International Journal of Medical Informatics 192 (2024) 105603

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

FRELSA: A dataset for frailty in elderly people originated from ELSA and evaluated through machine learning models

Matteo Leghissa *, Álvaro Carrera, Carlos Á. Iglesias

Universidad Politécnica de Madrid, Av. Complutense, 30, 28040, Madrid, Spain

A R T I C L E I N F O

Keywords:
FRELSA
Frailty
Machine learning
ELSA
Artificial intelligence
Fried frailty phenotype
Frailty prediction
MultiSURF

A B S T R A C T

Background: Frailty is an age-related syndrome characterized by loss of strength and exhaustion and associated with multi-morbidity. Early detection and prediction of the appearance of frailty could help older people age better and prevent them from needing invasive and expensive treatments. Machine learning techniques show promising results in creating a medical support tool for such a task.

Methods: This study aims to create a dataset for machine learning-based frailty studies, using Fried's Frailty Phenotype definition. Starting from a longitudinal study on aging in the UK population, we defined a frailty label for each subject. We evaluated the definition by training seven different models for detecting frailty with data that were contemporary to the ones used for the definition. We then integrated more data from two years before to obtain prediction models with a 24-month horizon. Features selection was performed using the MultiSURF algorithm, which ranks all features in order of relevance to the detection or prediction task.

Results: We present a new frailty dataset of 5303 subjects and more than 6500 available features. It is publicly available, provided one has access to the original English Longitudinal Study of Ageing dataset. The dataset is balanced after grouping frailty with pre-frailty, and it is suitable for multiclass or binary classification and prediction problems. The seven tested architectures performed similarly, forming a solid baseline that can be improved with future work. Linear regression achieved the best F-score and AUROC in detection and prediction tasks.

Conclusions: Creating new frailty-annotated datasets of this size is necessary to develop and improve the frailty prediction techniques. We have shown that our dataset can be used to study and test machine learning models to detect and predict frailty. Future work should improve models' architecture and performance, consider explainability, and possibly enrich the dataset with older waves.

1. Introduction

Frailty is an age-related syndrome characterized by loss of strength and exhaustion and associated with multi-morbidity. It is increasingly common worldwide due to the increase in life expectancy, which translates to higher costs for healthcare systems. Using 2014 data, it was estimated that frailty increased healthcare costs by 5.8 billion pounds per year in England [1].

Frailty increases the risk of adverse outcomes, including falls, hospitalization, and mortality, as well as healthcare costs and usage [2]. Pre-frailty can be treated with interdisciplinary primary healthcare to significantly improve the chances of delaying the onset of later stages of frailty [3].

Several different frailty definitions are accepted among clinicians, the most common ones being the Fried's Frailty Phenotype (FFP) [4], the Clinical Frailty Scale (CFS) [5], the electronic Frailty Index (eFI) [6], and the Frailty Trait Scale (FTS) [7]. Based on these and other definitions, many detection/screening tools have been proposed and adopted in different healthcare systems [8–12]. These tools often require the patients to take specific tests or answer questionnaires, which takes time and are not operations that can be frequently repeated. However, since frailty in its early stages is reversible [13], an automated tool supporting medical operations based on clinical data could prove very effective.

Machine Learning (ML) represents a promising path toward developing such a tool, which has been experimented with in many studies. The main ones are collected in a systematic review by the same authors

* Corresponding author.
E-mail addresses: matteo.leg@upm.es (M. Leghissa), a.carrera@upm.es (Á. Carrera), carlosangel.iglesias@upm.es (C.Á. Iglesias).

https://doi.org/10.1016/j.ijmedinf.2024.105603
Received 29 April 2024; Received in revised form 1 August 2024; Accepted 13 August 2024
Available online 19 August 2024

of this paper [14]. Specifically, about half of the studies collected in the review use Electronic Health Records (EHR) or data from longitudinal aging studies to train models and find frailty-related patterns, a solution that does not require specific reiterated actions from the patients. The review also shows that the FFP is the most common definition of frailty used in these studies and was adopted in about half of the publications. Since consistency and replicability of the results are essential to keep advancing the state of the art, we also adopted the FFP definition in this study.

This study aims to create a frailty dataset for future tests and improvements and evaluate it through seven ML models. This dataset should have a significant number of patients, relevant health-related variables, possibly non-health-related variables, a feature determining the patient's level of frailty, and the evolution of the subjects throughout the years.

To achieve all these goals, we started from the dataset of English Longitudinal Study of Ageing (ELSA) [15], a longitudinal study on the aging and well-being of the older population in the UK, constituted of nine waves of data collection, with a granularity of two years between waves. Specifically, we used wave 6 data to determine the cohort's frailty levels and trained different models to assess the dataset's detection capability. Furthermore, we used data from wave 5 to train the same type of models to predict the appearance of frailty in wave 6 two years later.

Other longitudinal studies have been used to study frailty and FFP precisely [16–20]. However, none of them has the number of participants and waves of data collection that ELSA has, which can be used in future works to improve the results and predictions.

2. Methods

In this section, we describe how we used the data from the ELSA study [15] to compute each patient's frailty level, define the FRailty English Longitudinal Study of Ageing (FRELSA) dataset, and study its characteristics.

As previously introduced, one of the most commonly accepted definitions of frailty is FFP, presented by Fried et al. in 2001 [4]. It classifies patients into three categories (non-frail, pre-frail, and frail) using five variables: weight loss, slowness, dominant hand grip strength, physical activity, and exhaustion level.

If the patient scores zero on the five criteria, it is considered *non-frail*; if they score one or two, they are in the *pre-frail* category, and if three or more of the criteria are present, they fall within the definition of *frail*. We decided to adopt this definition for our study because it is widely used in the medical field and is also the most common in previous works that tried to apply ML techniques to frailty in elderly people [14]. Hence, we needed to use the available data to classify the frailty level of the participants in the ELSA study.

ELSA is a longitudinal study on aging that was approved in 2000 and started with data collection of a first wave in 2002. The questionnaire covers general health, disability, life expectancy, economic position and resources in old age, retirement age, social networks, social support and participation at older ages, and household and family structure [21]. A broad range of topics has been included, such as family, work, economic issues, physical and mental health, social and psychological factors, behavior, cognition, and biology. The ELSA webpage [22]provides more details on the exact measurements performed and questionnaires posed to the subjects. Each section of the questionnaire results has been included in the feature selection process described in Sect. 3, meaning we looked for variables correlated with frailty in all the possible topics present in the ELSA data. The initial 2002 cohort consisted of 12099 individuals, whose ages ranged from 50 to 100, with a mean of 65. Every two years, a new data collection wave starts. To keep the age representation balanced, we follow up with all the eligible participants of the previous wave and include a refreshment cohort every time. The study stopped in 2019 after nine waves of data collection.

ELSA data is available under request for research purposes. More information is available on the ELSA website [22], including the procedure to obtain the data.

The FFP test is not directly present in the ELSA study, and no specific section is dedicated to frailty. Fortunately, in some of the data collection waves, it is possible to derive the frailty of most participants, which is what we did under the guidance of geriatric experts. It follows a description of the definition process of the FFP variable, summed up in Table 1.

In waves 2, 4, and 6, a nurse visited all willing participants of the cohort, collecting data essential to compute the FFP, such as Body Mass Index (BMI) and grip strength. In waves 8 and 9, nurse visits were also performed, but on mutually exclusive sets of the cohort; thus, we selected wave 6 data to work and label the participants using Fried's criteria, obtaining a bigger cohort. In wave 6, there were 10601 participants, of which 8054 underwent the nurse visit and were eligible to have the frailty level computed.

The first criterion of the FFP definition is *weight loss* in the last year. Unfortunately, the ELSA data do not include such information, and weight was only measured during the nurse visit, so even using data from the previous wave (wave 5, two years earlier) would not help. Hence, we translated the first criterion to being underweight (BMI < 18.5), the best available solution and had already been validated in other studies [23].

The second and third criteria are *slowness* and *dominant hand grip strength*, measured in every wave and the nurse visit, respectively; hence, the definition stayed precisely the same as the one Fried et al. gave [4]. Walking speed was measured only on participants 60 or older, which forced us to drop all the participants 50 to 59.

The fourth criterion is *physical activity level*, measured by the number of calories consumed weekly. We translated it using three questions about how often the participant does vigorous/moderate/mild sports or activities. For this criterion not to be met, the participant needs to answer at least "one to three days a month" to at least two questions or respond with a higher frequency to at least one.

The last criterion is *exhaustion*, measured using two questions from the Center for Epidemiological Studies-Depression (CES-D) scale [24], about how often in the previous week subjects felt like everything they did was an effort and they could not get going. To meet the criterion, the participant must answer "more than three days" to either one of the questions. The ELSA data contain the whole CES-D questionnaire, but instead of asking for frequency during last week, they were turned into Yes or No questions. Hence, if the ELSA participant answered "Yes" to either one of the questions, the criterion is met. The complete comparison of the two definitions (FFP and FRELSA) can be seen in Table 1. More information on the FFP criterion distribution among participants can be found in Appendix A

Because of missing data and measurements that the participants refused to provide, 5303 of the eligible 8054 patients visited could have their FFP computed. The distribution of participants is as follows: 2772 non-frail (0 criteria met), 2128 pre-frail (1 or 2 criteria met), and 403 frail (3 or more criteria met) subjects from wave 6. This kind of sample size is very satisfactory for the training of ML models, and allows for good generalization of the models' task [25].

We then added data from wave 5, collected two years earlier, filtering all the participants and keeping only the ones who were still in the study during wave 6 and were among the 5303 who could have their frailty label computed. We obtained 5135 eligible participants, meaning only 168 subjects we could label were absent in wave 5. This addition to the dataset allows us to evaluate the data through two models: frailty detection (using data from wave 6) and frailty prediction (using data from wave 5) models, described in the next section.

In Table 2, wave 6 and wave 5 data are divided by frailty levels. Demographic data for each category, such as sex, age, education, marital status, and self-reported health, are reported in percentage relative to each frailty category (computed with wave 6 data). As explained in

Table 1
A comparison of the original definition of FFP and the adaptation made using ELSA data.

	FFP		**FRELSA**	
	Male	**Female**	**Male**	**Female**
Weight loss	> 4.5 Kg or > 5% in the last year		BMI < 18.5	
Slowness	H ≤ 173 cm →≤ 0.653 m/s H > 173 cm →≤ 0.762 m/s	H ≤ 159 cm →≤ 0.653 m/s H > 159 cm →≤ 0.762 m/s	H ≤ 173 cm →≤ 0.65 m/s H > 173 cm →≤ 0.76 m/s	H ≤ 159 cm →≤ 0.65 m/s H > 159 cm →≤ 0.76 m/s
Hand grip	BMI ≤ 24 →≤ 29 kg BMI 24.1 ÷ 26 →≤ 30 kg BMI 26.1 ÷ 28 →≤ 31 kg BMI ≥ 28 →≤ 32 kg	BMI ≤ 23 →≤ 17 kg BMI 23.1 ÷ 26 →≤ 17.3 kg BMI 26.1 ÷ 29 →≤ 18 kg BMI ≥ 29 →≤ 21 kg	BMI ≤ 24 →≤ 29 kg BMI 24.1 ÷ 26 →≤ 30 kg BMI 26.1 ÷ 28 →≤ 31 kg BMI ≥ 28 →≤ 32 kg	BMI ≤ 23 →≤ 17 kg BMI 23.1 ÷ 26 →≤ 17.3 kg BMI 26.1 ÷ 29 →≤ 18 kg BMI ≥ 29 →≤ 21 kg
Phys Act	< 383 kcal/week	< 270 kcal/week	Answering *"One to three days a month"* to at least two of the following questions, or with higher frequency to at least one: How often do you do: - *Vigorous sports or activities* - *Moderate sports or activities* - *Mild sports or activities*	
Exhaustion	Answering *"More than 3 days"* to either question: How often in the last week did you feel this way: - *Everything I did was an effort* - *I could not get going*		Answering *"Yes"* to either question: Did you feel this way during the last week: - *Everything I did was an effort* - *I could not get going*	

Table 2
Wave 6 and 5 data demographics divided by frailty level.

	W6 (5303)			**W5** (5135)		
	non-frail (2772)	pre-frail (2128)	frail (403)	non-frail (2685)	pre-frail (2066)	frail (384)
Sex						
Male	49.5%	42.6%	39.0%	49.3%	42.6%	39.3%
Female	50.5%	57.4%	61.0%	50.7%	57.4%	60.7%
Age						
58 - 59	0.0%	0.0%	0.0%	11.7%	7.7%	5.0%
60 - 69	63.9%	43.6%	23.8%	59.8%	42.6%	23.2%
70 - 79	31.6%	38.3%	36.7%	25.9%	36.4%	38.5%
80 - 89	4.4%	16.1%	32.3%	2.6%	12.4%	28.6%
90+	0.1%	2.0%	7.2%	0.0%	0.9%	4.7%
Ed Qualification						
Higher ed. degree	42.4%	28.6%	17.3%	43.8%	29.5%	18.2%
Vocational ed.	27.9%	26.0%	22.1%	28.8%	26.7%	23.2%
Foreign/other	10.3%	13.4%	12.9%	10.6%	13.8%	13.5%
No qualification	16.1%	29.0%	42.9%	16.6%	29.9%	45.1%
Unknown	3.2%	3.0%	4.7%	0.1%	0.1%	0.0%
Marital status						
Married/partner	73.3%	61.1%	49.8%	74.0%	62.8%	49.7%
Not married/divorced	15.2%	16.0%	19.4%	14.8%	16.1%	19.0%
Widowed	11.4%	21.2%	34.0%	10.3%	19.1%	31.3%
Health (self-rep.)						
Excellent	15.4%	6.9%	0.5%	17.4%	8.2%	2.6%
Very good	39.3%	24.3%	7.4%	40.9%	27.3%	8.1%
Good	33.5%	35.6%	23.1%	32.2%	36.4%	27.9%
Fair	10.6%	26.8%	43.2%	8.3%	22.3%	36.7%
Poor	1.1%	6.4%	25.8%	0.6%	5.6%	24.2%
Unknown	0.0%	0.0%	0.0%	0.7%	0.3%	0.5%

Sec. 2, FFP could only be computed on people from wave 6 who were 60 or older, because of the walking speed, which was not tested on participants aged 50 to 59. Educational qualification data were unavailable for wave 6, so the values were taken from wave 5, assuming no participant obtained a higher qualification during the two years between waves. People only present in wave 6 fall under *Unknown*.

3. Results

To evaluate the frailty dataset FRELSA, we trained different architectures of classification models on the data to detect or predict the appearance of frailty. To do so, we grouped the classes of frail and pre-frail so that the dataset would be balanced (52% non-frail and 48% pre-frail/frail in waves 5 and 6), effectively leaving us with a binary classification problem. This choice is consistent with the task of helping clinicians with early detection and prediction of the first stages of frailty, which is one of the goals of the lines of future work following up on this study.

Feature selection is vital to obtain good results from the ML models. We did not want to use a wrapper or embedded method for feature selection because they are model-dependent, while we wanted to keep the process model-agnostic. The goal of this feature selection phase is not only to get the best performance out of the models but also to understand which features, among the thousands we have at our disposal,

Table 3
Metrics of the 10-fold cross-validation results for the detection models.

Wave 6	**Accuracy**	Std dev	**Precision**	Std dev	**Recall**	Std dev	**F-score**	Std dev	**AUROC**	Std dev
SVM lin	0.737	0.022	0.744	0.026	0.733	0.025	0.732	0.024	0.814	0.018
SVM rbf	**0.741**	0.020	0.748	0.018	0.736	0.022	0.736	0.022	0.810	0.024
DT	0.705	0.023	0.710	0.024	0.700	0.024	0.700	0.024	0.766	0.026
RF	0.734	0.017	0.738	0.017	0.730	0.017	0.730	0.018	0.810	0.012
LR	0.740	0.020	0.743	0.021	**0.737**	0.021	**0.737**	0.020	**0.817**	0.018
MLP	0.683	0.018	0.682	0.019	0.682	0.019	0.682	0.019	0.740	0.022
VC	0.739	-	**0.784**	-	0.625	-	0.700	-	0.734	-

Table 4
Metrics of the 10-fold cross-validation results for the prediction models.

Wave 5	**Accuracy**	Std dev	**Precision**	Std dev	**Recall**	Std dev	**F-score**	Std dev	**AUROC**	Std dev
SVM lin	0.731	0.015	0.742	0.016	0.726	0.013	0.725	0.014	**0.802**	0.009
SVM rbf	0.731	0.016	0.742	0.015	0.725	0.014	0.724	0.015	0.801	0.019
DT	0.702	0.026	0.708	0.025	0.696	0.025	0.695	0.026	0.755	0.018
RF	0.730	0.018	0.733	0.019	0.727	0.018	0.727	0.018	0.798	0.024
LR	**0.737**	0.015	0.742	0.016	**0.733**	0.016	**0.733**	0.016	**0.802**	0.017
MLP	0.668	0.016	0.667	0.016	0.666	0.015	0.666	0.015	0.723	0.015
VC	0.730	-	**0.783**	-	0.600	-	0.680	-	0.724	-

are the most relevant to the detection of frailty, with the definition we used. Also, because many inputs are missing in the data, we could not use classic statistic methods such as the chi-squared statistics [26], the ANOVA F-value [27], or mutual information [28], since they require complete data. Hence, we landed on using a relief-based method [29], a class of filter methods for feature selection capable of capturing feature interactions. Specifically, we used the MultiSURF algorithm, proposed in 2018 by Urbanowicz et al. [30] as a refined relief-based feature selection method for bioinformatics data. It ranks the importance of all the available features using a nearest-neighbor logic, giving particular importance to the features that determine a change in the target variable of two similar patients. More information on the ranking of the features can be found in Appendix B.

As mentioned, the ELSA dataset lacks data for various reasons, including the participants refusing to answer specific questions. All missing data are encoded in the ELSA dataset as negative numbers. Although the MultiSURF algorithm [30] used for the feature selection process can handle missing values, the ML baseline models described in the following paragraph work best with datasets that are not missing any entry. Since all the variables are treated as categorical features, we filled all the missing values in the dataset with the mode value of the feature. This is the most straightforward method to deal with missing inputs in categorical features and one of the most common [31].

To create a baseline of preliminary results, we selected six different standard ML architectures to train on the FRELSA data: Support Vector Machine (SVM) with linear and Radial Basis Function (RBF) kernels, a Linear Regression (LR), a Decision Tree (DT), a Random Forest (RF), and a Multi Layer Perceptron (MLP) classifiers, and lastly, we combined them with a Voting Classifier (VC).

Firstly, we trained the classifiers on wave 6 data only, and we will call them *detection models* since they are trained to detect frailty with data contemporary to the ones we used to compute FFP. The same architectures were then trained with wave 5 data to obtain *prediction models*. Since two years pass between waves, ideally, these predictors should be able to anticipate the appearance of (pre-)frailty within a 24-month horizon. After considering score metrics and the trade-off between computational complexity, we decided to use the first 50 features of the ranking given by the MultiSURF algorithm for detection and prediction models. All these models used 10-fold cross-validation, and parameters were optimized using a grid search algorithm, with the models' F-score as a reference metric. More information on the model's parameters can be found in Appendix C.

The code was written in Python using the Scikit-learn library [32], and is available in GitHub.[1] To run the code in the repository properly, you must have previously obtained the ELSA data, which are available for researchers upon request [22]. The direct association between the ELSA patients' unique identification variable (*"idauniq"*) and their frailty level can be found in the same repository.

Tables 3 and 4 collect the performance metrics of all the detection and prediction models, respectively. The results presented in the table are the averages of the 10-fold cross-validation training process. Standard deviations are reported as well. Overall, the simplest LR is the most consistent architecture, having the best F-score, standard deviation, and Area Under the Receiving Operating Characteristic (AUROC) in detection and prediction cases.

The MLP scored the worst overall F-score and AUROC due to the limited choice of available parameters and the basic training technique. The neural network architectures given to the grid search optimization algorithm were not the most elaborate, and there is room to improve the performance of this model in particular. However, these classifiers aim to evaluate the frailty label created in this study, described in Sect. 2, and not to obtain the best possible metrics. This way, a baseline of results is set, which will be improved with future work, as described in the following section.

4. Conclusions

The results of this study show that the ELSA study [15] data are suitable for studying frailty, specifically for predicting the appearance of FFP. The work presented in this paper created a new dataset starting from wave 6 and wave 5 of the ELSA data, FRELSA, which contains a new FFP label. Moreover, a ranking of the best features to detect and predict the appearance of frailty was obtained through the MultiSURF feature selection algorithm [30]. It was also shown that it is possible to train ML models to detect and predict the appearance of FFP in older populations. Detection and prediction within a two-year horizon model perform indistinguishably, which implies that the data and questionnaire used contain information on frailty long before its actual appearance.

Although the results are promising, the obtained metrics leave room for improvement, especially by designing and training a more sophisticated architecture. Moreover, all the models were trained with the first

[1] https://github.com/gsi-upm/FRELSA.

50 ranking features obtained by the MultiSURF method. Still, better predictive performance could be achieved with more model-specific feature selection.

One limitation of this study is the use of self-reported data. This type of data is unreliable because it is easily influenced by external factors such as the participant's state of mind or the way the question is asked. This is especially true of long questionnaires such as the ELSA. In contrast to the few objective measurements in the study, most of the data collected are questionnaire-type. In particular, two of the five variables we used to calculate FFP, exhaustion and physical activity level, are calculated from self-reported data and may have affected the accuracy of participants' labeling.

A limiting factor of this study, connected to the one mentioned above, is the missing nurse visit in wave 5. Because BMI and grip strength tests are unavailable, FFP cannot be consistently computed from wave 5. Having a frailty level at wave 5 would allow for a study of the evolution of the participants' frailty status. For instance, we could exclude from the training phase of the prediction models the subjects already pre-frail at wave 5 data collection time and obtain models for the more specific task of frailty onset in healthy subjects.

Another possible improvement for future work involves the explainability of the ML models, which was not considered in this study. Explainability is essential in the field of healthcare AI to favor the collaboration of technical and medical operators and avoid ethical and legal issues [33]. Therefore, moving forward in the topic of ML for the prediction of frailty, the developers should make an effort to consider the medical perspective and implement some tools to provide a human explanation of the models' results.

Finally, a line of work to further improve the FRELSA dataset is the addition of other frailty definitions. This is no trivial task, not only in terms of definition and computation but also of comparison with FFP. The other most common frailty definition used in these cases is the eFI [6], but it is a different approach to the frailty concept, and the classification problems resulting from such definition are very different and difficult to compare to the ones presented in this article.

Summary table

What was already known on the topic:

- Frailty is an increasingly common age-related syndrome associated with multi-morbidity.
- Fried's Frailty Phenotype (FFP) is one of the most common and widely accepted definitions of frailty.
- Machine Learning (ML) techniques are one of the most promising paths to support clinicians in frailty detection and prediction.
- ELSA is a relevant longitudinal study on aging in the UK with 9 waves of data collection of various natures.

What this study added to our knowledge:

- FRELSA is a new dataset generated starting from ELSA data using the FFP frailty definition.
- FRELSA can be used to train ML architectures for frailty detection and prediction.
- multiSURF algorithm was used to rank all the features ordered by relevance for frailty detection and prediction.
- Seven ML architectures were used to evaluate the dataset and create a baseline of results.

List of acronyms

AUROC	Area Under the Receiving Operating Characteristic
BMI	Body Mass Index
CES-D	Center for Epidemiological Studies-Depression
CFS	Clinical Frailty Scale
DT	Decision Tree
eFI	electronic Frailty Index
EHR	Electronic Health Records
ELSA	English Longitudinal Study of Ageing
FFP	Fried's Frailty Phenotype
FRELSA	FRailty English Longitudinal Study of Ageing
FTS	Frailty Trait Scale
LR	Linear Regression
ML	Machine Learning
MLP	Multi Layer Perceptron
RBF	Radial Basis Function
RF	Random Forest
SVM	Support Vector Machine
VC	Voting Classifier

Funding

Funding was received for this work.

All of the funding sources for the work described in this publication come from the AROMA / MIRATAR project, grant TED2021-132149B-C42 funded by MICIU/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR.

CRediT authorship contribution statement

Matteo Leghissa: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Álvaro Carrera:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization. **Carlos Á. Iglesias:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We want to thank all the MIRATAR (grant TED2021-132149B-C42 funded by MICIU/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR) project participants and colleagues from CIBER (Centro de Investigación Biomédica en Red), UCLM (Universidad de Castilla - La Mancha), and UC3M (Universidad Carlos III de Madrid).

Appendix A. Frailty criteria

This appendix provides information about the FFP criterion distribution of the FRELSA data. Specifically, in Table A.5 pre-frail and frail participants are divided by sex and age, and the percentages of frailty criteria they met are reported.

We can observe that weight loss seems to be the least relevant criterion in all the categories. On the other hand exhaustion is one of the most relevant, although its rate decreases with age, in both pre-frail and frail, males and females. The subjectivity of the matter, and the fact that participants in the younger age ranges might still be working, certainly affect the results. It would be interesting to know the impact of the self reported data methodology (i.e. questionnaire answers) on these distributions. This is one of the limitations of this work, which were discussed in Sect. 4.

Table A.5
Fried's frailty phenotype's criteria of pre-frail and frail participants, reported in percentages.

	Sex	Age	Weight loss	Slowness	Grip strength	Phys act	Exhaustion
Pre-frail	Male	60-69 (380)	0.79%	22.11%	39.74%	12.37%	46.05%
		70-79 (361)	0.28%	26.04%	60.11%	9.7%	30.47%
		80-89 (147)	0.68%	29.25%	77.55%	12.24%	22.45%
		90+ (18)	0.0%	27.78%	94.44%	16.67%	16.67%
	Female	60-69 (548)	2.37%	22.99%	44.34%	7.48%	46.9%
		70-79 (455)	2.42%	33.63%	50.11%	7.47%	40.0%
		80-89 (195)	0.51%	46.15%	64.62%	8.21%	31.28%
		90+ (24)	0.0%	41.67%	83.33%	8.33%	16.67%
Frail	Male	60-69 (38)	2.63%	92.11%	65.79%	71.05%	97.37%
		70-79 (54)	1.85%	92.59%	85.19%	66.67%	75.93%
		80-89 (55)	3.64%	90.91%	94.55%	60.0%	78.18%
		90+ (10)	0.0%	90.0%	100.0%	60.0%	70.0%
	Female	60-69 (58)	3.45%	89.66%	77.59%	51.72%	98.28%
		70-79 (94)	4.26%	96.81%	95.74%	37.23%	85.11%
		80-89 (75)	4.0%	96.0%	93.33%	44.0%	80.0%
		90+ (19)	10.53%	89.47%	94.74%	78.95%	73.68%

Appendix B. Input variables

Some of the variables used in the models' training are collected in this appendix. Specifically, Table B.6 collects the first 15 features for the wave 6 detection problem, ranked by the MultiSURF feature selection algorithm [30]. The same goes for Table B.7, in which wave 5 prediction variables are collected and described, using the name and description directly from the ELSA documentation.

Table B.6
Variables ranked by MultiSURF algorithm for wave 6 detection data. The name of the variable and the description are from the ELSA documentation.

Name	Description
hemob96	Mobility: whether said had none of the listed difficulties (walking 100 yds, sitting 2 hrs, getting up from a chair, climbing stairs, stooping kneeling or crouching, arm above the shoulder, pulling or pushing large obj, lifting weights, picking up coin).
HEBowC	Screening: used bowel testing kit.
hemobcs	Mobility: difficulty climbing several flights stairs without resting.
hemobst	Mobility: difficulty stooping, kneeling or crouching.
helwk	Whether has self-reported health problem/disability that limits paid work.
HeFunc	Difficulty walking 1/4 mile unaided.
indager	Definitive age variable collapsed at 90+ to avoid disclosure.
headl96	Whether said had none of the listed difficulties (dressing, walking, bathing, eating, getting in and out of bed, using the toilet, using a map, cooking, shopping, making calls, communicating, taking medications, housekeeping, managing money).
fastelig	BLOOD: Eligible for a fasting blood sample.
heill	Whether has a self-reported long-standing illness.
heaid96	Aids used: not use any of listed aids (cane, zimmer frame, wheelchair, scooter, special eating utensil, personal alarm, crutches)
PSAgF	Self-perceived age
HePain	Whether often troubled with pain
scptr4	Has taken a holiday abroad in the last 12 months
hemobli	Mobility: difficulty lifting or carrying weights over 10 pounds
scqolh	CASP19 scale: How often feels their health stops them doing what they want to do.
heaidca	Aids used: cane or walking stick
hemobch	Mobility: difficulty getting up from chair after sitting long periods
HOBB	At the present time do you have an internet connection
heslpx	Sleep: frequency doze off/nap in morning or afternoon
scqola	CASP19 scale: How often feels age prevents them from doing things they like
htfev	LUNG: Highest technically satisfactory value for FEV
MMFTRE2	Outcome of full tandem stand according to age
fastask	BLOOD: Respondent was asked to fast
HTPEF	LUNG: Highest technically satisfactory value for PEF
scactc	How often the respondent goes to an art gallery or museum
hehelf	Self-reported general health
NumMeds	DRUG: NumMeds
htfvc	LUNG: Highest technically satisfactory value for FVC
scint	On average, how often do you use the internet or email?
scqolo	CASP19 scale: How often feels full of energy these days
fffqqual	W3 Qualifications
DhCAg	Children Grid: Age of Child 1
PRFEV	LUNG: Predicted value for FEV
MMRROC	Chair rise: Outcome of multiple chair rises, split by age.
hebowtm	Screening: month of last bowel test
mmrrre	CHAIRRAISE: Outcome of multiple chair rises (number of rises completed)
MedCNJD	DRUG: Are they taking or using any medicines, pills, syrups, ointments, puffers or injections prescribed for them by a doctor or a nurse?
IasPa	Period code: How much do you receive from the state pension
scactd	How often the respondent goes to the theatre, a concert or the opera
hohavpc	Durables owned - computer
scchdt	How often the respondent sends or receives text messages to/from their children

Table B.6 (*continued*)

Name	Description
scacta	How often the respondent goes to the cinema
scqols	CASP19 scale: How often feels the future looks good to them
IAWork	Did respondent or spouse do any work for pay in the last year
iasiisa	Savings & investments (respondent or spouse): ISA
ExLo80	Expectation (%) that they will live to [age] [depends on current age]
PRPEF	LUNG: Predicted value for PEF
PScedC	Whether felt their sleep was restless during past week
PRFVC	LUNG: Predicted value for FVC

Table B.7
Variables ranked by MultiSURF algorithm for wave 5 prediction data. The name of the variable and the description are from the ELSA documentation.

Name	Description
hemob96	Mobility: whether said had none of listed difficulties (walking 100 yd, sitting 2 hrs, getting up from chair, climbing stairs, stooping kneeling or crouching, arm above shoulder, pulling or pushing large obj, lifting weights, picking up coin).
hemobcs	Mobility: difficulty climbing several flights stairs without resting.
helwk	Whether has self-reported health problem/disability that limits paid work.
indager	Definitive age variable collapsed at 90+ to avoid disclosure.
herpd	Shortness of breath: whether has when hurrying on level/walking up slight hill.
heactb	Frequency does moderate sports or activities.
hemobst	Mobility: difficulty stooping, kneeling or crouching.
heill	Whether has a self-reported long-standing illness.
scpt06	Respondent uses the internet and/or email.
heacta	Frequency does vigorous sports or activities.
headl96	Whether said had none of the listed difficulties (dressing, walking, bathing, eating, getting in and out of bed, using the toilet, using a map, cooking, shopping, making calls, communicating, taking medications, housekeeping, managing money).
mmwlkb	Time taken for second walk
mmwlka	Time taken for first walk
HeFunc	Difficulty walking 1/4 mile unaided.
scqolh	CASP19 scale: How often feels their health stops them doing what they want to do.
palevel	Physical activity summary
IAWork	Did respondent or spouse do any work for pay in the last year
scqolo	CASP19 scale: How often feels full of energy these days
HePain	Whether often troubled with pain
scqola	CASP19 scale: How often feels age prevents them from doing things they like
DhCAg	Children Grid: Age of Child 1
breths	Breathlessness, MRC respiratory questionnaire
fffqual	W3 Qualifications
hemobli	Mobility: difficulty lifting or carrying weights over 10 pounds
hemobch	Mobility: difficulty getting up from chair after sitting long periods
sptro1	Whether respondent aged 65+ gets lifts from family/friends not living with them
hehelf	Self-reported general health
scpt04	Respondent has taken a holiday abroad in the last 12 months
hehpa	Functioning: whether ever has help with mobility, ADL, IADL
scdeac	How well the following describes the respondent: active
hohavpc	Durables owned - computer
cfind	Total Cognitive Index (memory + executive)
scfeac	During the past 30 days to what degree the respondent feels active
heaid96	Aids used: not use any of listed aids (cane, zimmer frame, wheelchair, scooter, special eating utensil, personal alarm, crutches)
scactc	How often the respondent goes to an art gallery or museum
w5edqual	Highest Educational Qualification at ELSA W5
CASP19	CASP 19 index
wpact96	Activities during last month: none of these (paid work, self-employment, voluntary work, cared for someone, looked after home or family, attended a formal educational or training course)
hedent	Whether has natural teeth, dentures or neither
cfmeind	Memory function index
HeCda	Respondent gets pain or discomfort in either of the legs which comes on when they walk
HeBdiaAR	Wave when diagnosis of arthritis was first reported
heaidca	Aids used: cane or walking stick
hehps96	None of listed therapists/classes used to help with physical functioning difficulties (occupational therapy or physiotherapy, chiropody treatment, exercise class, other treatment or assistance)
hehsm96	Said no treatment to help with physical functioning difficulties
wpactw	Derived: prioritized value for work status in last month
cfani	Number of animals mentioned (CAPI interview)
sptro96	Whether respondent aged 65+ does not use any of the above means of transport (lifts from family/friends not living with them, taxi, door-to-door community transport, transport provided by hospital/day center/etc, transport provided by care home)
IasPa	Period code: How much do you receive from the state pension
pscedb	Whether felt everything they did during past week was an effort

Appendix C. Models' parameters

In this appendix, the models' hyperparameters are collected, specifically in Table C.8 are the detection models, and in Table C.9 the prediction models. These parameters were obtained through a grid search, with all the available parameter choices reported in said tables.

Table C.8
Best parameters grid search of detection all the detection models. Bold parameter values are the selected best.

Model	Parameter	Values		
SVM lin	C	0.1	1	10
SVM rfb	C	0.1	**1**	10
	gamma	0.01	**0.1**	-
DT	max_depth	**5**	10	20
RF	max_depth	**5**	10	20
	n_estimators	20	50	**100**
LR	C	**0.1**	1	10
MLP	activation	**tanh**	relu	-
	alpha	0.0001	**0.001**	-
	hidden_layer_sizes	(100, 50)	**(100, 75, 25)**	(100, 100, 75, 50, 25)

Table C.9
Best parameters grid search of detection all the prediction models. Bold parameter values are the selected best.

Model	Parameter	Values		
SVM lin	C	0.1	1	10
SVM rfb	C	0.1	**1**	10
	gamma	**0.01**	0.1	-
DT	max_depth	**5**	10	20
RF	max_depth:	5	**10**	20
	n_estimators	20	50	**100**
LR	C	0.1	1	**10**
MLP	activation	tanh	**relu**	-
	alpha	**0.0001**	0.001	-
	hidden_layer_sizes	(100, 50)	**(100, 75, 25)**	(100, 100, 75, 50, 25)

References

[1] L. Han, A. Clegg, T. Doran, L. Fraser, The impact of frailty on healthcare resource use: a longitudinal analysis using the clinical practice research datalink in England, Age Ageing 48 (5) (2019) 665–671, https://doi.org/10.1093/ageing/afz088.

[2] E.O. Hoogendijk, J. Afilalo, K.E. Ensrud, P. Kowal, G. Onder, L.P. Fried, Frailty: implications for clinical practice and public health, Lancet 394 (10206) (2019) 1365–1375, https://doi.org/10.1016/s0140-6736(19)31786-6.

[3] L. Gené Huguet, M. Navarro González, B. Kostov, M. Ortega Carmona, C. Colungo Francia, M. Carpallo Nieto, A. Hervás Docón, R. Vilarrasa Sauquet, R. García Prado, A. Sisó-Almirall, Pre frail 80: multifactorial intervention to prevent progression of pre-frailty to frailty in the elderly, J. Nutr. Health Aging 22 (10) (2018) 1266–1274, https://doi.org/10.1007/s12603-018-1089-2, https://www.sciencedirect.com/science/article/pii/S1279770723010655.

[4] L.P. Fried, C.M. Tangen, J. Walston, A.B. Newman, C. Hirsch, J. Gottdiener, T. Seeman, R. Tracy, W.J. Kop, G. Burke, M.A. McBurnie, Frailty in older adults: evidence for a phenotype, J. Gerontol., Ser. A 56 (3) (2001) M146–M157, https://doi.org/10.1093/gerona/56.3.M146.

[5] K. Rockwood, X. Song, C. MacKnight, H. Bergman, D.B. Hogan, I. McDowell, A. Mitnitski, A global clinical measure of fitness and frailty in elderly people, CMAJ, Can. Med. Assoc. J. 173 (5) (2005) 489–495, https://doi.org/10.1503/cmaj.050051.

[6] A. Clegg, C. Bates, J. Young, R. Ryan, L. Nichols, E. Ann Teale, M.A. Mohammed, J. Parry, T. Marshall, Development and validation of an electronic frailty index using routine primary care electronic health record data, Age Ageing 45 (3) (2016) 353–360, https://doi.org/10.1093/ageing/afw039.

[7] F.J. García-García, L. Carcaillon, J. Fernandez-Tresguerres, A. Alfaro, J.L. Larrion, C. Castillo, L. Rodriguez-Mañas, A new operational definition of frailty: the frailty trait scale, J. Am. Med. Dir. Assoc. 15 (5) (2014) 371.e7–371.e13, https://doi.org/10.1016/j.jamda.2014.01.004.

[8] L.J. Gleason, E.A. Benton, M.L. Alvarez-Nebreda, M.J. Weaver, M.B. Harris, H. Javedan, FRAIL questionnaire screening tool and short-term outcomes in geriatric fracture patients, J. Am. Med. Dir. Assoc. 18 (12) (2017) 1082–1086, https://doi.org/10.1016/j.jamda.2017.07.005.

[9] D.L. Vetrano, A. Zucchelli, G. Onder, L. Fratiglioni, A. Calderón-Larrañaga, A. Marengoni, E. Marconi, I. Cricelli, P. Lora Aprile, R. Bernabei, C. Cricelli, F. Lapi, Frailty detection among primary care older patients through the primary care frailty index (pc-fi), Sci. Rep. 13 (1) (Mar. 2023), https://doi.org/10.1038/s41598-023-30350-3.

[10] M. Raîche, R. Hébert, M.-F. Dubois, Prisma-7: a case-finding tool to identify older adults with moderate to severe disabilities, Arch. Gerontol. Geriatr. 47 (1) (2008) 9–18, https://doi.org/10.1016/j.archger.2007.06.004.

[11] P. Botolfsen, J.L. Helbostad, R. Moe-nilssen, J.C. Wall, Reliability and concurrent validity of the expanded timed up-and-go test in older people with impaired mobility, Physiother. Res. Int. 13 (2) (2008) 94–106, https://doi.org/10.1002/pri.394.

[12] B. Vellas, L. Balardy, S. Gillette-Guyonnet, G. Abellan Van Kan, A. Ghisolfi-Marque, J. Subra, S. Bismuth, S. Oustric, M. Cesari, Looking for frailty in community-dwelling older persons: the gerontopole frailty screening tool (gfst), J. Nutr. Health Aging 17 (7) (2013) 629–631, https://doi.org/10.1007/s12603-013-0363-6.

[13] M. Wleklik, I. Uchmanowicz, E.A. Jankowska, C. Vitale, M. Lisiak, M. Drozd, P. Pobrotyn, M. Tkaczyszyn, C. Lee, Multidimensional approach to frailty, Front. Psychol. 11 (2020), https://doi.org/10.3389/fpsyg.2020.00564.

[14] M. Leghissa, A. Carrera, C.A. Iglesias, Machine learning approaches for frailty detection, prediction and classification in elderly people: a systematic review, Int. J. Med. Inform. 178 (2023) 105172, https://doi.org/10.1016/j.ijmedinf.2023.105172.

[15] J. Banks, G.D. Batty, J. Breedvelt, K. Coughlin, R. Crawford, M. Marmot, J. Nazroo, Z. Oldfield, N. Steel, A. Steptoe, M. Wood, P. Zaninotto, English Longitudinal Study of Ageing: Waves 0-9, 1998-2019, 2023.

[16] F.J. Garcia-Garcia, G. Gutierrez Avila, A. Alfaro-Acha, M.S. Amor Andres, M. De Los Angeles de la Torre Lanza, M.V. Escribano Aparicio, S. Humanes Aparicio, J.L. Larrion Zugasti, M. Gomez-Serranillo Reus, F. Rodriguez-Artalejo, L. Rodriguez-Manas, The prevalence of frailty syndrome in an older population from Spain. The Toledo study for healthy aging, J. Nutr. Health Aging 15 (2011), https://doi.org/10.1007/s12603-011-0075-8. (Accessed 6 August 2023).

[17] L. Ferrucci, S. Bandinelli, E. Benvenuti, A. Di Iorio, C. Macchi, T. Harris, J. Guralnik, Subsystems contributing to the decline in ability to walk: bridging the gap between epidemiology and geriatric practice in the InCHIANTI study, J. Am. Geriatr. Soc. 48 (2001) 1618–1625, https://doi.org/10.1111/j.1532-5415.2000.tb03873.x. (Accessed 6 August 2023).

[18] ELSI, Brazilian longitudinal study of aging, http://elsi.cpqrr.fiocruz.br/en/home-english. (Accessed 6 August 2023), 2023.

[19] C.W. Won, S. Lee, J. Kim, D. Chon, S. Kim, C.-O. Kim, M.K. Kim, B. Cho, K.M. Choi, E. Roh, H.C. Jang, S.J. Son, J.-H. Lee, Y.S. Park, S.-G. Lee, B.J. Kim, H.J. Kim, J. Choi, H. Ga, K.J. Lee, Y. Lee, M. Kim, Korean frailty and aging cohort study (KFACS): cohort profile, BMJ Open 10 (4) (2020), https://doi.org/10.1136/bmjopen-2019-035573.

[20] TILDA, The Irish LongituDinal study on ageing, https://tilda.tcd.ie/data/, 2023. (Accessed 6 August 2023).

[21] A. Steptoe, E. Breeze, J. Banks, J. Nazroo, Cohort profile: the English longitudinal study of ageing, Int. J. Epidemiol. 42 (6) (2012) 1640–1648, https://doi.org/10.1093/ije/dys168.

[22] ELSA, English longitudinal study of ageing, https://www.elsa-project.ac.uk/, 2024. (Accessed 23 April 2024).

[23] R.S. Crow, M. Lohman, A. Titus, S. Cook, M. Bruce, T. Mackenzie, S. Bartels, J. Batsis, Association of obesity and frailty in older adults: Nhanes 1999–2004, J. Nutr. Health Aging 23 (2) (2019) 138–144, https://doi.org/10.1007/s12603-018-1138-x.

[24] L.S. Radloff, The ces-d scale: a self-report depression scale for research in the general population, Appl. Psychol. Meas. 1 (3) (1977) 385–401, https://doi.org/10.1177/014662167700100306.

[25] D. Rajput, W.-J. Wang, C.-C. Chen, Evaluation of a decided sample size in machine learning applications, BMC Bioinform. 24 (1) (Feb. 2023), https://doi.org/10.1186/s12859-023-05156-9.

[26] K. Pearson, X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, Philos. Mag. 50 (302) (1900) 157–175, https://doi.org/10.1080/14786440009463897.

[27] H. Scheffé, The Analysis of Variance, Wiley Classics Library ed. Edition, A Wiley Publication in Mathematical Statistics, Wiley-Interscience Publication, New York [u.a.], 1999, includes bibliographical references (p. 457-465) and indexes.

[28] A. Kraskov, H. Stögbauer, P. Grassberger, Estimating mutual information, Phys. Rev. E 69 (2004) 066138, https://doi.org/10.1103/PhysRevE.69.066138.

[29] R.J. Urbanowicz, M. Meeker, W. La Cava, R.S. Olson, J.H. Moore, Relief-based feature selection: introduction and review, J. Biomed. Inform. 85 (2018) 189–203, https://doi.org/10.1016/j.jbi.2018.07.014.

[30] R.J. Urbanowicz, R.S. Olson, P. Schmitt, M. Meeker, J.H. Moore, Benchmarking relief-based feature selection methods for bioinformatics data mining, J. Biomed. Inform. 85 (2018) 168–188, https://doi.org/10.1016/j.jbi.2018.07.015.

[31] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, O. Tabona, A survey on missing data in machine learning, J. Big Data 8 (1) (Oct. 2021), https://doi.org/10.1186/s40537-021-00516-9.

[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[33] J. Amann, A. Blasimme, E. Vayena, D. Frey, V. Madai, Explainability for artificial intelligence in healthcare: a multidisciplinary perspective, BMC Med. Inform. Decis. Mak. 20 (2020), https://doi.org/10.1186/s12911-020-01332-6.