mathematics

MDPI

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

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

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

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.

$$X_i = \begin{Bmatrix} X_i^1 \\ X_i^2 \\ X_i^3 \\ . \\ . \\ X_i^n \end{Bmatrix} \tag{1}$$

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

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

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

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.

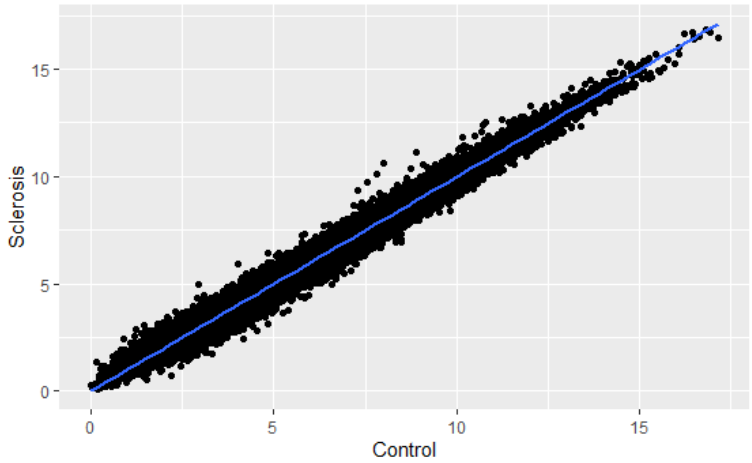

Figure 1. Gene expression in SSc-ILD vs. control patients.

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.

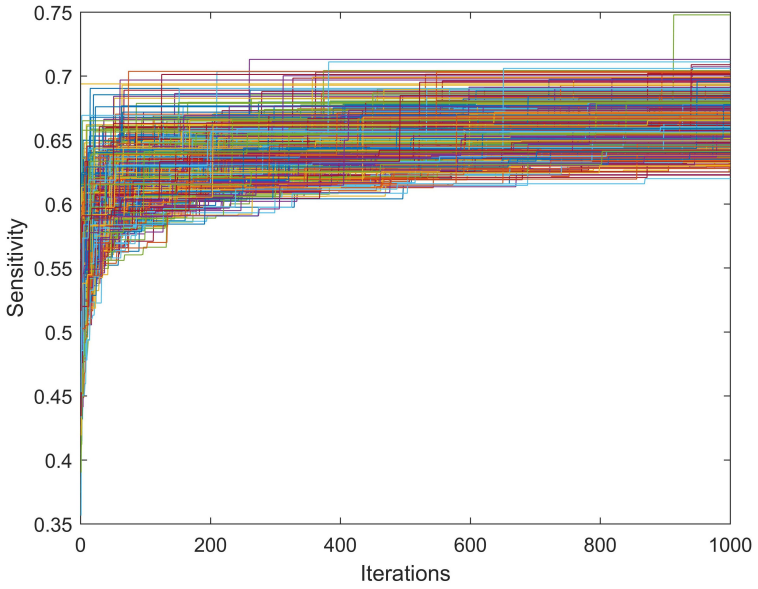

Figure 2. Sensitivity results of the models.

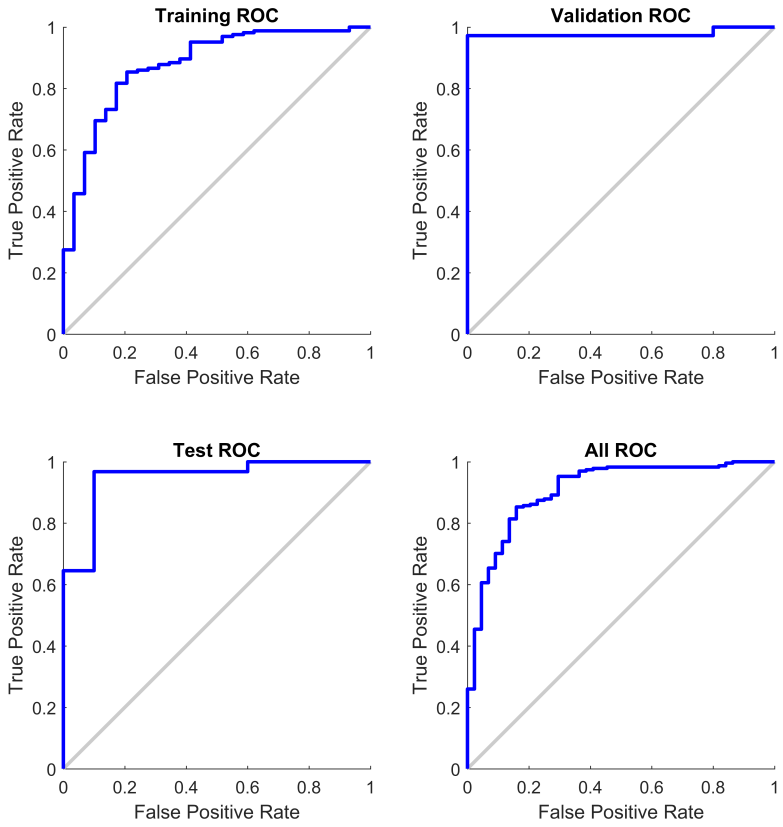

Figure 3. ROC sample for one iteration.

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

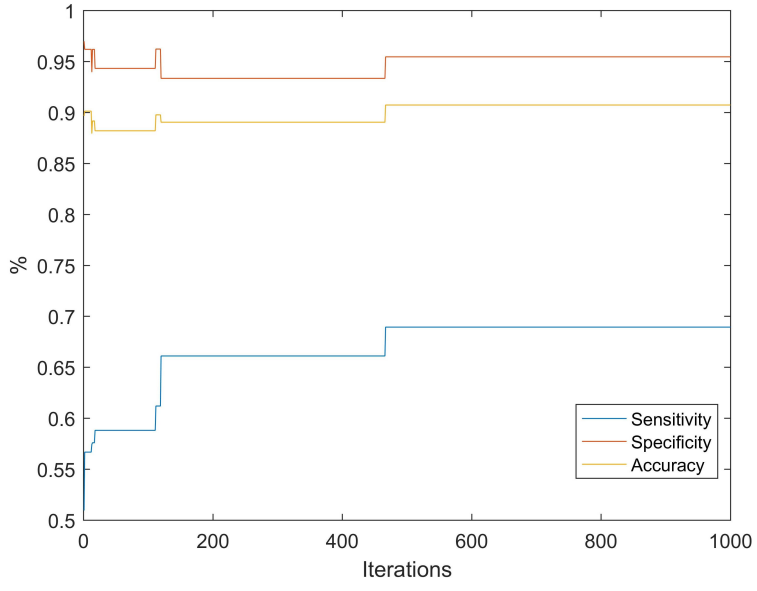

Figure 4. Sensitivity, specificity and accuracy of a sample model.

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.

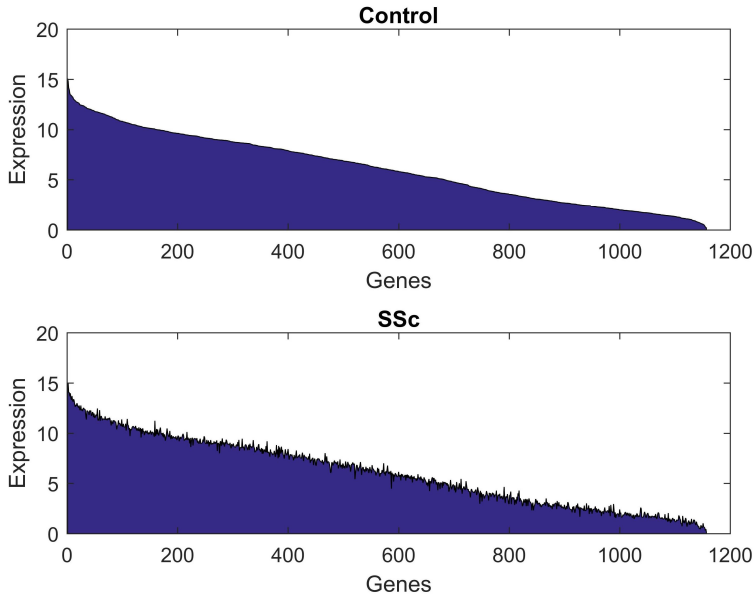

Figure 5. Mean gene expression for controls and patients.

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.

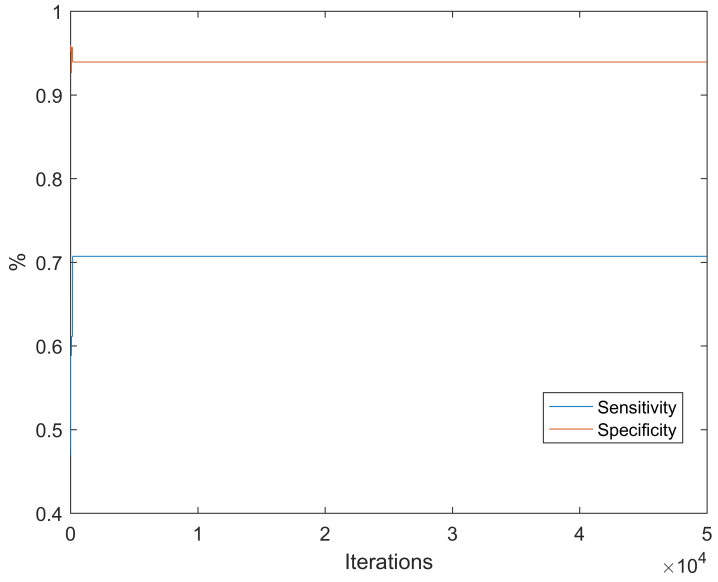

Figure 6. Sample of asymptotic analysis (50,000 iterations).

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]