Supporting Medical Decisions for Treating Rare Diseases Through Genetic Programming

Illya Bakurov[1(✉)], Mauro Castelli[1], Leonardo Vanneschi[1], and Maria João Freitas[2]

[1] Nova Information Management School (NOVA IMS),
Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisbon, Portugal
{ibakurov,mcastelli,lvanneschi}@novaims.unl.pt

[2] Raríssimas - Associação Nacional de Deficiências Mentais e Raras,
Rua das Açucenas, Lote 1, Loja Dta., 1300-003 Lisbon, Portugal
mjoao.freitas@rarissimas.pt

Abstract. Casa dos Marcos is the largest specialized medical and residential center for rare diseases in the Iberian Peninsula. The large number of patients and the uniqueness of their diseases demand a considerable amount of diverse and highly personalized therapies, that are nowadays largely managed manually. This paper aims at catering for the emergent need of efficient and effective artificial intelligence systems for the support of the everyday activities of centers like Casa dos Marcos. We present six predictive data models developed with a genetic programming based system which, integrated into a web-application, enabled data-driven support for the therapists in Casa dos Marcos. The presented results clearly indicate the usefulness of the system in assisting complex therapeutic procedures for children suffering from rare diseases.

Keywords: Genetic Programming · Geometric Semantic Genetic Programming · Medical decisions · Rare diseases

1 Introduction

The term *rare disease* is used to identify any disease that affects a tiny percentage of the population. From a regulatory perspective, rare diseases are defined as those diseases where less than 200,000 persons are affected in the USA or no more than one person over 2,000 is affected in the general population of the European Union (EU). Such diseases usually have a genetic basis, often affecting patients early in childhood or even since birth, and are frequently progressive, disabling and life threatening. Nowadays approximately 7,000 different rare diseases have been identified, and the number of people suffering from a rare disease in the USA and EU exceeds 55 million, highlighting the enormous social impact of these diseases. For all these reasons, in the last two decades there has been a substantial

P. Kaufmann and P. A. Castillo (Eds.): EvoApplications 2019, LNCS 11454, pp. 187–203, 2019.
https://doi.org/10.1007/978-3-030-16692-2_13

increase worldwide in the number of specialized medical and therapeutic centers for the care and treatment of patients affected by these diseases [1]. In Portugal there has been a development in the treatment and care of this type of patients, in particular after the foundation of the Portuguese Association of Mental and Rare Diseases in 2002. This association, called *Raríssimas*, is a non-profit organization whose mission is to support people affected by rare diseases and their relatives. In 2010, *Raríssimas*, with substantial contribution of private capitals and donations, gave birth to *Casa dos Marcos*, a medical and residential center. *Casa dos Marcos* is a highly specialized center, with a clinic with the capacity to receive 5,000 patients per year and a Physical Medicine and Rehabilitation unit.

Although extremely active, *Casa dos Marcos* is presently facing several challenges, many of which are shared by the analogous medical centers worldwide: (1) rare diseases are, by their very nature, diverse among each other and thus *unique*, which makes personalized therapies a must; (2) all the actions related to the planning of the therapies, the hospitalization and recovery of the patients and the organization of their everyday life in the medical center are, in large part, performed manually. These issues naturally demand for accurate and efficient computational systems. Many of the activities of the center, in fact, demand for efficient and effective predictive models able to support medical decisions.

The objective of this paper is to contribute towards the achievement of such an ambitious goal, using the development of a Machine Learning (ML) system able to generate six predictive models to forecast the effect of a specialized therapy on one global and five local factors. These models were integrated into a specially designed web-application to support the decision-making processes of *Casa dos Marcos*. The system which generated the models is based on Genetic Programming (GP), and more in particular on one of the newest developments of GP: The Evolutionary Demes Despeciation Algorithm (EDDA). GP holds tremendous potential for this type of application, for at least the following reasons. First, it has the potentiality of generating highly non-linear models of multiple features. Second, it can automatically perform feature selection during the learning phase. Additionally, GP produces models which enable subsequent interpretation and feature importance analysis. To the best of our knowledge, the one presented in this paper is the first GP-based system developed in the context of Rare Diseases.

The document is organized as follows. Section 2 introduces the reader to the context of this study. Section 3 presents the research track regarding practical applications of state-of-the-art ML tools in the field of medicine. Section 4 describes the methodological approach. Section 5 presents the experimental settings and discusses the results obtained. Finally, Sect. 6, summarizes the main findings of this work and provides some suggestions for future work.

2 Problem Description and Data

2.1 The Pediasuit Protocol

Pediasuit is a modern, intensive therapy suit. It is inspired by the *Penguin suit*, developed by the Soviet space program to neutralize the harmful effects of weightlessness and hypokinesis on the body of astronauts during space flights [2]. It is currently used by Casa dos Marcos during some therapeutic procedures. The *Pediasuit Protocol* is a therapeutic approach which uses *Pediasuit*, initially designed for people who happen to have neurological disorders, such as cerebral palsy, developmental delay, autism and other conditions that affect motor development and/or cognitive functions [2]. With the help of tension elastics, than can create an almost anti-gravity effect, patients are able to perform movements that would otherwise seem impossible and will be able to train their muscles and posture in a very specific manner. In this way, the therapy helps minimize pathological reflexes and promote the establishment of new, correct and functional movements.

The Pediasuit Protocol is one of the therapeutic treatments performed by Casa dos Marcos to assist patients, mostly children, with rare, predominantly neurodegenerative, diseases. Concretely, the therapists' goal consists of achieving an improvement, or, at least, a situation of freezing, regarding motor and mental functioning, taking into account the progressive degeneration of nerve cells. The Pediasuit Protocol lasts for four weeks, with daily sessions of two or four hours each. Moreover, since it requires the assistance of highly specialized physiotherapists, the price ranges from 1300 euros to 2500 euros, according to the number of hours per session [3].

Gross Motor Function Measure. The *Gross Motor Function Measure* (GMFM) is a clinical tool designed to evaluate changes in gross motor function in patients (usually children) with cerebral palsy, traditionally using 88 measures (GMFM-88) [4]. The measures of the GMFM-88 standard span a large set of motor activities, each with its own summary-measures (from now on called *factors*): lying and rolling (17 measures), sitting (20 measures), crawling and kneeling (14 measures), standing (13 measures) and walking, running and jumping (24 measures). For each factor, therapists ask patients to perform a set of specific movements and exercises (the measures), which accuracy is then evaluated on an ordinal scale ranging from 0 to 3, where 0 means “does not initiate”, 1 means “initiates”, 2 means “partially completes”, and 3 means “completes”. At the end, global assessment measures are calculated: one total and five local scores (one for each factor). The GMFM international standard was adopted by Casa dos Marcos to measure the improvement, in terms of motor functionality, of patients who attended the Pediasuit Protocol. Each patient is evaluated through GMGM-88 twice: one before the start of the therapy and another after its conclusion.

2.2 The Data

Casa dos Marcos provided one dataset, with records of 27 different patients undergoing the therapy. Through data pre-processing procedures, one patient was removed from the analysis because of the high amount of missing values. Each patient underwent the therapy at least once and, for each therapy, two examinations under the GMFM standard were performed. Besides the GMFM measures and six summary indicators, the dataset also contains three socio-demographic attributes: gender, birth date and diagnosis. In total, considering other operational attributes, 108 variables were provided.

Given a dataset with the information related to only 26 patients, developing predictive models with more than 100 attributes is not an easy task. Given this difficulty, to increase the size of the training data, we decided to consider one training instance per therapy. As a result, 41 data instances became available for the analysis. At first glance, it seems inaccurate to consider different therapies of the same patient as independent training data instances. However, given the problem specificity, this decision has two motivational arguments. First, most of the patients who have undergone the therapy present neurodegenerative diseases. Given this fact, a unique and unpredictable deviation is expected, in terms of GMFM, between subsequent therapies taken by the same patient. Second, the same patient is expected to react differently to equivalent therapies performed in different time periods, due to the disease progression status. Furthermore, the GMFM of such a patient, after one year, is expected to vary in a negative fashion.

3 Related Work

This section presents a selection of previous studies where GP was used to tackle challenging problems in the field of health-care and medicine. One of the first studies appeared in [5], where a constrained-syntax GP-based algorithm for discovering classification rules in medical data sets was proposed. To address that problem, the authors defined a GP framework containing several syntactic constraints to be enforced by the system using a disjunctive standard form representation, so that individuals represent valid rule sets that are easy to interpret. GP was compared against a decision-tree-building algorithm over five medical data sets, and it was able to obtain good results with respect to predictive accuracy and rule comprehensibility, by comparison with decision trees.

Another study where GP was used to solve a problem in the field of medicine was proposed in [6], where the authors tackled a problem related to the physiochemical properties of proteins. More in detail, the problem addressed in their work was the prediction of the physiochemical properties of proteins tertiary structure, with the objective of predicting the size of the residues considering the protein tertiary structure data. The authors employed a semantics-based GP framework to solve the problem successfully, and the system produced a superior performance with respect to other considered techniques like artificial neural networks and support vector machines.

In 2018, Ting et al. [7] used GP to analyze feature importance for metabolomics [8]. In their work the authors analyzed a population-based metabolomics dataset on osteoarthritis and developed a Linear GP (LGP) algorithm to search classification models that can best predict the disease outcome, as well as to identify the most important metabolic markers associated with the disease. The LGP algorithm produced satisfactory performance, also being able to identify a set of key metabolic markers that may be useful for achieving a better understanding of the biochemistry of the disease.

Other successful applications of GP and its variants are in the field of pharmacokinetics, where several contributions appeared in recent years [9–11]. In several of these contributions, one crucial aspect strengthened by the authors is the ability of GP in producing a human-understandable model, a fundamental feature in the field of medicine that also motivates the choice of GP as the elected ML method for solving the problem considered in this study.

The interested reader is referred to [12] for a recent overview of the main contribution of genetic and evolutionary computation in the medical field.

4 Methodology

4.1 Geometric Semantic Genetic Programming

In the current terminology adopted by a considerable part of the Genetic Programming (GP) [13] research community, the term *semantics* indicates the vector of output values of a solution, calculated on the training observations [14,15]. Under this perspective, a GP individual can be seen as a point in a multidimensional space (its semantics). This space, called *semantic space*, has a number of dimensions equal to the number of observations in the training set.

Geometric Semantic Genetic Programming (GSGP) [14] is a recently introduced variant of GP in which standard crossover and mutation are replaced by so-called *Geometric Semantic Operators* (GSOs). The former operators allow the algorithm to exploit semantic awareness and induce precise geometric properties on the semantic space. GSOs, introduced by Moraglio et al. [14], gained popularity in the GP community [15] because of their property of inducing a unimodal error surface (characterized by the absence of locally optimal solutions) for any supervised learning problem. The proof of this property can be found in [14].

Here, we report the definition of the GSOs, as given by Moraglio et al. for real functions domains, since these are the operators that will be used in the experimental phase. For applications that consider other types of data, the reader is referred to [14]. *Geometric Semantic Crossover* (GSC) generates, as the unique offspring of parents $T_1, T_2 : \mathbb{R}^n \rightarrow \mathbb{R}$, the expression: $T_{XO} = (T_1 \cdot T_R) + ((1 - T_R) \cdot T_2)$, where T_R is a random real function whose output values range in the interval $[0, 1]$. *Geometric Semantic Mutation* (GSM) returns, as the result of the mutation of an individual $T : \mathbb{R}^n \rightarrow \mathbb{R}$, the expression: $T_M = T + ms \cdot (T_{R1} - T_{R2})$, where T_{R1} and T_{R2} are random real functions with codomain in $[0, 1]$ and ms is a parameter called the mutation step.

This work considers the GSOs' implementation presented in [16].

A recognized drawback of GSGP consists of the potential weakness of GSC. Given that GSC generates an offspring whose semantics stands on the segment joining two points representing the parents (in the semantic space), it can only achieve the global optimum solution if the semantics of the individuals in the population "surround" the semantics of the global optimum. Using the terminology of [17,18], GSC only has the possibility of generating a globally optimal solution only if this solution lays within the semantic *convex hull* identified by the population. The need for overcoming this drawback has led to several methods to properly initialize a population of GSGP, like for instance the ones presented in [19–21].

4.2 Evolutionary Demes Despeciation Algorithm

Initialization is known to play a very important role for any population-based algorithm. The same happens in GP, where a wide variety of programs of various sizes and shapes are desirable [13]. With the introduction of GSOs, new techniques taking their particularities into consideration, have been developed [19]. The Evolutionary Demes Despeciation Algorithm (EDDA) is contextualized in this research track.

In Biology, demes are independent populations, or sub-populations, of individuals that actively interbreed and mature, and the term despeciation indicates the combination of demes of previously distinct species into a new population, where distinct biological lineage is blended. The despeciation phenomenon rarely occurs in Nature, but in some cases it is known to fortify populations. In EDDA, the initial population of GSGP is generated using the best individuals obtained from a set of independent sub-populations (demes), that evolved for few generations and under different evolutionary conditions: some demes use standard GP, while others use GSGP and each deme is being evolved under distinct search parameters [20]. EDDA was recently introduced in the GP community [22,23] and owes its success to its simplicity and wide scope of applications. Although EDDA was originally developed to take into consideration the particularities of GSGP, it can also be used to initialize any population-based algorithm.

GSGP using EDDA demonstrated its superiority over GSGP initialized with the traditional Ramped Half-and-Half (RHH) [13] method over six complex symbolic regression applications [20]. More specifically, on all problems, EDDA allowed for generation of solutions with comparable or even better generalization ability and of significantly smaller size than using RHH. The efficacy of EDDA depends on two main parameters: the proportion of GSGP demes in the system (n) and the number of generations to evolve each deme (m). Using an algorithm-specific notation, given two natural numbers n and m, where $n \in [0, 100]$, EDDA-n% represents a system where demes are left to evolve for m generations such that n% of the population was initialized using individuals from GSGP demes, while the remaining $(100 - n)$% was initialized using standard GP demes. The pseudo-code in Fig. 1 explains the process.

```
EDDA-$n$% (evolving demes for $m$ generations):

1. Create an empty population $P$ of size $N$;
2. Repeat $N * (n/100)$ times:
   (a) Create an empty deme;
   (b) Randomly initialize this deme using a classical initialization algorithm (RHH
       used here);
   (c) Evolve individuals from 2.b) for $m$ generations using GSGP;
   (d) After finishing 2.c), select the best individual from the deme and store it in $P$;
3. Repeat $N * (1 - n/100)$ times:
   (a) Create an empty deme;
   (b) Randomly initialize this deme using a classical initialization algorithm (RHH
       used here);
   (c) Evolve individuals from 3.b) for $m$ generations using standard GP;
   (d) After finishing 3.c), select the best individual from the deme and store it in $P$;
4. Retrieve $P$ and use it as the initial population of GSGP
```

Fig. 1. Pseudo-code of the EDDA-n% system, in which demes are left to evolve for m generations.

In the pseudo-code in Fig. 1, points 2(b), 2(c), 3(b) and 3(c) implement the *evolution of demes*, while points 2(d) and 3(d) implement the *despeciation* phase. In the former step, the different demes evolve independently; in the latter phase, individuals coming from different demes, and thus from different evolutionary dynamics and histories, are joined in a new population (P in the pseudo-code). To evolve this new population, GSGP is preferred over standard GP because in several application domains GSGP is known to outperform standard GP [24]. For this reason, in this study, after the *despeciation* phase, we used GSGP to conduct the main evolutionary process (MEP).

5 Experiments

Given that the therapy's impact can be assessed through six possible summary measures (factors), to predict it's impact on a given patient, six different supervised-learning models have to be created for each factor.

5.1 Experimental Settings

The training dataset for each one of the six supervised-learning problems consisted of 41 training data instances and 97 input features. The terminal set consisted of all input features and nine constants defined in $[-1,\ 1]$ as $\{-1.0,\ -0.75,\ -0.5,\ -0.25,\ 0.0,\ 0.25,\ 0.5,\ 0.75,\ 1.0\}$. The function set contained the primitive functions $\{+,\ -,\ *,\ /,\ sin,\ cos,\ \surd, ln\}$, where $\sqrt{x}$ and $ln(x)$ return x if its value does not fall within their respective domain. Similarly, a/x returns a if x equals to zero.

The fitness was calculated as the Root Mean Squared Error (RMSE) between predicted and expected outputs. Tournament selection with a selection pressure of 10% was used to select the parents. Similarly to [20,24], the probability of applying a given variation operator was randomly drawn at the beginning of each generation and the mutation step, in the case of geometric semantic mutation, was randomly generated, with uniform probability in [0, 1], at each mutation event. Survival was elitist, i.e., the best individual was copied unchanged into the next population at each generation.

For all the considered algorithm executions, populations of two-hundred individuals were used (both in the initialization and in the main algorithm). Consequently, since we used EDDA to initialize the population, there were two-hundred demes consisting of two-hundred individuals each. Given the restricted number of training data instances allied to the high dimensionality of the problem, each initial population contained a set of 97 individuals, each of which composed by one single terminal (one distinct input feature). This was done in order not to misuse potentially relevant features during the phase of random initialization of the demes. At each deme, tree initialization of the remaining 103 individuals was performed by using RHH, with maximum initial depth equal to 5. Individuals growth was not limited throughout the whole evolutionary process.

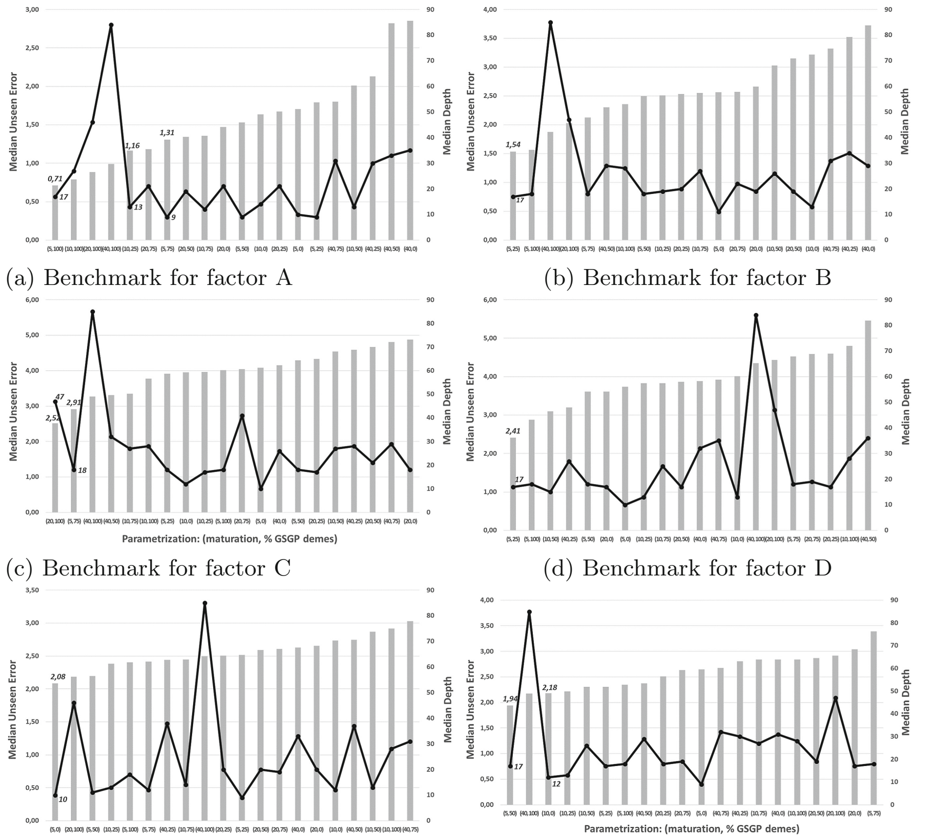
For each factor, EDDA was studied in 5 different configurations, as in [20]: EDDA-0, 25, 50, 75, 100%. For each on of these, we studied a version in which each deme was evolved for 5, 10, 20 and 40 generations. As such, 20 benchmarks were considered for each GMFM factor (target), where each benchmark consisted of a pair *(maturation, EDDA-n%)*. This totaled to 120 benchmarks to tune EDDA. For each benchmark, at the end of the evolution of each deme, the best individual (in terms of training error) was selected to seed the initial population of GSGP.

Given that there were only 41 data instances, Leave-One-Out (LOO) cross-validation was applied. More concretely, at the beginning of each run, one different data instance was left out to assess generalization, while the remaining $n-1$ were used to train the model. For this reason, 41 independent runs of EDDA were performed.

5.2 Experimental Results

Figure 2 exhibits all 120 benchmarks conducted to study the impact of maturation and proportion of the GSGP demes in EDDA for each target (GMFM factor). Each sub-figure summarizes all 41 runs of each benchmark, conducted for a given target, through a median validation error (gray bars with the left vertical axis) and median depth (black lines with the right vertical axis). Both summary statistics were obtained from the analysis of the best individual at the end of each run. Notice that in each sub-figure, the benchmarks were sorted in ascending order by median validation error.

The following paragraph provides a discussion of the results presented in Fig. 2. It is important to notice that our choice of the EDDA parameters was guided not only by validation error but also by the size of the individuals.



(a) Benchmark for factor A

(b) Benchmark for factor B

(c) Benchmark for factor C

(d) Benchmark for factor D

(e) Benchmark for factor E

(f) Benchmark for global factor

Fig. 2. Evolution of the median best validation error (gray bars measured through left vertical axis) and median depth (black lines measured through right vertical axis) for factors: A (a), B (b), C (c), D (d), E (e) and global factor (f).

Three arguments motivated this. First, Occam's razor, i.e., given models with similar training performance, the simplest (in our case, smallest) model should be preferred. Second, the need for delivering final individuals through a web-application, used by the therapists. Third, models interpretability is important in this application, mainly because the domain experts have to trust the models, and smaller models should be easier to interpret (even though we are aware that this may not always be true). Here are, the chosen parameters with a short motivation for each case reported in Fig. 2:

- sub-plot (a): it was decided to opt for the parameterization provided by benchmark number 7, i.e. (5, 75), due to its noticeably lower median depth compared to parameterizations with slightly better generalization ability;

- sub-plot (b): the first benchmark, i.e. (5, 25), because it exhibited the best combination of summary statistics, i.e., lowest median error and depth;
- sub-plot (c): the second benchmark, i.e. (5, 75), since it demonstrated a significantly lower median depth at almost no penalty in terms of generalization ability;
- sub-plot (d): the first benchmark, i.e. (5, 25), because it exhibited the best combination of the summary statistics.
- sub-plot (e): the first benchmark, i.e. (5, 0), because it demonstrated the best combination of summary statistics;
- sub-plot (f): the first benchmark, i.e. (5, 50), since it exhibited the best combination of summary statistics.

After selecting the initialization parameters, detailed analysis of the evolutionary process, succeeding EDDA initialization, was performed and its results are discussed in the continuation.

Figure 3 provides visualization of the evolutionary process conducted for each factor, for 50 generations. Each sub-figure summarizes 41 runs through the median training and validation error (gray and black solid lines on the left vertical axis) and median depth (dashed gray line on the right vertical axis).

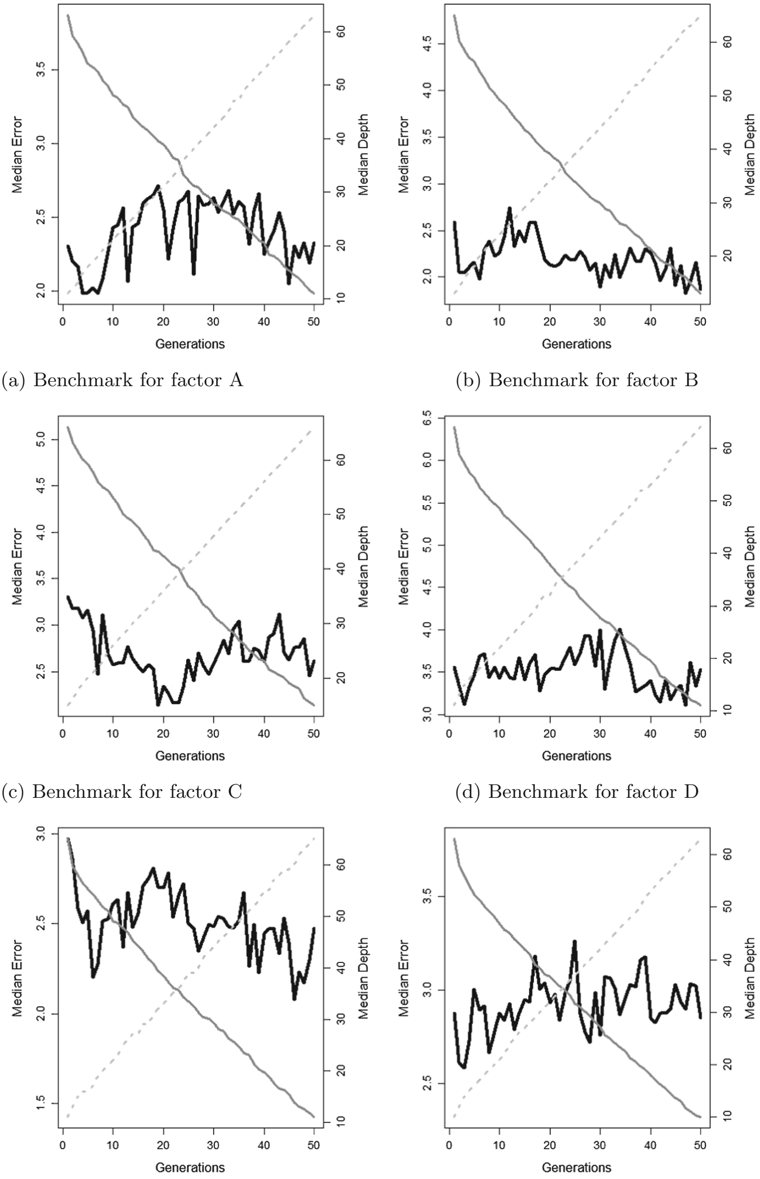
Analysis of Fig. 3 suggests that the evolutionary process should not take more than 5 generations, excepting factor C, where it is reasonable to conduct it for 20 generation. Further evolution, in median terms, does not seem to contribute for the generalization ability of the final individuals, because the validation error does not decrease anymore or starts to increase.

It is worth noticing that, after applying EDDA initialization technique, the evolutionary process turns out to be a mean of recombination of (potentially local) optimal solutions which (potentially) surround global optima. For this reason, the evolution does not require *many* iterations.

Table 1 exhibits the main features of the selected individuals (predictive models) for each of six factors. The second column of the table provides the generalization ability, measured as the RMSE calculated on validation set and averaged across 41 runs. The third column provides information about efficiency of the models, calculated as standard deviation of RMSE. Finally, the last fourth column provides the depth of each individual.

Table 1. Assessment of generalization ability of evolved individuals for prediction of the six factors.

Factor	$\overline{RMSE}$	s_{RMSE}	Depth
A	3.83	4.36	13
B	4.54	4.89	21
C	4.71	4.94	20
D	5.93	6.23	13
E	2.95	2.84	17
Global	3.78	3.91	14



(a) Benchmark for factor A

(b) Benchmark for factor B

(c) Benchmark for factor C

(d) Benchmark for factor D

(e) Benchmark for factor E

(f) Benchmark for global factor

Fig. 3. Evolution of the median best error (left vertical axis) and median depth (right vertical axis), on training and validation data for factors: A (a), B (b), C (c), D (d), E (e) and the global factor (f). The legend for the all sub-plots in the figure is: ―― training error ―― validation error ┈┈┈ depth.

Advantages of EDDA. In order to fully understand the practical advantage of EDDA, consider Fig. 4, which exhibits one of the six evolved predictive models. Only the fact individuals can fit in one-third of an A4 page is fascinating. Compared to a classical initialization algorithm (like RHH) applied to evolve a GSGP population, the final solutions for exactly the same problem and comparable generalization ability, would hardly fit on all the pages of this Proceedings volume. This issue is extremely important, because, for example, we were able to settle all six models in a web-application, where therapists only need to input GMFM values of a given patient.

```
(+ (+ (+ (+ (∗ (+ (+ (+ (+ (∗ sin(X1) (/ X7 (− X44 (− (−
    X43 X83) (ˆ(1/2) X39)))))) (+ (+ X45 X13) (+ X37 X96)))
     (+ cos(X66) (− (ˆ(1/2) (∗ (∗ sin(X60) X8) sin(X39)))
    (+ X48 X40)))) (cos(+ X33 (+ (− X35 (sin() X75)) (+
    0.0 (+ (sin() X8) X52)))))) (∗ 0.75 (− (LF cos(X65)) (
    LF (/ X80 (+ X56 (∗ 0.25 (∗ sin(X48) X89)))))))) (LF
    (− (ˆ(1/2) X39) X79))) (∗ (− 1.0 (LF (− (ˆ(1/2) X39)
    X79))) (+ (∗ (+ X96 (/ (∗ (/ (+ (ˆ(1/2) (/ X54 X94))
    X55) X66) X6) X1)) (LF (+ X83 X80))) (∗ (− 1.0 (LF (+
    X83 X80))) (+ (+ (/ X69 X53) (/ X69 X53)) X96))))) (∗
    0.91 (− (LF (+ X94 (/ (∗ (+ X60 (+ X91 X14)) −0.75) (/
     (− X54 X10) X41)))) (LF X79)))) (∗ 0.81 (− (LF 0.0) (
    LF (cos(sin(cos(ˆ(1/2) (∗ 0.5 −0.75))))))))) (∗ 0.74
    (− (LF (− (sin(/ (ˆ(1/2) X23) X34)) (+ (/ (∗ sin(−1.0)
     X26) X52) X53))) (LF (/ X74 (/ X68 (− (ˆ(1/2) X62) (/
     sin(X66) (+ X40 X58)))))))))
```

Fig. 4. The best model we have been able to evolve for the prediction of the global factor, expressed in Polish prefix notation

Considering the model reported in Fig. 4, Fig. 5 exhibits its most relevant features (predictors). The following list provides the five most frequent features, among the ones reported in Fig. 5, and their interpretation in the context of Pediasuit Protocol:

- **Factor E:** one of the main groups in GMFM which encloses 24 measures into the *Walking, Running and Jumping* category. This summary measure is the most advanced in terms of GMFM. If therapists are able to improve patients motor functioning in this category, they will be able to improve patients motor functioning as a whole;
- **B26:** one of the 88 GMFM measures, embedded in the category *Sitting*. Following [4], it can be described as the ability of a patient, while sitting on a mat, to touch a toy placed 45° behind his/her right side and return to start;
- **C38:** another GMFM measure, embedded in the category *Sitting*. Following [4], it can be described as the ability of the patient to creep 1.8 m forward;

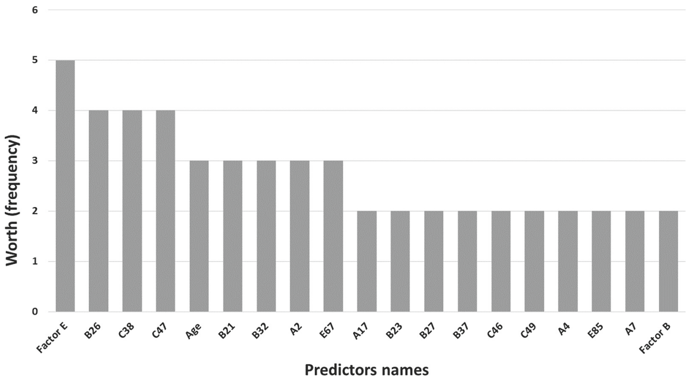


Fig. 5. The features (predictors) that appear more frequently in the model of Fig. 4, together with their worth factor (frequency).

- **C47:** another GMFM measure, embedded in the category *Sitting*. Following [4], it can be described as the ability of the patient to crawl backward down four steps on their hands and knees/feet. The fact B26, C38 and C47 appear to be the most frequent states their importance in therapeutic context. We speculate that a change in therapeutic design, which takes more in consideration improvement in these specific (or similar) motor activities, can produce a better improvement in patients health.
- **Age:** the age of the patient taken before performing the therapy. Our findings demonstrate that patients in different age groups respond differently to the therapy: those whose age is below nine years old show twice as good improvement, measured in terms of the global factor, as the patients whose age is equal or higher than nine years old. This evidence, according to the therapists, has to do with the fact that the body of a younger patient is, naturally, in a more dynamic phase of growth, and this reflects positively on their gross motor functioning improvement after the therapy.

More Than Prediction. In the context of the Pediasuit Protocol, more than merely providing a prediction, the six models that were developed bring other practical benefits:

- they can be a reference for the efficiency of the therapy. Concretely, if, after attempting the therapy, the final GMFM evaluation of a given patient will differ highly from the expectation, then the therapeutic approach at hand (its intensity, aim, coordination, etc.), might be inadequate for that particular patient;

– they can be used to prioritize the queue of patients. Concretely, and hypothesizing the lack of internal resources to address the needs of all the patients, models can help to decide which patients should attempt the therapy first. For example, a patient who may present an extremely debilitated motor functioning and whose expected improvement will be high, can receive higher priority over a patient whose motor functioning is significantly better and the expected improvement is smaller (this should require the agreement of both parties);
– they can be used as a simple informative tool, answering a elementary question: *"will this treatment help my beloved and how much?"* People who take their relatives to such therapy have hope in improvement. This hope can be concretized by providing an estimate of improvement in terms of motor function through different factors;
– they can be used to confirm the diagnosis. In the field of rare diseases, one of the main challenges for the doctors is to identify the diagnosis correctly. This comes from the simple fact that the diseases are rare and, for some of them, there is no concise and extensive framework for their identification;
– they can be used as means of therapeutic design improvement. One of the main features of GP is that individuals can be interpreted. By studying their structure, one can identify the most relevant predictors and their impact. As such, one can say which attributes are determinant for the improvement of the patients at each GMFM factor.

6 Conclusion

Patients affected by rare diseases need very specific and personalized therapies. Casa dos Marcos is the largest specialized medical and residential center for rare diseases in the Iberian Peninsula and has the capacity for hosting several thousands of patients per year. Nowadays, the design and development of personalized therapies and treatments is widely done manually by a team of specialists. Besides conceivably slow, this manual process is also subjective, and thus prone to errors. This motivates the impelling demand for intelligent computational systems, able to support and speedup the decision process. Machine learning is clearly a reasonable option, and predictive models, able for instance to give information about possible reactions of patients to therapies, can be of paramount importance for the development of Casa dos Marcos and the improvement of their everyday work.

This paper summarizes the main outcomes of a project that our research team developed in collaboration with Casa dos Marcos: a Genetic Programming (GP) based system, able to generate predictive models for important motor functioning factors concerning patients who happen to have rare diseases, after some specific therapies.

The presented system uses a very recent development of GP, called Evolutionary Demes Despeciation Algorithm (EDDA). EDDA integrates both the standard version of GP and Geometric Semantic GP (GSGP) in the initialization phase, aiming at capturing the advantages of both these techniques, while at the same time mitigating their respective flaws. One of the major advantages of EDDA is the ability to generate models with comparable accuracy of the ones generated by GSGP, but at the same time with a much smaller size. This is an extremely important characteristic, since the models generated by GSGP are known to be very accurate, but also extremely large. Exploiting this ability, in this paper, we have been able to show and comment on the best model evolved by EDDA, something that would have been unimaginable for GSGP. The reported model is very informative on the effect of the therapy and experts have validated the small subset of features that it uses. Being able to see and, at least partially, interpret the evolved model has been of fundamental importance because it has allowed the personnel of Casa dos Marcos to trust our system. Also, thanks to this increased trust, the models evolved by the presented system are now integrated in a web application, that we have developed, and is nowadays presently in use in Casa dos Marcos.

Acknowledgments. This work was partially supported by national funds through FCT (Fundação para a Ciência e a Tecnologia) under project DSAIPA/DS/0022/2018 (GADgET) and project PTDC/CCI-INF/29168/2017 (BINDER).

References

1. Rare disease resources & FAQs. https://rarediseases.org/for-patients-and-families/information-resources/resources-faqs/
2. Scheeren, E.M., Mascarenhas, L.P.G., Chiarello, C.R., Costin, A.C.M.S., Oliveira, L., Neves, E.B.: Description of the pediasuit protocol^{TM}. Fisioterapia em movimento **25**(3), 473–480 (2012)
3. Centro de desenvolvimento e reabilitação da casa dos marcos. http://rarissimas.pt/centro-de-desenvolvimento-e-reabilitacao-da-casa-dos-marcos/
4. Russell, D.J., Rosenbaum, P.L., Cadman, D.T., Gowland, C., Hardy, S., Jarvis, S.: The gross motor function measure: a means to evaluate the effects of physical therapy. Dev. Med. Child Neurol. **31**(3), 341–352 (1989)
5. Bojarczuk, C.C., Lopes, H.S., Freitas, A.A., Michalkiewicz, E.L.: A constrained-syntax genetic programming system for discovering classification rules: application to medical data sets. Artif. Intell. Med. **30**(1), 27–48 (2004)
6. Castelli, M., Vanneschi, L., Manzoni, L., Popovič, A.: Semantic genetic programming for fast and accurate data knowledge discovery. Swarm Evol. Comput. **26**, 1–7 (2016)
7. Hu, T., Oksanen, K., Zhang, W., Randell, E., Furey, A., Zhai, G.: Analyzing feature importance for metabolomics using genetic programming. In: Castelli, M., Sekanina, L., Zhang, M., Cagnoni, S., García-Sánchez, P. (eds.) EuroGP 2018. LNCS, vol. 10781, pp. 68–83. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-77553-1_5
8. Beger, R.D., et al.: For “Precision Medicine and Pharmacometabolomics Task Group”-metabolomics society initiative: metabolomics enables precision medicine: “a white paper, community perspective”. Metabolomics **12**(9), 149 (2016)

9. Castelli, M., Vanneschi, L., Popovič, A.: Parameter evaluation of geometric semantic genetic programming in pharmacokinetics. Int. J. Bio-Inspired Comput. **8**(1), 42–50 (2016)
10. Castelli, M., et al.: An efficient implementation of geometric semantic genetic programming for anticoagulation level prediction in pharmacogenetics. In: Correia, L., Reis, L.P., Cascalho, J. (eds.) EPIA 2013. LNCS (LNAI), vol. 8154, pp. 78–89. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40669-0_8
11. Vanneschi, L., Castelli, M., Manzoni, L., Silva, S.: A new implementation of geometric semantic GP and its application to problems in pharmacokinetics. In: Krawiec, K., Moraglio, A., Hu, T., Etaner-Uyar, A.Ş., Hu, B. (eds.) EuroGP 2013. LNCS, vol. 7831, pp. 205–216. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-37207-0_18
12. Smith, S.L., Cagnoni, S.: Genetic and Evolutionary Computation: Medical Applications. Wiley, Chichester (2011)
13. Koza, J.: Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, Cambridge (1992)
14. Moraglio, A., Krawiec, K., Johnson, C.G.: Geometric semantic genetic programming. In: Coello, C.A.C., Cutello, V., Deb, K., Forrest, S., Nicosia, G., Pavone, M. (eds.) PPSN 2012. LNCS, vol. 7491, pp. 21–31. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-32937-1_3
15. Vanneschi, L., Castelli, M., Silva, S.: A survey of semantic methods in genetic programming. Genet. Program Evolvable Mach. **15**(2), 195–214 (2014)
16. Castelli, M., Silva, S., Vanneschi, L.: A c++ framework for geometric semantic genetic programming. Genet. Program Evolvable Mach. **16**(1), 73–81 (2015)
17. Castelli, M., Manzoni, L., Gonçalves, I., Vanneschi, L., Trujillo, L., Silva, S.: An analysis of geometric semantic crossover: a computational geometry approach. In: IJCCI (ECTA), pp. 201–208 (2016)
18. Oliveira, L.O.V., Otero, F.E., Pappa, G.L.: A dispersion operator for geometric semantic genetic programming. In: Proceedings of the Genetic and Evolutionary Computation Conference, pp. 773–780. ACM (2016)
19. Pawlak, T.P., Krawiec, K.: Semantic geometric initialization. In: Heywood, M.I., McDermott, J., Castelli, M., Costa, E., Sim, K. (eds.) EuroGP 2016. LNCS, vol. 9594, pp. 261–277. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-30668-1_17
20. Vanneschi, L., Bakurov, I., Castelli, M.: An initialization technique for geometric semantic GP based on demes evolution and despeciation. In: IEEE Congress on Evolutionary Computation (CEC), pp. 113–120. IEEE (2017)
21. Bakurov, I., Vanneschi, L., Castelli, M., Fontanella, F.: EDDA-V2 – an improvement of the evolutionary demes despeciation algorithm. In: Auger, A., Fonseca, C.M., Lourenço, N., Machado, P., Paquete, L., Whitley, D. (eds.) PPSN 2018. LNCS, vol. 11101, pp. 185–196. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99253-2_15
22. Bartashevich, P., Bakurov, I., Mostaghim, S., Vanneschi, L.: PSO-based search rules for aerial swarms against unexplored vector fields via genetic programming. In: Auger, A., Fonseca, C.M., Lourenço, N., Machado, P., Paquete, L., Whitley, D. (eds.) PPSN 2018. LNCS, vol. 11101, pp. 41–53. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99253-2_4

23. Bartashevich, P., Bakurov, I., Mostaghim, S., Vanneschi, L.: Evolving PSO algorithm design in vector fields using geometric semantic GP. In: Proceedings of the Genetic and Evolutionary Computation Conference Companion, GECCO 2018, Kyoto, Japan, 15–19 July 2018, pp. 262–263 (2018)
24. Vanneschi, L., Silva, S., Castelli, M., Manzoni, L.: Geometric semantic genetic programming for real life applications. In: Riolo, R., Moore, J.H., Kotanchek, M. (eds.) Genetic Programming Theory and Practice XI. GEC, pp. 191–209. Springer, New York (2014). https://doi.org/10.1007/978-1-4939-0375-7_11