsupervised learning for data science.
Springer, 2019.

[161] K Velten, R Reinicke, and K Friedrich.
Wear volume prediction with artificial neural networks.
Tribology International, 33(10):731–736, 2000.

[162] Neha Sharma, Vibhor Jain, and Anju Mishra.
An analysis of convolutional neural networks for image classification.
Procedia computer science, 132:377–384, 2018.

[163] Akbar K Waljee, Kay Sauder, Anand Patel, Sandeep Segar, Boang Liu, Yiwei Zhang, Ji Zhu, Ryan W Stidham, Ulysses Balis, and Peter DR Higgins.
Machine learning algorithms for objective remission and clinical outcomes with thiopurines.
Journal of Crohn's and Colitis, 11(7):801–810, 2017.

[164] Richard M Zur, Yulei Jiang, Lorenzo L Pesce, and Karen Drukker.
Noise injection for training artificial neural networks: A comparison with weight decay and early stopping.
Medical physics, 36(10):4810–4818, 2009.

[165] Hongjun Lu, Rudy Setiono, and Huan Liu.
Effective data mining using neural networks.
IEEE transactions on knowledge and data engineering, 8(6):957–961, 1996.

[166] Torgyn Shaikhina and Natalia A Khovanova.
Handling limited datasets with neural networks in medical applications: A small-data approach.
Artificial intelligence in medicine, 75:51–63, 2017.

[167] Derek A Pisner and David M Schnyer.
Support vector machine.
In *Machine learning*, pages 101–121. Elsevier, 2020.

[168] Jair Cervantes, Farid Garcia-Lamont, Lisbeth Rodríguez-Mazahua, and Asdrubal Lopez.
A comprehensive survey on support vector machine classification: Applications, challenges and trends.
Neurocomputing, 408:189–215, 2020.

[169] Johannes Smolander, Matthias Dehmer, and Frank Emmert-Streib.
Comparing deep belief networks with support vector machines for classifying gene expression data from complex disorders.
FEBS Open Bio, 9(7):1232–1248, 2019.

[170] V Rodriguez-Galiano, M Sanchez-Castillo, M Chica-Olmo, and MJOGR Chica-Rivas.
Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines.

Ore Geology Reviews, 71:804–818, 2015.

[171] Taylor R Lee, Warren T Wood, and Benjamin J Phrampus.
A machine learning (knn) approach to predicting global seafloor total organic carbon.
Global Biogeochemical Cycles, 33(1):37–46, 2019.

[172] Shichao Zhang, Xuelong Li, Ming Zong, Xiaofeng Zhu, and Debo Cheng.
Learning k for knn classification.
ACM Transactions on Intelligent Systems and Technology (TIST), 8(3):1–19, 2017.

[173] Shruti Jain and Ayodeji Olalekan Salau.
An image feature selection approach for dimensionality reduction based on knn and svm for akt proteins.
Cogent Engineering, 6(1):1599537, 2019.

[174] Berndt Müller, Joachim Reinhardt, and Michael T Strickland.
Neural networks: an introduction.
Springer Science & Business Media, 1995.

[175] James A Anderson.

An introduction to neural networks.
MIT press, 1995.

[176] W Thomas Miller, Richard S Sutton, and Paul J Werbos.
Neural networks for control.
MIT press, 1995.

[177] Rebecca Roelofs, Vaishaal Shankar, Benjamin Recht, Sara Fridovich-Keil, Moritz Hardt, John Miller, and Ludwig Schmidt.
A meta-analysis of overfitting in machine learning.
Advances in Neural Information Processing Systems, 32, 2019.

[178] Yaohao Peng and Mateus Hiro Nagata.
An empirical overview of nonlinearity and overfitting in machine learning using covid-19 data.
Chaos, Solitons & Fractals, 139:110055, 2020.

[179] Ali Akbar Movassagh, Jafar A Alzubi, Mehdi Gheisari, Mohamadtaghi Rahimi, Senthilkumar Mohan, Aaqif Afzaal Abbasi, and Narjes Nabipour.

Artificial neural networks training algorithm integrating invasive weed optimization with differential evolutionary model.

Journal of Ambient Intelligence and Humanized Computing, pages 1–9, 2021.

[180] Nicolaos Karayiannis and Anastasios N Venetsanopoulos.

Artificial neural networks: learning algorithms, performance evaluation, and applications, volume 209.

Springer Science & Business Media, 1992.

[181] David Meyer and FT Wien.

Support vector machines.

The Interface to libsvm in package e1071, 28(20):597, 2015.

[182] Vojislav Kecman.

Support vector machines–an introduction.

In *Support vector machines: theory and applications*, pages 1–47. Springer, 2005.

[183] Chih-Chung Chang and Chih-Jen Lin.

Libsvm: a library for support vector machines.

ACM transactions on intelligent systems and technology (TIST), 2(3):1–27, 2011.

[184] Pardis Birzhandi, Kyung Tae Kim, Byungjun Lee, and Hee Yong Youn.
Reduction of training data using parallel hyperplane for support vector machine.
Applied Artificial Intelligence, 33(6):497–516, 2019.

[185] William S Noble.
What is a support vector machine?
Nature biotechnology, 24(12):1565–1567, 2006.

[186] Haneen Arafat Abu Alfeilat, Ahmad BA Hassanat, Omar Lasassmeh, Ahmad S Tarawneh, Mahmoud Bashir Alhasanat, Hamzeh S Eyal Salman, and VB Surya Prasath.
Effects of distance measure choice on k-nearest neighbor classifier performance: a review.
Big data, 7(4):221–248, 2019.

[187] Balaji Rajagopalan and Upmanu Lall.

A k-nearest-neighbor simulator for daily precipitation and other weather variables.

Water resources research, 35(10):3089–3101, 1999.

[188] Yihua Liao and V Rao Vemuri.

Use of k-nearest neighbor classifier for intrusion detection.

Computers & security, 21(5):439–448, 2002.

[189] Archana Singh, Avantika Yadav, and Ajay Rana.

K-means with three different distance metrics.

International Journal of Computer Applications, 67(10), 2013.

[190] Hang Xu, Wenhua Zeng, Xiangxiang Zeng, and Gary G Yen.

An evolutionary algorithm based on minkowski distance for many-objective optimization.

IEEE transactions on cybernetics, 49(11):3968–3979, 2018.

[191] Patrick JF Groenen and Krzysztof Jajuga.

Fuzzy clustering with squared minkowski distances.

Fuzzy Sets and Systems, 120(2):227–237, 2001.

[192] Qingqing Zhai, Jun Yang, Min Xie, and Yu Zhao.
Generalized moment-independent importance measures based on minkowski distance.
European Journal of Operational Research, 239(2):449–455, 2014.

[193] Nigel Williams, Sebastian Zander, and Grenville Armitage.
A preliminary performance comparison of five machine learning algorithms for practical ip traffic flow classification.
ACM SIGCOMM Computer Communication Review, 36(5):5–16, 2006.

[194] Kamran Shaukat, Suhuai Luo, Vijay Varadharajan, Ibrahim A Hameed, Shan Chen, Dongxi Liu, and Jiaming Li.
Performance comparison and current challenges of using machine learning techniques in cybersecurity.
Energies, 13(10):2509, 2020.

[195] Iftikhar Ahmad, Mohammad Basheri, Muhammad Javed Iqbal, and Aneel Rahim.

Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection.

IEEE access, 6:33789–33795, 2018.

[196] https://www.mathworks.com/products/matlab.html.

Matlab.

Accessed on July 2023.

[197] https://www.python.org/.

Python.

Accessed on July 2023.

[198] https://www.r project.org/.

R.

Accessed on July 2023.

[199] https://www.kofo.mpg.de/en/research/services/orca.

Orca.

Accessed on July 2023.

[200] https://gaussian.com/.

Gaussian.

Accessed on July 2023.

[201] https://gaussian.com/gaussview6/.

Gauss view.

Accessed on July 2023.

[202] https://pymol.org/2/.

Pymol.

Accessed on July 2023.

[203] https://pypi.org/project/avogadro/.

Avogadro.

Accessed on July 2023.

[204] https://www.nwchem sw.org/.

Nwchem.

Accessed on July 2023.

[205] https://usegalaxy.org/.

Galaxy.

Accessed on July 2023.

[206] https://ambermd.org/AmberTools.php.

Ambertools.

Accessed on July 2023.

[207] https://blast.ncbi.nlm.nih.gov/Blast.cgi.

Blast.

Accessed on July 2023.

[208] https://megasoftware.net/.

Mega.

Accessed on July 2023.

[209] https://openbabel.org/wiki.

Open babel.

Accessed on July 2023.

[210] https://chemistry.artsandsciences.baylor.edu/research/chemdraw.

Chemdrawl.

Accessed on July 2023.

[211] https://www.cp2k.org/.

Cp2k.

Accessed on July 2023.

[212] https://genome.ucsc.edu/.

Usc genome browser.
Accessed on July 2023.

[213] CZ Cai, LY Han, Zhi Liang Ji, X Chen, and Yu Zong Chen.
Svm-prot: web-based support vector machine software for functional classification of a protein from its primary sequence.
Nucleic acids research, 31(13):3692–3697, 2003.

[214] Rachel Karchin, Kevin Karplus, and David Haussler.
Classifying g-protein coupled receptors with support vector machines.
Bioinformatics, 18(1):147–159, 2002.

[215] Jesus K Yamamoto-Furusho.
Genetic factors associated with the development of inflammatory bowel disease.
World Journal of Gastroenterology: WJG, 13(42):5594, 2007.