






Detection of Frailty Using Genetic Programming

The Case of Older People in Piedmont, Italy

Adane Tarekegn¹(✉) , Fulvio Ricceri^{2,3} , Giuseppe Costa^{2,3}, Elisa Ferracin³,
and Mario Giacobini⁴(✉) 

¹ Department of Mathematics “Andrea Peano”, University of Turin, Turin, Italy
adanenega.tarekegn@unito.it

² Department of Clinical and Biological Sciences, University of Turin, Turin, Italy
{fulvio.ricceri, giuseppe.costa}@unito.it

³ Unit of Epidemiology, Regional Health Service ASL TO3, Grugliasco, TO, Italy
elisa.ferracin@epi.piemonte.it

⁴ Data Analysis and Modeling Unit, Department of Veterinary Sciences,
University of Turin, Turin, Italy
mario.giacobini@unito.it

Abstract. Frailty appears to be the most problematic expression of elderly people. Frail older adults have a high risk of mortality, hospitalization, disability and other adverse outcomes, resulting in burden to individuals, their families, health care services and society. Early detection and screening would help to deliver preventive interventions and reduce the burden of frailty. For this purpose, several studies have been conducted to detect frailty that demonstrates its association with mortality and other health outcomes. Most of these studies have concentrated on the possible risk factors associated with frailty in the elderly population; however, efforts to identify and predict groups of elderly people who are at increased risk of frailty is still challenging in clinical settings. In this paper, Genetic Programming (GP) is exploited to detect and define frailty based on the whole elderly population of the Piedmont, Italy, using administrative databases of clinical characteristics and socio-economic factors. Specifically, GP is designed to predict frailty according to the expected risk of mortality, urgent hospitalization, disability, fracture, and access to the emergency department. The performance of GP model is evaluated using sensitivity, specificity, and accuracy metrics by dividing each dataset into a training set and test set. We find that GP shows competitive performance in predicting frailty compared to the traditional machine learning models. The study demonstrates that the proposed model might be used to screen future frail older adults using clinical, psychological and socio-economic variables, which are commonly collected in community healthcare institutions.

Keywords: Frailty · Prediction · Genetic Programming · Imbalanced data

1 Introduction

An increase in longevity results in older people struggling with age-related diseases and functional conditions [1]. This presents enormous challenges towards establishing new approaches for maintaining health at a higher age. An essential aspect of age-related health problems of the general patient condition is the onset of frailty. Even though there are a wide number of studies that have been developed to conceptualize and operationalize frailty, a gold standard definition of frailty still lacks [3–5]. Frailty in elderly people was first characterized as a physical phenotype by Fried et al. [6]. According to this study, frailty is defined on the basis of five physical components: exhaustion, weight loss, slow gait speed, weakness, and low levels of physical activity. People who meet three or more of the above mentioned physical components are classified as frail. Those people who meet one or two criteria as pre-frail and people who meet none of these criteria are classified as not frail. This research was only phenotypic and didn't consider other causes such as psychological and cognitive factors to measure frailty. On the other hand, Rockwood et al. [7] developed a model to detect frailty based on accumulated deficits. In [9,10], the comparison of the frailty phenotype and the frailty index models were also widely discussed. As indicated in the literature, several frailty scores based on different frailty concepts have been developed. However, each of the available tools intended to detect frailty poorly agrees with each other when applied to the same population [11].

The frailty syndrome is associated with a high risk for injurious falls, urgent hospitalization, preventable hospitalization, disability, fracture, access to emergency admissions with red code, and mortality. Using predictive modeling, administrative data allows the detection of potential risk factors and can be used as a clinical decision support system, which provides health professionals with information on the probable clinical patient outcome. This enables the physicians to react quickly and to avoid the likely adverse effects in advance. The identification of elderly people at risk of frailty is essential to provide appropriately tailored care and effectively manage healthcare resources [2].

Most existing studies in the relevant literature for detection of frailty rely on clinical information to investigate the effects of frailty outcomes in the elderly, although these detailed and accurate clinical data may not be adequately available [29]. Models that incorporated patient-level factors such as medical comorbidities and basic demographic data with variables from clinical assessment scores and included numerous social factors have gained good explanatory results. However, prediction remains a poorly understood and complex endeavor, especially when it comes to using available large administrative data. Administrative databases can be used as a better source to implement models able to define, detect, and measure frailty [12]. In [13,14] retrospective studies based on logistic regression models are proposed to develop frailty risk index and validate their content using health record data. There are also few models that are derived from a single source of information, like primary care electronic health record data and only insurance claims data [15]. More recent work on frailty was proposed by F. Bertini et al. [16] using logistic regression. In this paper, they

proposed a frailty prediction model using a broad set of socio-clinical and socio-economic variables. Their model was designed to detect and categorize frailty according to the expected risk of hospitalization or death. In general, the frailty indexes proposed in most literature have focused on the possible risk factors associated with frailty in the elderly population, but predicting who is at risk of frailty problems is still requires further investigation. In our work, we proposed a frailty prediction model using Genetic Programming (GP) to detect frailty based on different outcomes of frailty conditions, including mortality, disability, hospitalization, fracture, and access to the emergency department with red code.

To date, various literature on frailty pays particular attention to the statistical methods to detect and predict frailty. However, evolutionary algorithms, such as GP, could also have the capability to address the frailty problems. The ability of GP to produce high performance results depends on the nature of the problem as there is no single algorithm that works best for every problem. As a result, we compared the results of GP with the other commonly used machine learning models in terms of prediction performance on the six different problems of frailty: mortality, access to the emergency department with red code, disability, fracture, urgent hospitalization and preventable hospitalization. On each of the six problems, the results of GP were compared with support vector machine, random forest, artificial neural network and decision tree. The detailed descriptions of these machine learning methods can be found in [8].

2 Methods

2.1 Data Source

We used medical administrative data, which capture patient demographics, healthcare utilization, chronic conditions, and recorded diagnoses to develop predictive models for frailty. The data is based on the Piedmontese Longitudinal Study, an individual record linkage that is available for about 4 millions of Piedmont (Italy) inhabitants between the Italian 2011 census and the administrative and health databases (enrollees registry, hospital discharges, drug prescriptions, outpatient clinical investigation database, and health exemptions) and that is included in the Italian Statistical National Plan. About one million patients aged 65 and above are included in the study. For each patient, a total of 64 different variables are recorded describing histories of frailty related conditions and outcomes. 58 different input variables and 6 different output variables for each subject are included in the dataset. All outcomes and comorbidity variables are represented by Boolean values. The demographic variables such as age, marital status, citizenship, education level, income status, family size, and others are specified using the dummy variables. The ‘age’ variable is grouped into six categories, with 65–69 used as the first category. The output variables are described as outcomes or measurable changes in the health status of patients. All the 58 input variables were collected in 2016, while the 6 output variables were collected in 2017. So, GP model development was based on using the 2016

variables as input and the 2017 variables as unwanted output. Table 1 presents the description of the 6 output variables in the dataset.

2.2 Data Transformation

The dataset is large in volume and multidimensional, consisting of 58 input variables and 6 different output variables that are assigned simultaneously to each elderly person. This type of data is what we call ‘multi-output’ dataset. The way the data set is organized is such that one patient can have multiple outcomes. In particular, we identified 6 different outcomes that are associated with frailty conditions namely, mortality, disability, urgent hospitalization, fracture, preventable hospitalization, and access to the emergency department (ED) with red code. This multi-output dataset is transformed into six single-output problems associated with each output variable. Decomposing the original data into six independent datasets helps to study each output independently for the given number of similar risk factors. Transforming the original problem into single independent problems is a straightforward way to implement using GP since it involves transforming the data rather than the algorithm. Additionally, with this method, we can take full advantage of GP since it considers learning problems that contain only one output, i.e., each instance is associated with one single nominal target variable characterizing its property. The six problems with their respective datasets are analysed independently. The descriptive statistics of each dataset are presented in Table 1.

Table 1. Descriptive statistics of datasets in each problem.

Problem (variable)	Category	Code	Number	Percent
Mortality	No	0	1,053,790	96.18
	Yes	1	41,823	3.82
Access to ED with red code	No	0	1,088,124	99.32
	Yes	1	7,489	0.68
Disability	No	0	1,064,186	97.13
	Yes	1	31,427	2.87
Fracture	No	0	1,088,530	99.35
	Yes	1	7,083	0.65
Urgent hospitalization	No	0	1,056,695	96.45
	Yes	1	38,918	3.55
Preventable hospitalization	No	0	1,076,541	98.26
	Yes	1	19,072	1.74

2.3 Learning from Imbalanced Data

Imbalanced data sets are common in medicine and other domains, such as fraud detection [25]. The issue of imbalanced datasets has gathered wide attention from researchers during the last several years [25, 34]. It occurs when the samples represented in a problem show a skewed distribution, i.e., when there is a majority (or negative samples) and a minority (or positive samples). Analyzing such a complex nature of the dataset becomes an issue in the machine learning community including genetic programming [24] and it is observed that most of the traditional machine learning algorithms are very sensitive with imbalanced data [26, 27]. Usually, accurate classification of minority class samples is more important than majority class samples especially in medical diagnosis [24]. The datasets of the six problems in Table 1 (mortality, access to ED with red code, disability, fracture, urgent hospitalization and preventable hospitalization) are imbalanced because the negative class (class ‘0’) contains more samples than the other (class ‘1’). For all datasets, the imbalanced rate ranges approximately between 1%–4% (that is, the percent range of the data samples that belong to the positive class). In such cases, it is challenging to create an appropriate testing and training datasets for the GP, given that GP is built with the assumption that the test dataset is drawn from similar distribution as the training dataset [17]. Providing imbalanced data to a classifier will produce undesirable results such as much lower performance and increasing the number of false negatives. Among the techniques that deal with imbalanced data, we used the data-level approach to rebalance the class distribution. This is done by either employing under-sampling or oversampling to reduce the imbalance ratio in the dataset [18]. Under-sampling balances the dataset by reducing the size of the abundant class [19, 20], while over-sampling duplicates samples from the minority class [21, 22]. This would possibly improve the performance of classification, as long as the re-sampling does not cause information loss. The oversampling technique is used when the data set is quite small in size. In our case, since the amount of collected data is sufficient, we adopted under-sampling to rebalance the sample distribution. We applied this strategy for all problems with their respective dataset. After performing the undersampling of the majority class, we found a balanced proportion between the positive and negative classes for each dataset, as shown in Table 2.

Table 2. Positive and negative classes in each dataset

Dataset	Class category			
	Positive class		Negative class	
	Count	Percent	Count	Percent
Mortality	41823	50%	41823	50%
Access to ED with red code	7489		7489	
Disability	31427		31427	
Femur fracture	7083		7083	
Urgent hospitalization	38918		38918	
Preventable hospitalization	19072		19072	

3 Experiments

In the present study, we investigated the applicability of GP in the prediction of frailty among patients in elderly people, as explained in the previous section. The experiments include learning a binary classification of the data to frail and non-frail classes by considering the profiles of each individual patient over two years. In analysing the data for prediction, the output variables represent an occurrence in the next year, and the GP predictive model is proposed to detect and classify frailty according to the expected risk of urgent hospitalization, preventive hospitalization, disability, fracture, emergency admissions with a red code and death within a year. The GP model is trained using the training dataset (70%) and tested using test dataset (30%). The training dataset was used for building the model, and the test dataset was used to evaluate the prediction capabilities.

To build an effective predictive model, it is essential to train the model and perform testing using a dataset that comes from the same target distribution. All the six different datasets were randomly split into training and testing using the following steps.

1. Split the samples with negative class into 70% training and 30% testing.
2. Split the samples with positive class into 70% training and 30% testing set.
3. Combine the 70% samples with negative class obtained from step 1 and the 70% samples with positive class obtained from step 2.
4. Combine the 30% samples with negative class obtained from step 1 and the 30% samples with positive class obtained from step 2.
5. Perform a chi-square test with a significance level of 0.05 between the training set obtained from step (3) and the test set obtained from step (4). A statistical test was needed to check if the training set and testing set are representative of each other. A Chi-square independence test is used to determine if there is evidence of a difference between the training set (70%) and the test set (30%) with respect to the 58 categorical input variables. The produced test results are assessed based on the chi-square statistic, and statically significant results were found with respect to all variables.

3.1 GP Parameter Setup

In GP, setting the control parameters is an important first step to manipulate data and to obtain good results. In our datasets, we tried several experiments for classification tasks by using the control parameters of GP proposed in HeuristicsLab [33], such as population size, selection method, number of elite individuals, initialization method, number of generations, crossover probability rates, and mutation probability rates. Due to the stochastic nature of GP, 30 runs were performed in all problems, each with a different random number generator seed. For our frailty problem, we specifically focused on the two common parameters of GP: Maximum number of generations and Population size. In order to investigate the effect of few generation over larger population and small population over more generations and also to get an advantage from either of these

GP parameter settings, we run two different algorithms of GP (GP1 and GP2) under varying population size and the maximum number of generations, keeping all other parameters set to default. The maximum number of generations and population size for GP1 is set to be 1000 and 100, respectively. In GP2, we set a maximum number of generations to be 100 and population size 1000. For all frailty problems, GP1 and GP2 were applied, and for each experiment, 30 runs were performed with the same initial configurations of parameters. We clearly observed that the runs with a population size of 1000 and generation 100 are related to the immense runtime requirements, comparing with the runs of population size 100 and generation 1000. In fitness, it is apparent that a large population running for a small number of generations behaves differently from the small population running for a large number of generations. The fitness of GP1 and GP2 across generations were compared for mortality and fracture problems using mean squared error (MSE). The MSE is used as fitness to compare the quality of the two models (GP1 and GP2), and it was observed that GP2 produced lower error rates, which is ranging from 0.18 to 0.25 for mortality and from 0.19 to 0.25 for fracture problems. While for GP1 the MSE is much higher, which is ranging from 0.20 to 0.30 for mortality problem and from 0.22 to 0.29 for fracture problem. The results show that a large population is more likely than a small population to make more significant improvements in fitness from one generation to the next, given that it generates more new trees in each generation. Generally, for frailty problems, it seems that results with GP2 are more stable and that larger population is a better choice than many generations. As a result of this, we preferred GP with larger population size and smaller number of generations for the prediction of frailty conditions. The summary of parameters used for running GP2 experiments is presented in Table 3.

Table 3. GP parameters used in the experiment.

Parameter Name	Value
Algorithm	GP2
Maximum number of generations	100
Population size	1000
Mutation rate	15%
Crossover rate	90%
Solution creator	Ramped Half-and-Half
Maximum tree depth	10
Maximum tree length	100
Elites	1
Terminal set	Constant, variables

4 Results

In this section, we investigated the performance of GP for the prediction of frailty status in terms of the six problems or outcomes. The predictors common to all problems and which were also included in the final model produced by GP were the age, the number of urgent hospitalization, charlson comorbidity index, dementia and mental disease. The final prediction model of each problem generated by GP is a binary parse tree representing the classification model.

4.1 GP Prediction Performance

The different frailty prediction models obtained from GP were evaluated in terms of overall accuracy, sensitivity and specificity on the training and test dataset. In the context of this study, sensitivity measures the frail subjects who are correctly identified as having the event and specificity refers to the nonfrail subjects who are correctly identified as not having the event. The three performance measures were considered for mortality, urgent hospitalization, preventable hospitalization, disability, fracture, and access to ED with a red code. Detecting these adverse outcomes among a large number of subjects is important when applied in real-world practice. Hence, the true positive rate (TPR), also called sensitivity, was the main metric to consider. The overall accuracy (Acc) and true negative rate (TNR), also called specificity, were measured as additional performance metrics. The accuracy, TPR, and TNR were formulated using the true positives (TP), false positives (FP), true negatives, and false negatives (FN) [28].

In analysing GP for classification, the most important aspect is to know the number of samples that are classified correctly and those, which are classified incorrectly. The results averaged from 30 runs of GP experiments are presented in Table 4 on the training set and Table 5 on the testing set. In these problems, using sensitivity and specificity allows to correctly identify those with the disease condition (frail people) and to correctly identify those without the disease (non frail people), respectively. The standard deviation (SD) for mean sensitivity, specificity and accuracy are also calculated, since each problem is run 30 times, as shown in Tables 4 and 5. For mortality problem GP produced the best performance in all measurements. For access to ED with red code, the overall

Table 4. Performance of GP on the training set.

Problem	Sensitivity (SD)	Specificity (SD)	Accuracy (SD)
Mortality	0.75(0.05)	0.75(0.06)	0.75(0.02)
Access to ED with red code	0.76(0.24)	0.45(0.37)	0.58(0.09)
Disability	0.72(0.04)	0.69(0.05)	0.72(0.02)
Fracture	0.71(0.04)	0.67(0.14)	0.74(0.08)
Urgent Hospitalization	0.65(0.22)	0.63(0.29)	0.64(0.13)
Preventable Hospitalization	0.71(0.18)	0.63(0.33)	0.67(0.11)

Table 5. Performance of GP on the testing set.

Problem	Sensitivity (SD)	Specificity (SD)	Accuracy (SD)
Mortality	0.75(0.05)	0.76(0.06)	0.75(0.02)
Access to ED with red code	0.73(0.24)	0.43(0.36)	0.58(0.08)
Disability	0.70(0.04)	0.73(0.05)	0.71(0.02)
Fracture	0.71(0.14)	0.67(0.08)	0.72(0.04)
Urgent Hospitalization	0.66(0.22)	0.62(0.29)	0.63(0.13)
Preventable Hospitalization	0.73(0.18)	0.64(0.33)	0.68(0.11)

accuracy and specificity of GP are slightly lowered. For the remaining problems the performance of GP is at an acceptable level. These results confirmed the predictive capability of GP on frailty problems.

4.2 Performance of Other Non-GP Classifiers

In this section, we assessed the theoretical and performance comparison of GP with the statistical and machine learning methods. In the literature, there are some studies which compare GP with other statistical and machine learning methods [23, 35]. The studies suggest that GP may be better at representing the potentially non-linear relationship of (a smaller subset of) the strongest predictors, although the complexity of the GP-derived model was found to be much higher. The fact that GP required fewer predictors to achieve similar performance may have an advantage in practical application of the developed clinical prediction models. Therefore, a prediction model that requires fewer inputs, especially if the information relating to these inputs is in practice recorded easily and to a good quality, would considerably increase adoption and utility. Comparison of GP with statistical models, such as cox regression techniques, was attempted by [30] in terms of the performance of a cardiovascular risk score using a prospective cohort study of patients with symptomatic cardiovascular disease. The predictive ability of cox regression model and GP was evaluated in terms of their risk discrimination and calibration using the validation set. Their findings indicated that the discrimination of both models was comparable. Using the calibration of these models, which was assessed based on calibration plots and the generalization of the Hosmer-Lemeshow test statistic, was also similar, but with the Cox model is better calibrated to the validation data. In [36], a comparison of GP and NN in metamodeling of discrete-event simulation was studied. The results of this study concluded that GP provides greater accuracy in validation tests, demonstrating a better generalization capability than NN, despite the fact that GP when compared to NN requires more computation in model development. Most machine learning methods are usually straightforward to implement and work well with minimum resources; however their blackbox nature makes them non user friendly. On the other hand, GP results are often human friendly and provide an explicit mathematical formula as its output, although developing such

an efficient algorithm and realizing its full potential to solve real-world problems can be challenging. GP algorithms are expected to require a computing time that grows exponentially with the size of the problem [32]. In this study, GP prediction capability was compared with the well-known machine learning classifiers on mortality, disability, fracture, access to ED with red code and hospitalization problems.

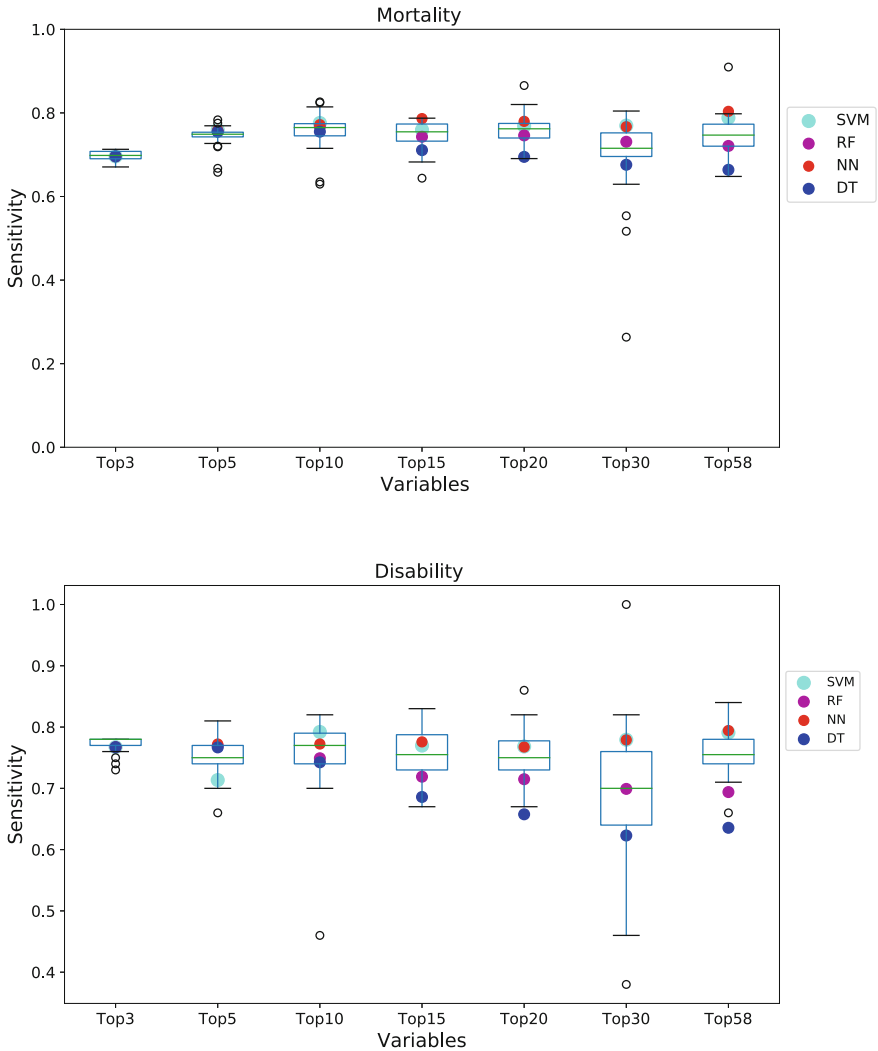


Fig. 1. Performance of GP on Mortality (upper plot) and Disability (lower plot) problems compared to the performance of SVM, RF, NN and DT. The box plots represent the 30 runs of GP with performance measured using sensitivity and the coloured points represent the sensitivity of SVM, RF, NN and DT. Top3 represents the top three variables and so on for each problem.

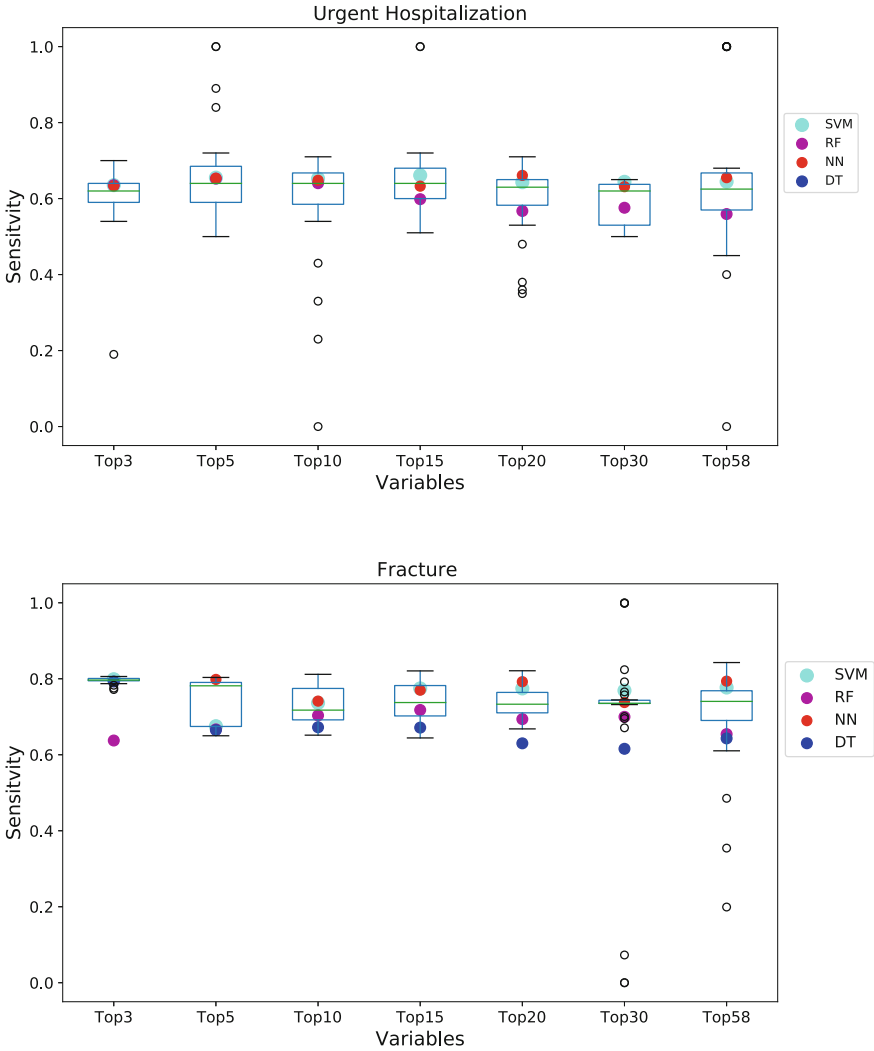


Fig. 2. Performance of GP on Urgent hospitalization (upper plot) and Fracture (lower plot) problems compared to the performance of SVM, RF, NN and DT. The box plots represent the 30 runs of GP with performance measured using sensitivity and the coloured points represent the sensitivity of SVM, RF, NN and DT. Top3 represents the top three variables and so on for each problem.

The most commonly used classifiers such as support vector machines (SVM), artificial neural networks (NN), random forests (RF) and decision trees (DT) were applied in all problems. The results obtained in each problem using the non-GP classifiers are compared with the results of GP using sensitivity. The comparison is based on the ability to identify the positive subjects in the frailty

problems using their respective datasets. The performance of predictions by the different classifiers is shown in Figs. 1 and 2. The figures depict the performance of all classifiers using sensitivity on the testing part of the data. From the figures, the performance values were obtained using different subset of ranked features, the boxplots represent the performance at every 30 runs of GP, and the different colored dots represent the performance of the other machine learning algorithms. In all plots, the x-axis represents the number of features and y-axis represents the performance of GP using sensitivity.

Looking at each box plot of GP in Figs. 1 and 2, we can observe that some runs are outliers in each problem due to the stochastic nature of GP. For example, in urgent hospitalization, there are three runs beyond the whiskers for the top 5 and top 10 variables. These runs are outliers of the 30 runs of GP, plotted as points. In all problems with all variables, the performance of SVM, RF, NN, and DT are displayed under the upper quartile of the GP box plots, indicating the maximum performance obtained from the 30 runs of GP is always greater than the performance of the machine learning models. Comparing all algorithms, decision tree followed by random forest has the lowest performance in all problems for the number of variables greater 10. The average sensitivity of GP overlaps with the performance of NN. However, the accuracy of GP is lowered compared to SVM and NN.

For making the fairest comparison possible between GP and other machine learning models, a pairwise statistical test between the 30 runs of GP and each individual machine learning model was also performed. The statistical test used was the Wilcoxon signed rank test. The Wilcoxon statistical test is a nonparametric test that ranks the differences in performances of GP and other algorithms over each frailty problem. The test was based on the sensitivity score of each algorithm in each problem on the test data at the significance level of 0.01. From the test results, it is found that the results between SVM and GP are statistically significant only in disability, urgent hospitalization and preventable hospitalization problems. Combining the experimental results (Figs. 1 and 2) and Wilcoxon-rank test results, it is concluded that for mortality and fracture problems SVM outperforms GP in sensitivity score, while for access to ED with a red code SVM performs lower than GP. GP outperforms DT in all problems except for urgent hospitalization. NN has a similar performance with GP for all problems excluding mortality and femur fracture.

4.3 Feature Selection Comparison of GP and Chi-Square

The performance of GP feature selection is compared with the well-known Chi-Square feature selection method. The top three variables (age, Charlson index, and the number of urgent hospitalization) selected by GP are also selected by chi-square as top three variables in the mortality problem. After three variables, there is slightly a little difference in the position of variables. Table 6 presents the prediction accuracy of the classification model using the features selected by GP and Chi-square for all problems. For each problem, the best average accuracy of the 30 runs of GP is taken to compare the classification performance

of GP and Chi-square feature selection methods. From this table, Chi-square performed the best in the mortality problem with an accuracy of 76% followed by GP with an accuracy of 75%, a difference of only 1%. This condition holds also for disability and fracture problems. For urgent hospitalization, both GP and chi-square produce a similar performance. The results show that GP can perform the feature selection task with competitive results.

Table 6. Prediction accuracy via feature selection of GP and Chi-square

Problem	GP feature selection	Chi-Square feature selection
Mortality	0.75	0.76
Urgent hospitalization	0.64	0.64
Disability	0.72	0.73
Preventable hospitalization	0.68	0.71
Red code emergency	0.58	0.68
Fracture	0.71	0.73

5 Discussions and Conclusions

The goals of this study were to develop models to predict the risk of hospitalization, disability, mortality, fracture and emergency admissions among the older people in Piedmont, Italy. In this study, we inspected the possibility of using an administrative dataset to detect frailty in older adults using Genetic programming (GP), which was used as a potential tool for developing a prediction model. Six different models were developed, and the performance of each model relies on the input data provided to the learning algorithm. The performances of models created by GP were assessed by splitting the data into training set and test set. The test set was untouched during the entire training and model selection process and only used for the final model evaluation.

To find what works for our frailty problems, we performed several experiments by varying the parameter values of genetic programming. Typically, we tried to discover the optimal parameter choice between two genetic parameters: the population size and the number of generations. In order to get the efficient GP algorithm that best fits our data, many runs of small populations over many generations and large populations over a few generations are compared. For classification problems, the results demonstrated that large populations running for a small number of generations achieve better fitness than small population running for a large number of generations. After selecting the best GP algorithm for our data, several experiments with 30 runs of GP are conducted by adjusting the remaining parameters. The performance of the models obtained by GP is evaluated using sensitivity, specificity, and accuracy. From the results obtained, it is evident that GP algorithms perform well in separating the positive cases

from the negative cases of frailty outcomes. The overall classification performance for both training and testing are comparable with the existing machine learning techniques like artificial neural network, random forest and support vector machines. Overall, the results are encouraging, and further studies on frailty can be investigated to extend the findings on multiple outcomes simultaneously using evolutionary algorithms.

Overall, GP demonstrated substantial potential as a method for the automated development of clinical prediction models for diagnostic and prognostic purposes. The experiments of GP on administrative data acquired from different hospital discharges and drug prescriptions provide comparable accuracy to conventional models in the assessment of the risk of mortality, disability, fracture, access to the emergency department with red code and hospitalization.

References

1. Kojima, G., Liljas, A., Iliffe, S.: Frailty syndrome: implications and challenges for health care policy. *Risk Manag. Healthc. Policy* **12**, 23–30 (2019). <https://doi.org/10.2147/RMHP.S168750>
2. Comans, T.A., Peel, N.M., Hubbard, R.E., Mulligan, A.D., Gray, L.C., Scuffham, P.A.: The increase in healthcare costs associated with frailty in older people discharged to a post-acute transition care program. *Age Ageing* **45**, 317–320 (2016). <https://doi.org/10.1093/ageing/afv196>
3. Clegg, A., Young, J., Iliffe, S., Rikkert, M.O., Rockwood, K.: Frailty in elderly people. *Lancet* **381**, 752–762 (2013). [https://doi.org/10.1016/S0140-6736\(12\)62167-9](https://doi.org/10.1016/S0140-6736(12)62167-9)
4. Wennberg, D., Siegel, M., Darin, B., Filipova, N.: Combined predictive model: final report and technical documentation (2006)
5. Lally, F., Crome, P.: Understanding frailty (2007). <https://doi.org/10.1136/pgmj.2006.048587>
6. Fried, L.P., et al.: Frailty in older adults: evidence for a phenotype. *J. Gerontol. Ser. A Biol. Sci. Med. Sci.* **56**, M146–M157 (2001). <https://doi.org/10.1093/gerona/56.3.M146>
7. Rockwood, K., et al.: A global clinical measure of fitness and frailty in elderly people. *CMAJ* **173**, 489–495 (2005). <https://doi.org/10.1503/cmaj.050051>
8. Kotsiantis, S.B., et al.: Machine learning: a review of classification and combining techniques. *Artif. Intell. Rev.* **26**, 159–190 (2006). <https://doi.org/10.1007/s10462-007-9052-3>
9. Rockwood, K., Andrew, M., Mitnitski, A.: A comparison of two approaches to measuring frailty in elderly people. *J. Gerontol. Ser. A Biol. Sci. Med. Sci.* **62**, 738–743 (2007). <https://doi.org/10.1093/gerona/62.7.738>
10. Blodgett, J., Theou, O., Kirkland, S., Andreou, P., Rockwood, K.: Frailty in NHANES: comparing the frailty index and phenotype. *Arch. Gerontol. Geriatr.* **60**, 464–470 (2015). <https://doi.org/10.1016/j.archger.2015.01.016>
11. Theou, O., Brothers, T.D., Mitnitski, A., Rockwood, K.: Operationalization of frailty using eight commonly used scales and comparison of their ability to predict all-cause mortality. *J. Am. Geriatr. Soc.* **61**, 1537–1551 (2013). <https://doi.org/10.1111/jgs.12420>
12. Katz, A., Wong, S., Williamson, T., Taylor, C., Peterson, S.: Identification of frailty using EMR and admin data: a complex issue. *Int. J. Popul. Data Sci.* **3** (2018). <https://doi.org/10.23889/ijpds.v3i4.832>

13. Chen, C.-Y., Wu, S.-C., Chen, L.-J., Lue, B.-H.: The prevalence of subjective frailty and factors associated with frailty in Taiwan. *Arch. Gerontol. Geriatr.* **50**, S43–S47 (2010). [https://doi.org/10.1016/s0167-4943\(10\)70012-1](https://doi.org/10.1016/s0167-4943(10)70012-1)
14. Lee, D.H., Buth, K.J., Martin, B.J., Yip, A.M., Hirsch, G.M.: Frail patients are at increased risk for mortality and prolonged institutional care after cardiac surgery. *Circulation* **121**, 973 (2010). <https://doi.org/10.1161/CIRCULATIONAHA.108.841437>
15. Homer, M.L., Palmer, N.P., Fox, K.P., Armstrong, J., Mandl, K.D.: Predicting falls in people aged 65 years and older from insurance claims. *Am. J. Med.* **130**, 744.e17–744.e23 (2017). <https://doi.org/10.1016/j.amjmed.2017.01.003>
16. Bertini, F., Bergami, G., Montesi, D., Veronese, G., Marchesini, G., Pandolfi, P.: Predicting frailty condition in elderly using multidimensional socioclinical databases. *Proc. IEEE* **106**, 723–737 (2018). <https://doi.org/10.1109/JPROC.2018.2791463>
17. Amari, S.: Machine learning. In: Amari, S. (ed.) *Information Geometry and Its Applications*. AMS, vol. 194, pp. 231–278. Springer, Tokyo (2016). https://doi.org/10.1007/978-4-431-55978-8_11
18. Japkowicz, N., Stephen, S.: The class imbalance problem: a systematic study. *Intell. Data Anal.* **6**, 429–449 (2018). <https://doi.org/10.3233/ida-2002-6504>
19. Barandela, R., Sánchez, J.S., García, V., Rangel, E.: Strategies for learning in class imbalance problems. *Pattern Recogn.* **36**, 849–851 (2003). [https://doi.org/10.1016/S0031-3203\(02\)00257-1](https://doi.org/10.1016/S0031-3203(02)00257-1)
20. McCarthy, K., Zabar, B., Weiss, G.: Does cost-sensitive learning beat sampling for classifying rare classes? In: *Proceedings of the 1st International Workshop on Utility-based Data Mining - UBDM 2005*, pp. 69–77. ACM Press, New York (2005). <https://doi.org/10.1145/1089827.1089836>
21. Chen, J.X., Cheng, T.H., Chan, A.L.F., Wang, H.Y.: An application of classification analysis for skewed class distribution in therapeutic drug monitoring - the case of vancomycin. In: *Proceedings - IDEAS Workshop on Medical Information Systems: The Digital Hospital, IDEAS 2004-DH* (2005)
22. Orriols, A., Bernadí-Mansilla, E.: Class imbalance problem in UCS classifier system: fitness adaptation. In: *2005 IEEE Congress on Evolutionary Computation, IEEE CEC 2005, Proceedings* (2005)
23. Azimlu, F., Rahnamayan, S., Makrehchi, M., Kalra, N.: Comparing genetic programming with other data mining techniques on prediction models. In: *2019 14th International Conference on Computer Science & Education (ICCSE)*, pp. 785–791. IEEE (2019). <https://doi.org/10.1109/ICCSE.2019.8845381>
24. Amal, S., Periwal, V., Scaria, V.: Predictive modeling of anti-malarial molecules inhibiting Apicoplast formation. *BMC Bioinf.* **14**, 55 (2013). <https://doi.org/10.1186/1471-2105-14-55>
25. Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., Bing, G.: Learning from class-imbalanced data: review of methods and applications. *Expert Syst. Appl.* **73**, 220–239 (2017). <https://doi.org/10.1016/j.eswa.2016.12.035>
26. Kang, Q., Chen, X.S., Li, S.S., Zhou, M.C.: A noise-filtered under-sampling scheme for imbalanced classification. *IEEE Trans. Cybern.* **47**, 4263–4274 (2017). <https://doi.org/10.1109/TCYB.2016.2606104>
27. Leevy, J.L., Khoshgoftaar, T.M., Bauder, R.A., Seliya, N.: A survey on addressing high-class imbalance in big data. *J. Big Data* **5**(1), 1–30 (2018). <https://doi.org/10.1186/s40537-018-0151-6>
28. Han, J., Kamber, M., Pei, J.: *Data Mining*. Elsevier, Amsterdam (2012). <https://doi.org/10.1016/C2009-0-61819-5>

29. Volrathongchai, K., Brennan, P.F., Ferris, M.C.: Predicting the likelihood of falls among the elderly using likelihood basis pursuit technique. In: AMIA Annual Symposium, Proceedings (2005)
30. Bannister, C.A., Halcox, J.P., Currie, C.J., Preece, A., Spasić, I.: A genetic programming approach to development of clinical prediction models: a case study in symptomatic cardiovascular disease. *PLoS One* (2018). <https://doi.org/10.1371/journal.pone.0202685>
31. Bannister, C.A., Currie, C.J., Preece, A., Spasic, I.: Automatic development of clinical prediction models with genetic programming: a case study in cardiovascular disease. *Value Health* **17**, A200–A201 (2014). <https://doi.org/10.1016/j.jval.2014.03.1171>
32. Poli, R., Koza, J.: Genetic programming. In: Burke, E., Kendall, G. (eds.) *Search Methodologies*, pp. 143–185. Springer, Boston (2014). https://doi.org/10.1007/978-1-4614-6940-7_6
33. HeuristicLab homepage. <https://dev.heuristiclab.com/trac.fcgi/wiki>
34. Vluymans, S.: Learning from imbalanced data. In: *Studies in Computational Intelligence*, pp. 81–110 (2019). https://doi.org/10.1007/978-3-030-04663-7_4
35. Ulloa-Cazarez, R.L., López-Martín, C., Abran, A., Yáñez-Márquez, C.: Prediction of online students performance by means of genetic programming. *Appl. Artif. Intell.* **32**, 858–881 (2018). <https://doi.org/10.1080/08839514.2018.1508839>
36. Can, B., Heavey, C.: A comparison of genetic programming and artificial neural networks in metamodeling of discrete-event simulation models. *Comput. Oper. Res.* **39**, 424–436 (2012). <https://doi.org/10.1016/j.cor.2011.05.004>