Predicting depression in old age: Combining life course data with machine learning

Carlotta Montorsi[a,b,d]

carlotta.montorsi@liser.lu

Alessio Fusco[a], Philippe Van Kerm[a,b], Stèphane P.A. Bordas[c]

[a] Luxembourg Institute of Socio-Economic Research, Department of Living Conditions, 11, Porte des Sciences L-4366, Esch-sur-Alzette, Luxembourg

[b] University of Luxembourg, Department of Social Sciences, 11, Porte des Sciences L-4366, Esch-sur-Alzette, Luxembourg

[c] University of Luxembourg, Department of Engineering, 6, Avenue de la Fonte L-4364, Esch-sur-Alzette, Luxembourg

[d] Insubria University, Department of Economics, 71, via Monte Generoso 21100, Varese, Italy

Abstract

Depression in old age has negative individual and societal consequences. With ageing populations, understanding life course factors that raise the risk of clinical depression in old age may reduce healthcare costs and guide resource allocation. In this paper, we estimate the risk of self-reported depression by combining adult life course trajectories and childhood conditions in supervised machine learning algorithms. Our contribution is threefold. Using data from the Survey of Health, Ageing and Retirement in Europe (SHARE), we first implement and compare the performance of six alternative machine learning algorithms. Second, we analyse the performance of the algorithms using different life-course data configurations. While we obtain similar predictive abilities between algorithms, we achieve the highest models' performance when employing high-dimensional and less structured data. Finally, we use the SHAP (SHapley Additive exPlanations) method to extract the most decisive depressive patterns by gender. Age, health, childhood conditions, and low education predict most depression risk later in life. In addition, we identify new predictive patterns in high-frequency emotion-enhancing life events and low utilization of dental care services.

Key words: depression, life course data, machine learning, ageing population, SHARE.

JEL classification: I14, I31, C53, C55,

Highlights

- We predict old age depression risk by combining life course data and machine learning models
- Models are estimated on SHARElife data
- Models performed best when incorporating high-dimensional and less structured data
- Depression in women is more foreseeable than in men
- Our method allows uncovering under explored depression predictors such as repeated emotion-enhancing life events and dental care utilization

1 Introduction

Population ageing is one of the critical challenges of our times (United Nations, Department of Economic and Social Affairs, 2019). The share of the EU population above the age of 65 will reach almost 30% by 2050 (starting from 19.2 % in 2016); therefore, understanding well-being in old age is a priority. Several dimensions account for well-being. Mental health is a crucial aspect of it, with mental illness having detrimental individual consequences - such as a negative impact on productivity (Conti and Burton, 1994)- and bearing important costs for society – the annual cost of depression and anxiety amounts to USD 1 trillion for the global economy (Health, 2020; see also Sobocki et al., 2006).

However, to date, mental health among the elderly surprisingly received less attention than in other age groups. Due to discrimination and stigma associated with ageing, mental disorders in old age are under-treated and under-diagnosed in primary care settings (WHO, 2005). From a policy perspective, it appears crucial to provide preventive tools that could help identifying populations at risk and anticipate the onset of depression in old age.

Predicting depression is, however, a challenging task. Depression in old age is shaped in non-trivial ways by life and work events individuals are confronted with - the places where they have lived, other individuals they have met, as well as the institutions they have been exposed to (Colman and Ataullahjan, 2010; Falkingham et al., 2020). The complexity and high dimensionality of the mechanisms at play when we examine how life course experiences influence old age outcomes challenge traditional modeling techniques: it is the non-linear combination of individual biographies, institutional contexts, and possibly cultural influences that matter.

We approach the complexity of depression underlying mechanisms by adopting supervised machine learning algorithms (SML). Relying on the predictive power of SML opens possibilities for learning about how individual history influences later life outcomes and identifying key "past events" or combinations or sequences of events that may predict depression. Indeed, out-of-shelves machine learning models proved to excel at capturing complex non-linear interactions and generally outperform conventional linear prediction models (see Leist et al., 2022)

This paper's first contribution is to develop a predictive model of clinical depression in old age, combining two main ingredients: a large set of past life course data and machine learning tools. A growing number of studies applied machine learning approaches to analyse mental disorders. These studies focus on symptoms of depression (Librenza-Garcia et al., 2021), they use costly health records and invasive medical screening (Nemesure et al., 2021 and Garriga et al., 2022), or they look at post-therapeutic-treatment outcomes (Sajjadian et al., 2021). By contrast, our predictive approach relies almost exclusively on past life course histories and is based on more accessible and less invasive survey data source. To our knowledge, this paper is the first attempt to explore the predictive power of life histories in combination with machine learning predictive tools. While there is a growing literature using ML in economics and social sciences focusing on objective outcomes,[1], our paper contributes to the few examples in the litera-

[1]ML models have been used in the fields of criminal justice (see Berk, 2012), economic wellbeing measurement at a granular level using mobile data or satellite imagery (see Engstrom et al., 2016), income distribution and means test estimation in developing countries (see McBride and Nichols, 2018), high school drop outs (see Sansone, 2019) or inequality of opportunity measurement (see Brunori and Neid-

ture attempting to predict a subjective variable. [2]

The second contribution of this paper is methodological. We compare the performance of six algorithms and assess the sensitivity of the results to the structure of the underlying data. In modeling clinical depression, we gradually change the input data set configuration from highly structured and low-dimensional to unstructured and high-dimensional. The scope of creating different predictors' configurations is to assess the extent to which changing the data structure affects the predictive ability of the model and at which cost in terms of interpretability of the results. Until now, the mental well-being literature has not addressed this question, assuming that a single, structured, low-dimensional information set is the best predictor of depression. Our initial hypothesis is that functions of different complexity behave similarly as long as the data are low dimensional and well structured. However, we expect simpler models such as the logistic regression to fail to handle the increased dimensionality of the information set. More complex machine learning capabilities may overcome the data-dimensionality issue, resulting in a more accurate prediction model.

We address these questions using data from the Survey of Health, Ageing, and Retirement (SHARE). From here, we extract life course information relying on sequence analysis (see Abbott, 1995). Similar to DNA molecules representation, this approach represents an individual life history as an ordered string of characters, named *sequence*. A sequence strength relies on its holistic perspective over the life course, which enables capturing complex dynamics and life transitions.

We construct individual life sequences over six life domains from 15 to 49 years, and we extract meaningful information following the methodology proposed in Wahrendorf et al. (2013), Studer and Ritschard (2016), and Bolano and Studer (2020). Thus, on top of childhood information, demographics, and a small collection of adulthood life events, we create three different life course predictor sets of increasing dimensionality. The first

höfer, 2021).

[2] ML models have been used in the context of analysis of affective forecasting (see Wilson and Gilbert, 2005), prediction of happiness, health, and depression from a combination of high-frequency data and surveys (see Jaques et al., 2015), daily stress prediction from mobile phone and weather conditions data (Bogomolov et al., 2014), and to predict depression among university students (see Choudhury et al., 2019).

and lowest dimensional sequence representation is that of sequences' cluster membership. These clusters categorize individuals based on similarities in the states and transitions over the life sequence; this predictor set counts around 113 variables. The second set decodes life history information on the basis of four socially meaningful sequence features: the timing of life turning points (e.g. when transitioning from single to married), the sequencing of trajectory states (e.g., whether having children before marriage or vice-versa), the duration spent in each state (e.g., for how long an individual has been married) and the entropy of the trajectory (e.g., a standardized measure of the number of changes that occurred along the sequence). Overall, this set counts around 330 variables. Finally, the third set describes life trajectories with more than 800 binary variables in a highly unstructured form. Each binary variable represents a combination of age and life trajectory state.

We benchmark the predictive capacity of life course information with two different predictor sets. The first is a minimal predictor set including only demographic variables, e.g., country of residence, age, birth cohort, interview year, educational level, age at first childbirth, and migrant status. The second is the predictive capacity from clinical studies targeting depression risk (see Garriga et al., 2022).

A shortcoming of many ML approaches is the difficulty of interpreting predictions. Lack of interpretability might result from the intrinsic black-box character of ML methods such as neural networks or ensemble methods such as Gradient Boosting. As a third contribution of the paper, we follow the recent literature on "interpretable machine learning" (IML), and we use SHAP (SHapley Additive exPlanations) to interpret the predictions made from the Gradient Boosting algorithm. For both genders, we extracted acknowledged predictors of depression, such as age, health, childhood conditions and educational level. In addition, we identify two under-explored predictors for depression risk. The first is the entropy of life trajectory which refers to the number of state changes that occur in a life dimension. Individuals confronted with more life or work events display a higher entropy. This dimension has been largely neglected in the epidemiology of depression literature which has not looked directly at this lifetime measure. These findings

open new possibilities for further investigating the meaning of repeated life changes in shaping health outcomes. The second is low lifetime utilization of dental care services.

In terms of predictive performance, we find that more complex ML algorithms work as well as standard logistic models. However, we achieve slightly higher model performance when we use high dimension data combined with an off-shelf Gradient Boosting model. Compared to the minimal benchmark, the predictive accuracy, judged by Area Under the Precision-Recall curve (PR-AUC), increases by around four percentage points for all models when using life course data in the forms of high-dimensional sequences' features. This improvement confirms that the data structure matter and favors increasing the complexity of the input to identify individuals at risk for depression more precisely. Independently of the algorithm, a PR-AUC of 0.68 for females and 0.46 for males is a reliable maximum given the type of available information. Our predictive results are consistent with other studies targeting similar results (see Bogomolov et al., 2014, Garriga et al., 2022, and Bakkeli, 2022). However, in these articles, the information used for prediction comes from expensive and highly protected electronic health records or behavioral metrics, e.g., from mobile phone activity, thus less accessible by policymakers.

In addition, these studies used information measured at the time of diagnosis of depression, thereby restricting the ability to anticipate depression in at-risk populations. We rely on past life history information from survey data. Therefore, our findings have the dual advantages of using more accessible information and supporting genuinely preventive actions.

The remainder of this article is organized as follows. Section 2 describes the source survey, and section 3 outlines the methodological process. Section 4 elaborates on the predictive findings. Section 5 presents the concluding discussion.

2 Data

2.1 The sample

Our analysis draws data from the bi-annual Survey of Health, Ageing, and Retirement in Europe (SHARE). The SHARE survey has collected individual-level data on health, socio-economic status, and social and family networks of more than 123,000 individuals aged 50+ from 2004 to 2020 (Börsch-Supan, 2019). A feature of SHARE that makes it particularly suitable for our application is the retrospective questionnaire called SHARELIFE. SHARELIFE was included in the third (2008 – 2009) and the seventh waves (2017).[3] The questionnaire has modules on several individual life dimensions, such as childhood conditions, partnerships and parenting, employment trajectories, migration, housing, financial histories, and much more. SHARELIFE collects retrospective information using the so-called "life-grid approach." The life-grid approach supplements the interviewers' questions with a graphical longitudinal representation of respondents' life. The interviewer fills the grid during the interview, ending with a set of information for each respondent's age. A problem of retrospective data is the recall bias (see Havari and Mazzonna, 2015), if respondents systematically remember incorrectly the life events that happened in their past. To limit the influence of recall bias, we include only individuals aged 50 to 88 when answering the SHARELIFE interviews and exclude individuals who had difficulties responding to the retrospective questionnaire. The final respondent sample includes 63952 respondents (35946 females, 28006 males).

The countries covered are Austria, Belgium, Switzerland, Czech Republic, Germany, Denmark, Spain, France, Greece, Italy, Poland, Sweden, the Netherlands, Luxembourg, Hungary, Portugal, Slovenia, Estonia, and Croatia. The analysis is conducted by pooling these countries together and stratifying the sample by gender.

[3] In the seventh wave only respondents who did not participate in the third wave answer the questionnaire (82% of wave seven respondents Börsch-Supan, 2019)

2.2 Mental well-being measures

The outcome of interest is a binary indicator of clinical depression. We construct this measure from the 12 EURO-D items of depressive symptoms stored in the mental health module of SHARE (waves 1 to 7). The construct, face, and content validity and reliability of this measure have widely been validated by the literature (see Prince et al., 1999, Walker and Schimmack, 2008, and Diener et al., 2013). EURO-D scores range between a minimum of 0 and a maximum of 12, with a score of at least 4 indicating "clinical depression" and below 4 "no depression".[4]

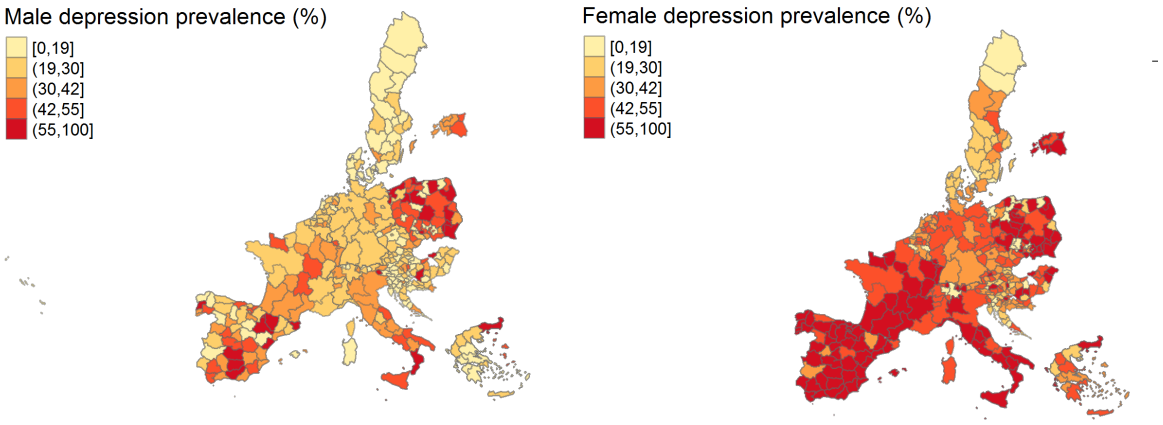
For some individual, we have repeated observations. We define them as depressed if they have at least one measure of depression over the observation period i.e., if applicable, we selected the individual observation where depression is positive. If they have more than one depression measurement, we randomly select one. This selection criterion results in a relatively high depression prevalence, 46% for females and 29% for males.

Figure 1 and Appendix A illustrate the distribution of depression prevalence within the analyzed countries at NUTS3 and NUTS2 levels, stratified by gender. In line with the literature (see Van de Velde et al., 2010), the depression prevalence appears not equally distributed.

First, females display a higher propensity to be depressed than males. Second, with the exception of Luxembourg, we observe a gradient in depression prevalence across more economically developed and less developed countries. Denmark, Sweden, and Switzerland have the lowest depression prevalence. At the top of the distribution, there are Hungary, Portugal, Poland, Spain, and Luxembourg.

[4]For more information on this depression threshold validation see Prince et al. (1999). As a sensitivity test, we also analyzed different threshold values,e.g., 3, 5, and 6. We observe a decrease in models' predictive ability at higher threshold values.

FIGURE 1: Map of depression (%) among 50+ at the NUTS3 level, by gender



We consider these differences by first stratifying the sample by gender. Thus, we include country dummies and country-age interactions.

2.3 Predictor sets

Mental disorders threatening successful aging may result from complex circumstances and events taking place throughout the entire life span (see Colman and Ataullahjan, 2010, Currie and Almond, 2011, Layard et al., 2014, Pakpahan et al., 2017). Central to life course epidemiology theories is the idea that health-related states are shaped by endogenous and exogenous forces interacting through time. The effect of these forces is different along the life cycle. There are some periods of development, so-called "sensitive," where specific experiences may exert a marked influence over future history (Bornstein, 1989).

Scholars emphasize how parental socioeconomic status, health, cognitive and non-cognitive skills, stressful situations (e.g., parental divorce or separation), and parental mental distress have strong indirect associations with later-life mental health (see Atkins et al., 2020, Flèche et al., 2021, and Zheng et al., 2021). Education is one of the most important mediators between adverse childhood experiences and mental health. A higher level of education helps moderate the negative effect of disadvantageous childhood conditions, and this effect is more substantial for women (Arpino et al., 2018). Rural-urban contextual differences may also count as depressive risk factors, but results in this area are inconclusive (Blazer et al., 1985). Access to health care services and dental care during

life received much less attention than other experiences, leaving unclear how these factors influence depression (Kisely, 2016). Housing arrangement also matters for mental well-being, where a longer duration of renting worsens well-being as opposed to owning (Vanhoutte et al., 2017). Finally, we highlight one common element rooted in mental health studies: the large gender gap characterizing depression levels. The difference persists across countries. Scholars found explanations in female hormonal fluctuations across the life course, particularly during puberty, prior to menstruation, following pregnancy, and at perimenopause. Recent evidence suggests the decisive role of differences in genetic endowment across genders. However, whether and which differences in socioeconomic life-course risk factors explain the gap remains unclear.

We start from this large body of literature to frame the selection of predictors in our SML models. Within this frame, we adopt a life course perspective and create life trajectories for all life dimensions on which information is available in SHARELIFE.

Our sets include information on childhood conditions, demographic characteristics, and a small collection of adult life event indicators that we use as control variables. We report descriptive statistics for this set in Appendix B. Qualitatively, females in our sample have been out of the labor force, have kids, and have established their households at younger ages more often than males.

Around 20% of respondents have at least one missing value in the selected childhood, demographic or life-course variables. We impute missing values separately by country to preserve differences in mean and covariance structures and encode missing values patterns in this baseline predictor set. [5] As a sensitivity analysis (available upon request), we repeated our exercise, dropping observations with missing variables; the results were unchanged.

We codify adult life trajectories following the literature on sequence analysis. Our trajectories covers years from ages 15 to 49.

Sequences are objects made up of an ordered list of successive elements chosen from

[5] We imputed missing data with the R package "missForest" (see Stekhoven and Bühlmann, 2012), which relies on an iterative method based on the Random Forest algorithm. This non-parametric algorithm has the advantage that it can handle mixed types of variables

a finite list of states, named alphabet (see Abbott, 1995). Sequences' strength relies on their holistic perspective over the life course, which enables capturing complex dynamics and life transitions. Notwithstanding the extolled potential highlighted in their first formulation, their use in scientific applications has been limited (see Studer and Ritschard, 2016, Aisenbrey and Fasang, 2010 and Liao et al., 2022).

We construct life sequences for six variables: work status, housing arrangement, family, health, residence location, and general life events. The family sequence combines information on partner history, children's history, and cohabitation history. We obtain the work, health, and housing arrangement sequences from the gateway portal harmonized sequences (gwd).[6]

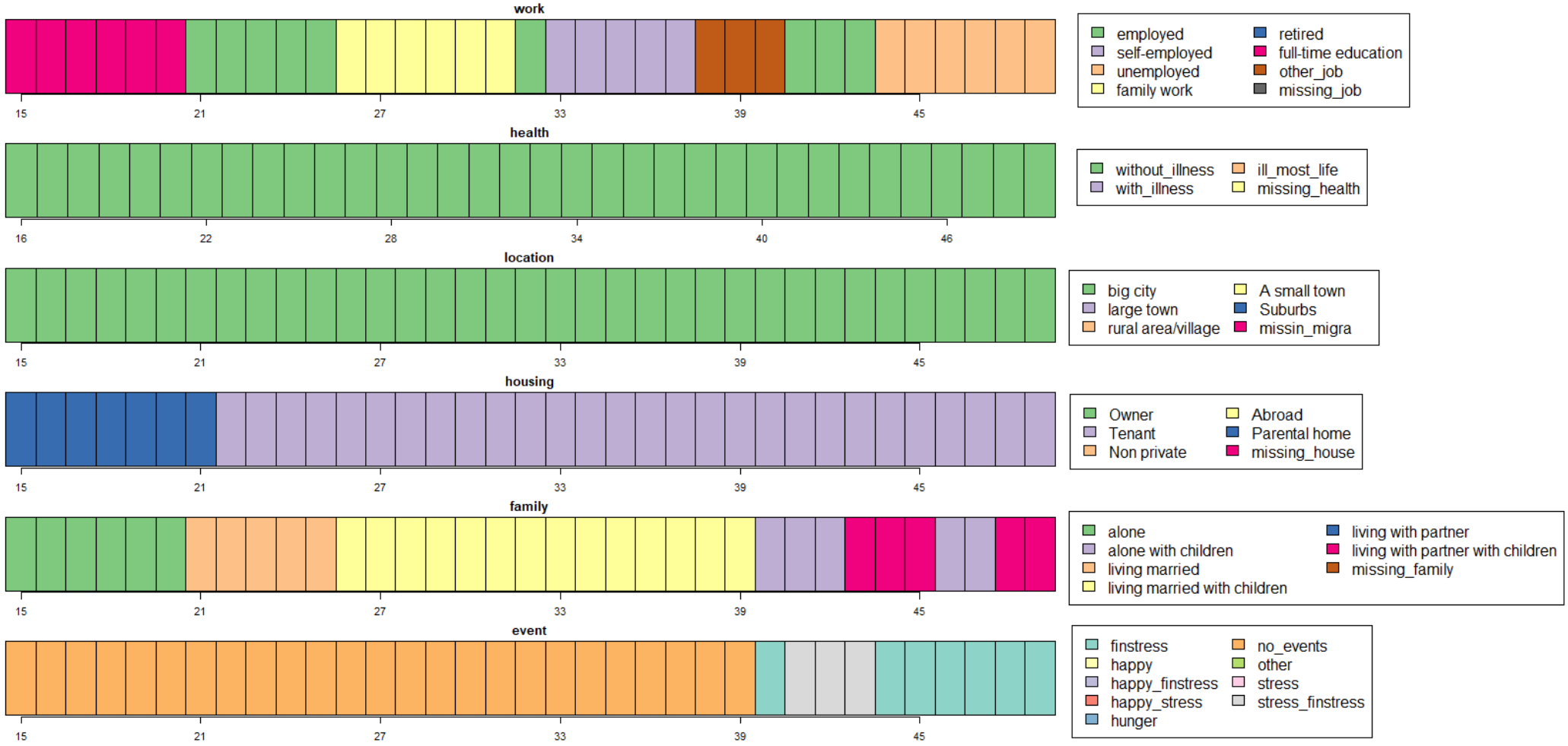
Figure 2 exemplifies the six life sequences we construct for each individual in our sample. Appendix C reports the alphabet and definitions for the six variables. We add additional states in case of missing data. Missing information in the life trajectories is rare except for the location of residence. We construct sequences and corresponding figures in R using the "TraMineR" package (see Gabadinho et al., 2011).

Different ways of coding sequences give rise to more or less structured predictor sets. The lower dimensional sequences' configuration is that of typologies or clusters, as detailed in 2.3.1. The intermediate dimensionality derives from sequences' features that we illustrate in section 2.3.2. Finally, the more unstructured decoding case results when creating m binary variables of potential states for each sequence's year. For example, each year, the family sequence can assume six states, i.e., married, single, with a partner, and with or without children. Thus, for each year of the sequence, we create a binary variable "Age 15 with a partner," "Age 15 with partner and children," "Age 15 married," "Age 15 married with children [...]", "Age 16 ..." indicating whether or not the observed person is in one of these states at a given age.

Finally, we pre-processed all data before feeding them into the algorithms (for more information on pre-processing, see Appendix D).

[6]This analysis uses information from the Harmonized SHARE Life History dataset and codebook, Version B as of February 2020, developed by the Gateway to Global Aging Data in collaboration with the University of Duesseldorf. The development of the Harmonized SHARE Life History was funded by the National Institute on Ageing (R01 AG030153, RC2 AG036619, R03 AG043052)

FIGURE 2: Representation of six life dimensions for an individual. Each rectangle represents an age and each colour represents a different state



2.3.1 Sequences' cluster

Typologies or clusters are the most common sequence configuration employed in social science applications. To construct sequence clusters, we follow a standardized procedure. We start by creating individual sequences for each life course variable. Next, we assess the dissimilarities between sequences for each life dimension and create a distance matrix. Several measures exist to estimate dissimilarities among sequences (see Gabadinho et al., 2011). In our empirical application, we use the Dynamic Hamming Distance (DHD) proposed by Lesnard (see Lesnard, 2010).[7] Once we get a matrix of sequences' dissimilarities, we perform cluster analysis (Ward methods) to regroup the more similar sequences into clusters. To select the number of clusters, we used the "nbClust" package in R (see Charrad et al., 2014) and retained the solution that maximizes the silhouette index (see Kaufman and Rousseeuw, 2009). The optimal number of clusters differs among the analyzed dimensions and across genders. In the family trajectories, the algorithm detected four clusters for both gender; in the work trajectories, the algorithm detected two clusters for males and four clusters for females; in health trajectories, the algorithm identified three clusters for both genders; in the residence location trajectory, it detected five clusters for both gender; lastly, in the general life trajectories, the algorithm finds two clusters for males and three clusters for females. Graphical illustrations and information on the adopted clusters solution are in section Appendix E.

2.3.2 Sequences' features

The second sequence configuration we employ is that of sequence features. We extract four features with established social roles: sequencing, duration, timing, and entropy (see Billari et al., 2006 and Studer and Ritschard, 2016, and Bolano and Studer, 2020).

Sequencing refers to the order in which the states appear along the sequence. The

[7] This measure belongs to the class of "edit" distances, which equates distance to the minimal cost of transforming one sequence into another. What determines this cost are the number of operations required to transform one sequence into another and the cost of each operation. There are two primary operations: substitution and insertion-deletion (indel). The distinct characteristics of the DHD distance are its state-dependent and time-variant substitution costs (e.g., the cost of changing from the state of "in education" to the condition of "employment" differs whether we are at the beginning of or at the end of the working career)

social norms attached to this sequential aspect are well-documented. For example, the social consequences of having a child before marriage differ from when the first childbirth occurs after marriage.

To capture indicators of the sequencing, we employ "frequent sub-sequence mining." A sub-sequence is frequent if it occurs in more than 10% of sequences. Given a sequence *s*, e.g., A-B-C, a sub-sequence *z* is any subset of *s* that respects the ordering of *s*, e.g., A-B, B-C, A-C, A, B, C is all sub-sequences of *s*. Frequent sub-sequence are not mutually exclusive since the pattern A-B-C does not exclude the pattern A-B. We extract a list of frequent sub-sequences for each life course variable. We generate indicator variables for each extracted sub-sequence indicating the presence or absence of the sub-sequence in each trajectory.

The second extracted sequences aspect is the *spell duration*. The duration represents an individual's overall time in a specific sequence's state, for example, how long it has been married. This sequencing feature mirrors the concept of exposure to a given event. It has a crucial role in life course studies. For example, Mossakowski et al.(2009) estimate a negative effect of unemployment spell duration on mental health and well-being (see Mossakowski, 2009).

The concept of *timing* refers to the age at which a transition from one state to another occurred. The timing of events plays a relevant social role, given the presence of age-related social norms. For example, the critical period model emphasized the differential impact on the mental health of experiencing unemployment at the beginning or middle of a working career. The same applies to childbirth or marriage age. In our analysis, we included a timing indicator that refers to the time of each transition to different states over five years. For example, if an individual gets married at age 25, we created an indicator variable "20-25.married" that captures the transition to married and the time of its occurrence.

Finally, we included a measurement of within-sequence *entropy*. The within-sequence entropy measures the stability of the states along the trajectory. This measure does not account for states' order as well as it does not distinguish between positive and adverse

conditions.

The life course literature has largely overlooked the dimension of entropy when predicting future life outcomes. However, the definition of entropy is consistent with the concept of life changes. The entropy is equal to zero when the individual has experienced no life changes throughout the trajectory and one when the same amount of time has elapsed in each possible variable's states. Life changes were discussed in various areas of academia (see Haslam et al., 2021, Lin and Ensel, 1989 and Rahe, 1975). Changes in life, by increasing uncertainty in life, call into question the sense of autonomy and self-continuity, impairing wellness and mental health. Moreover, the negative effects of life changes arise especially when they are not accompanied by substantial social support (Lin and Ensel, 1989).

Following the literature on sequence analysis, we measure within sequence entropy (normalized) by the Shannon entropy formula:

$$h(p_1, \ldots, p_a) = \frac{-\sum_{i=1}^{a} p_i log_2(p_i)}{log_2 a}$$

where a is the size of the sequence alphabet and p_i is the proportion of occurrence's of the ith state in the considered sequence. The sequence features' configuration counts around 310 predictors, combining sub-sequencing, duration, timing, and entropy. Because of this increased number of variables, the predictor set will likely suffer from the multicollinearity problem. In addition, it is more likely that a non-linear rather than linear model describes the relations between inputs and output. Both are problematic for traditional techniques, such as logistic regression, but are handled by out-of-shelves SML algorithms.

3 Methods

Machine learning algorithms are data-driven methods that help discover patterns otherwise neglected by traditional knowledge. For an extensive illustration of these methods, refer to Hastie et al. (2009) (Hastie et al., 2009). In their more general formulation, Su-

pervised Machine learning (SML) algorithms are methods for automatically building a predictive function $\mathscr{F}$ that maps $X \in \mathscr{X}$, the predictor set, to a prediction $\hat{y} \in \mathscr{Y}$. The predictive function $\mathscr{F}$ is what we estimate from the data. Depending on the researcher's prior assumption, it could assume a more or less flexible form.

In general, any SML model requires optimizing two types of parameters: the structural parameters and the model hyper-parameters. Optimizing the structural parameters typically involves minimizing a loss function. For example, in a simple linear regression framework, the estimated coefficient β minimizes the residual sum of squares. The optimization of the model hyper-parameters is called the tuning stage. It is a crucial step of any SML's model-building procedure. In our predictive exercise, we optimize the model hyper-parameters through a stratified ten folds cross-validation with random search or grid search (or both) in the hyper-parameter space. All our models are trained to maximize the PR-AUC in the tuning phase. This metrics selection serves to overcome the accuracy paradox and train the models to maximize their depression detection. [8]

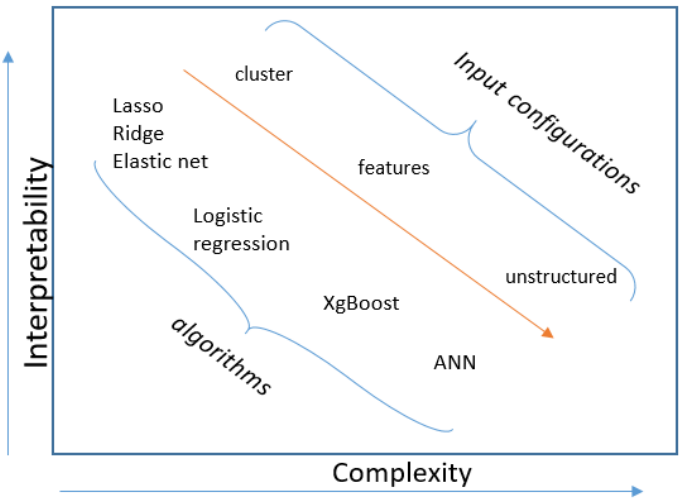
As pointed out by Athey and Imbens (2019) (Athey, 2019), estimating a wide range of machine learning models is always advisable, as they all have their advantages and disadvantages. In exploring SML methods, we started from the mainstream approaches based on logistic regression models fit by maximum likelihood. We proceeded along the trade-off between model complexity and interpretability. We applied shrinkage methods, Extreme Gradient Boosting (XGBoost), and Artificial Neural Networks (ANN). For an explanation of the model's specific characteristics and optimal models' hyper-parameters, refer to Appendix F and G, respectively.

Figure 3 illustrates the framework of input configurations and machine learning models explored in this article. As the complexity of the model changes, we combine a gradual increase in the amount and type of information provided at the input. Moving along the

[8] We divide the sample into a training (80% of the total sample) and a test sample (20%). Then, we divide the training data set into ten folds of equal size, preserving the percentage of samples for each target class in each fold. We repeat the same procedure ten times for each fold and hyper-parameter configuration: we keep out one fold (validation set) and train the model on the remaining nine folds. We evaluate the trained model in the held-out fold and retrieve the predictive scores. Once we used all ten folds, we chose the hyper-parameters combination that maximizes the predictive score across folds. The selected model is then used to estimate the training and test error

diagonal, from the top left to the bottom right, both learning models and input configurations increase their complexity.

FIGURE 3: Models-Inputs framework



3.1 Assessing predictive performance

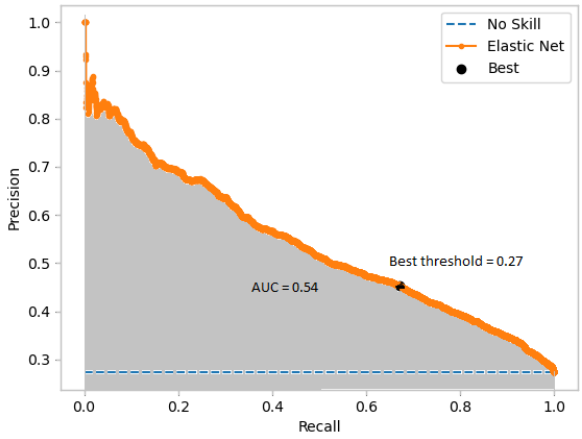
When predicting health diseases, the recall (sensitivity or true positive rate) and the Area Under the Precision-Recall curve (PR-AUC) are standard model selection and evaluation criteria (Steyerberg et al., 2010).

The recall represents the proportion of depressed people that the model correctly identified. It is an essential metric for disease diagnostic tools, as it measures the reliability of the diagnostic tool in detecting the disease. We chose to optimize this metric as predicting an individual is not at risk of depression when developing depression symptoms is more costly than the opposite mistake.

The precision represents the proportion of genuinely depressed among those predicted as depressed.

At last, the PR curve, the orange line in the figure 4, relates, at all possible threshold probability values, the recall and the precision.

FIGURE 4: Precision-Recall curve



The baseline of the PR curve is the horizontal line with the y-value equal to the depression prevalence in the sample. The AUC measures the area under the PR curve. When the AUC is near one, the classifier separates classes perfectly. An AUC near zero indicates the worst separability measure. Larger values of this metric indicate better model performance. Optimizing this metric allows for achieving the optimal trade-off between precision and recall.

4 Results

Following this paper's objectives, for each gender, we present each model's predictive performance according to the various data structures used. Then, we extract the most predictive life course factors for old age depression.

4.1 Predictive Performance

Figure 5 and 6 illustrate the Area-Under-the precision-recall curve across models, input structures, and genders. The box plots illustrate the distribution of the errors in the training samples. The red dots indicate the score in the test sample. Comparing these two errors is informative about overfitting, which occurs when the algorithms perform well on the training data but poorly out-of-sample.

We benchmarked our models' performance against two different baselines. The first

is a minimal baseline where we only use demographic information, e.g., age, interview year, interview season, country of residence, education, cohort, children, and migrant status. The second is the model from clinical studies that use health records and medical screenings.

Figure 5 and 6 illustrate the Area-Under-the precision-recall curve across models, input structures, and genders. The box plots illustrate the distribution of errors in the training sample. The red dots indicate the score in the test sample. Comparing these two errors is informative about overfitting. Overfitting happens when the algorithms perform well on the training data but poorly out-of-sample.

We benchmarked our models' performance against two different baselines. The first is a minimal baseline where we only use demographic information, e.g., age, interview year, interview season, country of residence, education, cohort, children, and migrant status. The second is the model from clinical studies that use health records and medical screenings.

The first input structure we explore is that of the sequences' clusters (purple box-plot). The predictor set counts around 113 predictors, remaining relatively small not to create multicollinearity issues. The best-performing models are the Gradient Boosting (XGBoost) and regularized regressions, which reach around 0.681 PR-AUC in the females' sample and 0.453 PR-AUC in the males' sample.

We then change the life sequences' configuration structure from clusters to sequence features (green box-plot). The predictor set now counts around 303 predictors. The PR-AUC in the training and test sets increases in all classifiers. The best model is the Gradient Boosting, which settles at a PR-AUC of 0.682 for females and 0.46 for males.

Finally, we try the unstructured sequence configuration (yellow box-plot). This predictor set counts around 327 predictors. The PR-AUC is higher than the cluster configuration and slightly smaller or equal to the sequences' features configuration. In this highly multicollinear setting, the noise in the input structure increases substantially. The models find less relevant patterns in the data to improve predictive performance.

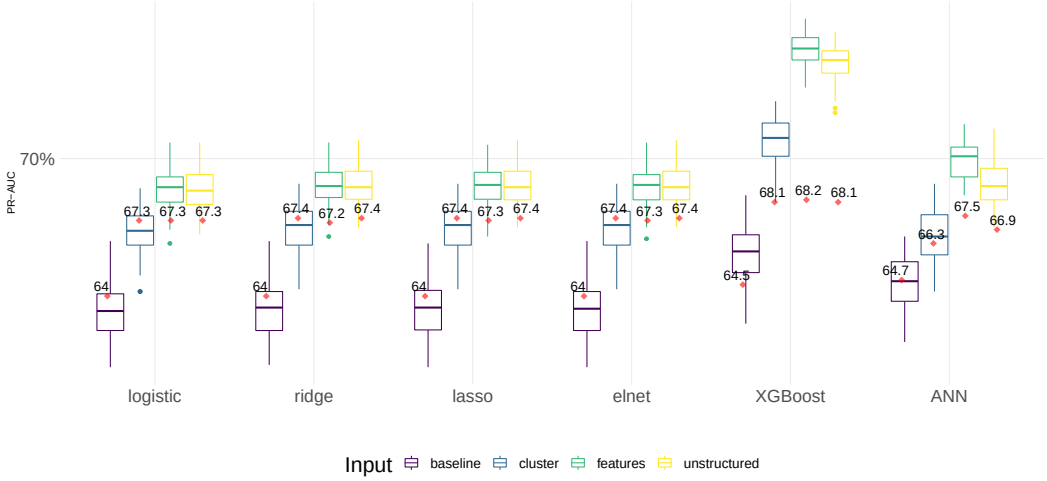
The similarity in predictive performance across different algorithms proves that com-

plex algorithms are comparable to the simpler traditional logistic regression model for the data at hand. We found similar conclusions in other clinical studies (for a systematic review, see Christodoulou et al., 2019).

Compared with the baseline predictor set, models trained with life course predictors achieve better predictive performance. The PR-AUC improves by 4 percentage points for all algorithms, implying that life course information does serve in increasing the ability for depression risk detection.

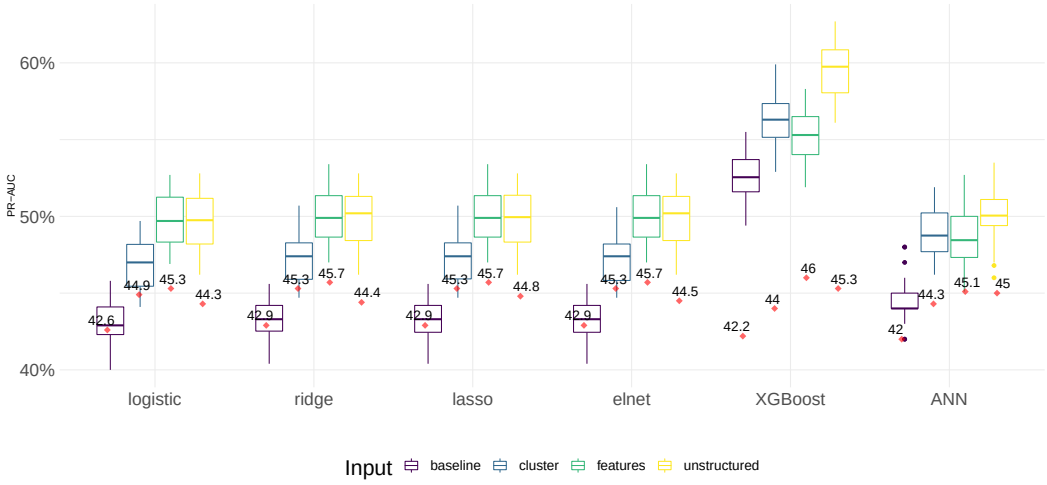
We highlight the significant difference in predictive performance across gender. Using the same type of life course information, the models achieve a PR-AUC of around 22 percentage points higher in the females' sample than the males' sample. This result highlights a need to differentiate depression diagnosis procedures across genders, as it would be highly insufficient to look only at socio-demographic factors to detect male depression.

FIGURE 5: Area-Under-Precision-Recall curve across models and input configurations, female sample



Note: Red dots indicates TEST error, box plots are the training errors. Colors represent different predictor set

FIGURE 6: Area-Under-Precision-Recall curve across models and input configurations, male sample



Note: Red dots indicates TEST error, box plots are the training errors. Colors represent different predictor set

The comparison of our life-course approach with current medical studies reveals that life histories have a slightly lower predicting ability than concurrent medical screening and health records (see Librenza-Garcia et al., 2021 and Garriga et al., 2022). These clinical studies reach a ROC-AUC of around 0.71-0.75. Instead, we reach a maximum ROC-AUC of around 0.695 for females and 0.672 for males (see Appendix H for other predictive performance metrics).

For this prediction result, we illustrate two potential explanations.

The first explanation targets the inner nature of the target variable we analyzed and the predictors we included. The expected out-of sample test error, for a given value x_0 and a given learning algorithm $f(x)$, can be always written as the sum of three fundamentals quantities:

$$E(y_0 - \hat{f}(x_0))^2 = \underbrace{\text{Var}(\hat{f}(x_0))}_{\text{variance}} + \underbrace{[\text{Bias}(\hat{f}(x_0))]^2}_{\text{bias}} + \underbrace{\text{Var}(\epsilon)}_{\text{irreducible error variance}} \tag{4.1}$$

The variance of the models' is given by the changes in the model's parameters estimates when changing the training set x_0. The bias refers to the error in fitting a real-life problem with an oversimplified function.

In any machine learning problem, the goal is to minimize the value on the left-hand

side of equation 4.1 by estimating a model $f(x)$ with low variance and bias. However, our prediction will still have some errors even after achieving this goal. This error is irreducible because no matter how well we estimate the model, we cannot reduce the out-of-sample prediction error below $\text{Var}(\epsilon)$.

This error value depends on the nature of the data and the amount of information available to the researcher. This paper targets a self-reported depression indicator that may be related to unobserved respondents' characteristics. For example, respondents can drastically change their depression perception if a dramatic accident occurs a few days before the interview. This situation will remain unobservable no matter how many predictors we include in the predictor set. Another element we do not control for but that correlates with depression is genetic endowment (Levinson, 2006).

Our model-building procedure, based on repeated cross-validations, ensures that we have taken all the necessary steps to reduce models' variance and bias. Hence, our prediction error likely comes from the irreducible component.

The second potential explanation addresses the dimensionality of the sample used to train the algorithms. Indeed, black-box SML models as neural networks need big data of millions of observations to exploit their predictive ability fully. The empirical sample in this analysis counts around 60 000 observations. This relatively small sample size may harm the model's detection ability. To test this explanation, we train our models using increasing fractions of the training data, from 10% to 90%. For each training fraction, we compute the test PR-AUC. We notice that with 40% of the training data, the test AUC reaches almost the same score as the whole training set (see Appendix I), suggesting training-size independence.

As an extension of the main results presented here, in Appendix J, we report ML models' performances for different EURO-D depression thresholds, i.e., 3, 5, and 6, in a pooled sample. For all algorithms, increasing the EURO-D depression thresholds deteriorates the classification performance. The fewer depressed examples in the sample, the less information is made available for the algorithms to learn significant depression patterns. This reduction affects the models' detection ability of depressed cases.

4.2 SHAP values across genders

This section sheds light on the complexity behind the ML algorithms' predictions. We sought to understand how variables contributed to generating the final individual predicted probabilities for each gender. We employ the Shapley Additive exPlanations (SHAP) method for this aim. This new method has provided reliable and consistent results in previous research (see Lundberg and Lee, 2017). SHAP relies on the Shapley values concept, which originates from the collaborative game theory (Shapley, 1953).

Contrary to other variable importance metrics, the SHAP framework is the only explanation method that can, in principle, explain any predictive model, i.e., it is a model-agnostic tool (see Lundberg et al., 2020 and Molnar, 2020). However, like other interpretability tools, SHAP does not detect the presence of collinear variables. When two variables are collinear, the SHAP score of one of the two is equal to zero.

The general idea underlying the SHAP framework is to estimate, for any given model, a simpler explanation model, which corresponds to an interpretable approximation of the initial model. Given a vector $\boldsymbol{x}$ of p predictor variables, $\boldsymbol{x} = [x_1, \ldots, x_p]$, and a trained model f, SHAP approximate the model f with a simple explanation model g that has the following form:

$$g(\boldsymbol{z}) = \phi_0 + \sum_{i=1}^{p} \phi_i z_i. \tag{4.2}$$

In equation 4.2, $\boldsymbol{z} = [z_1, \ldots, z_p]$ is a coalition vector, where z_i is equal to 1 if the variable x_i is present and 0 if the variable is absent, p is the number of predictors, and $\phi_i \in R$ is the variable i contribution to the model predictions, i.e., the Shapley value. The Shapley value ϕ_i is then estimated through the following equation:

$$\phi_i(f, x) = \sum_{\boldsymbol{z} \subset \boldsymbol{x}} \frac{|\,\boldsymbol{z}\,|!(p - |\,\boldsymbol{z}\,| - 1)!}{p!} [f(\boldsymbol{z}) - f(\boldsymbol{z} \backslash i)] \tag{4.3}$$

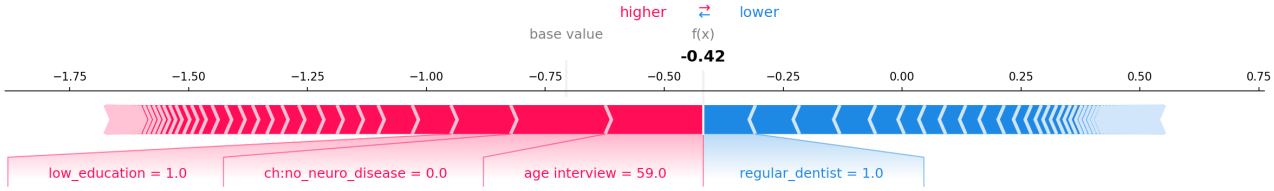
where $|\,\boldsymbol{z}\,|$ is the number of non-zero entries in $\boldsymbol{z}$.

The exact estimation of each ϕ_i may be computationally infeasible. However, Lundeberg et al. (Lundberg and Lee, 2017) introduced efficient algorithms to estimate such values in the case of Gradient Boosting (see Lundberg et al., 2018) and Neural Networks.

Our model-input exploration has yet to reveal a clear winner across the model. However, it shows a higher predictive power of life sequences' features, i.e., timing, duration, sequencing, and entropy. In what follows, we illustrate SHAP values for this sequence configuration only.

To understand the meaning of SHAP values, Figure 7 shows which features contributed to the model's prediction for a single randomly selected observation (a not depressed Slovenian female 59 years old).

FIGURE 7: A SHAP force plot of a single individual



Note: In **bold** is the predicted odd ratio, which correspond to 0.39 probability of being depressed. Red represents features that pushed the model probability score higher, blue represents features that pushed the score lower

The bold number -0.42 represents the predicted odds of being depressed, which in probability terms translates to 0.39. We color the features essential to predicting this observation in red and blue. Red represents features that pushed the model probability score higher, and blue represents features that moved the score lower. Features that had more of an impact on the score locate closer to the dividing boundary between red and blue. The bar represents the impact size. For this random individual, what contributes more to the increase in the depression score is having a low education level, having had a neurological disease in childhood, and her age at the time of the interview. What pushes down the risk of depression is to be gone to a dentist regularly.

Aggregating the results for all test predictions, Figure 8 illustrates the SHAP summary plot for the top twenty predictors for the Gradient Boosting for males and females.

The summary plot combines variable contribution with variable effects. Each point on the summary plot is a Shapley value for a feature and an instance. The feature's position on the y-axis is determined by the absolute average Shapley value and on the x-axis

by the Shapley value. The color represents the variable's value from low (blue color) to high (pink color). The number on the right of each variable name corresponds to the average SHAP value across all observations.

Comparing SHAP values across genders highlights idiosyncratic and common factors and traditional and original features. In line with the literature (see Clark and Lee, 2021 and Layard et al., 2014), for both genders, we found that material deprivation in childhood ("childhood: no basic facilities," and "childhood": rooms per capita"), low education, low subjective childhood health ("ch. health"), and low utilization of dental care services ("regular_dentist") predict higher depression in later life. These childhood-specific variables appear in all input configurations and all predictive algorithms. This result indicates that no matter the amount of adult life course information we provide to train the algorithms, childhood conditions matter most.

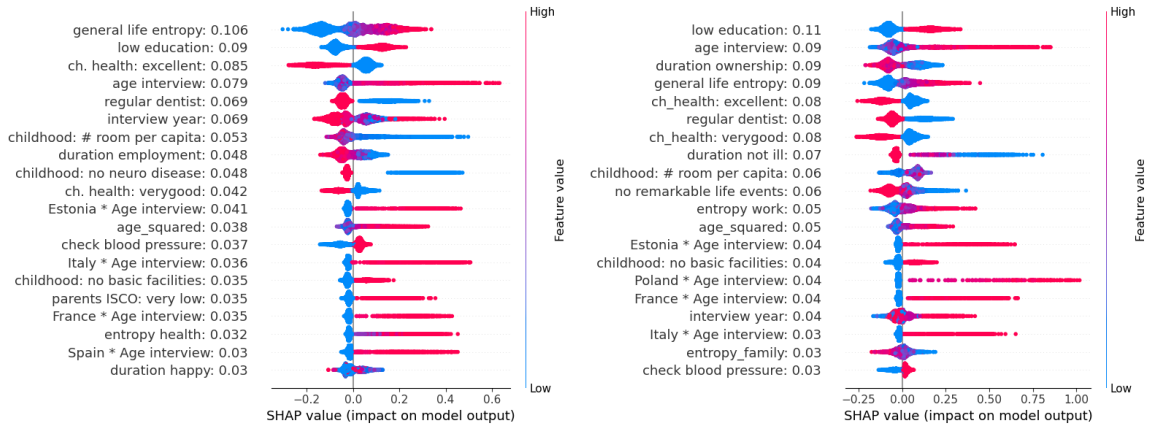
Interestingly, we extract a new predictive variable from the black box methods. For both genders, the entropy within the general life sequence ("general life entropy") increased the likelihood of depression later in life. The general life entropy refers to the number of remarkable life events creating emotional responses adults have undergone across their life courses. This finding opens room for further investigation of the role of emotional shocks in life course analysis. It also highlights the importance of monitoring emotional stressors throughout the lifetime.

Moreover, we extracted two distinctive male predictors: the entropy in the work and family sequences. It results that high-frequency changes in work status and low-frequency changes in family status predict higher depression risk for males but not females. As idiosyncratic female predictors, we find that low parents' ISCO and low employment duration increase the likelihood of depression.

Another finding concerns the heterogeneity in the predicting power of age across countries. The SHAP values for the age-countries interaction variables are high in all models and for each gender. However, age's contribution to the likelihood of depression differs across countries. In countries such as Italy, Poland, Hungary, Portugal, and Spain, being older contributes to an increased risk of depression. In countries such as Sweden,

Denmark, and Switzerland, the effect is the opposite, with higher ages associated with a lower risk of depression. This result also prompts further investigation of predictors that may vary by country.

FIGURE 8: Shapley values for Gradient Boosting, female (right) and male (left)



Note: Each point on the summary plot is a Shapley value for a feature and an instance. The position on the y-axis is determined by the feature rank and on the x-axis by the Shapley value. Colors indicate relationship between variables and depression probability: blue to red signifies positive correlation with depression, while red to blue signifies negative correlation. The number on the right of each variable name corresponds to the average SHAP value across all observations

5 Discussion

At the heart of this analysis was a comparison of supervised machine learning techniques applied to various life course data configurations to predict the risk of depression in people over fifty years old. Three key findings emerged from our analysis.

First, socioeconomic life trajectories help to anticipate outbreaks of depression later in life, but not perfectly. Secondly, large-scale life course information is more suitable for prediction tasks. We have yet to identify a clear winner across ML models. However, all models achieved the highest predictive performance when combined with sequential features, considering the duration, timing of state transitions, state sequencing, and entropy within the life course. With this data configuration, the models predict depression risk with a PR-AUC of 0.68 for females and 0.46 for males in the training and test sample. Compared to the minimum set of minimal predictors, where we only use demographic information, life course information improves predictive performance by about 4 per-

centage points for both genders. However, life course information yields lower predictive performance than other clinical studies using medical records (see Garriga et al., 2022). Overall, this finding indicates that for optimum depression risk detection, we need more sensitive data.

The third relevant aspect of this analysis stems from the gendered stratification and the extraction of depression predictors. We merged respondents from nineteen European countries and trained our models independently for each gender. We did so because of the substantial gender gap we observed in the depression item. Females suffer more from depression in all the analyzed countries. A concise explanation of this gender gap is still missing. Some new evidence suggests that it may be related to the genetic endowment of females and males. However, stratifying by gender was the straightforward way to shed light on potential differences in depression patterns.

SHAP feature extraction combined with gender stratification revealed established patterns of depression and new variables that traditional linear models neglect. For both genders and in all models, material deprivation in childhood, poor health in childhood and adulthood, and low education predicts a higher probability of depression later in life. The duration of ownership or lease predicts a lower probability of depression. As new predictive characteristics, we identify general life and work sequences' entropy. Entropy is the frequency of changes in condition over the life sequence. Men and women who go through successive periods of happiness, stress, financial stress or hunger are more likely to experience depression at a later age. Greater entropy in the work sequence and lower entropy in the family sequence increase the probability of depression for men only.

Similarly to previous well-being studies (see Clark and Lee, 2021, Zheng et al., 2021), our SHAP values results stressed the long-lasting influence of early childhood conditions on later-life well-being outcomes. Therefore, an essential point of intervention for policies addressing depression is early childhood disadvantages.

Second, our results suggests that many life changes might increase mental health problems. Thus highlighted is the unique importance and meaning of change in itself. Moreover, these results highlight the primary importance of gender, as they show that

predictors of depression tend to vary between men and women. These differences have implications for the design of mental health interventions.

The country-age interactions show significant heterogeneity across countries. In countries like Hungary, Poland, and Italy, increasing age predicts a higher probability of depression. On the other hand, in Denmark, Switzerland, and Sweden, increasing age predicts a lower likelihood of depression. This heterogeneous effect remains unexplained, but it may be related to the differences in European social welfare and pension systems.

Most of the existing literature focuses on treatments for depression; this article proposes to look at the preconditions of depression and thus shift the discussion from treatment to preventive and personalized healthcare. Our analysis, although not wholly, shows that past life trajectories may foreshadow later life depression outbreaks. These results motivate social insurance systems to increase their role in preventing diseases from alleviating pressures on the healthcare systems ex-post.

We show that a high general life entropy increases the likelihood of clinical depression in old age. Today, we live with high social and economic uncertainties, e.g., the Covid pandemic, the Ukraine war, and the economic downturn. These macro phenomena affect individuals by creating financial and work instabilities, displacements, and insecurities, thus increasing entropy in all life dimensions. Our findings, therefore, highlight the importance of welfare systems interventions to reduce or smooth the impact of such abrupt events. This topic is of both theoretical and empirical interest. More broadly, this study prompts further investigations seeking explanations and rationales for the findings uncovered here.

6 Acknowledgements

We gratefully acknowledge financial support from the Luxembourg National Research Fund under the PRIDE programme (PRIDE17/12252781-DTU-DRIVEN) and from the Luxembourg Ministry of Higher Education and Research.

The SHARE data collection has been funded by the European Commission, DG RTD through FP5 (QLK6-CT-2001-00360), FP6 (SHARE-I3: RII-CT-2006-062193, COMPARE:

CIT5-CT-2005-028857, SHARELIFE: CIT4-CT-2006-028812), FP7 (SHARE-PREP: GA N°211909, SHARE-LEAP: GA N°227822, SHARE M4: GA N°261982, DASISH: GA N°283646) and Horizon 2020 (SHARE-DEV3: GA N°676536, SHARE-COHESION: GA N°870628, SERISS: GA N°654221, SSHOC: GA N°823782, SHARE-COVID19: GA N°101015924) and by DG Employment, Social Affairs and Inclusion through VS 2015/0195, VS 2016/0135, VS 2018/0285, VS 2019/0332, and VS 2020/0313. Additional funding from the German Ministry of Education and Research, the Max Planck Society for the Advancement of Science, the U.S. National Institute on Aging (U01_AG09740-13S2, P01_AG005842, P01_AG08291, P30_AG12815, R21_AG025169, Y1-AG-4553-01, IAG_BSR06-11, OGHA_04-064, HHSN271201300071C, RAG052527A) and from various national funding sources is gratefully acknowledged.

References

Gateway to Global Aging Data, Produced by the Program on Global Aging, Health,and Policy, University of Southern California with funding from the National Institute on Aging (R01 AG030153). https://g2aging.org/. Accessed: 2021-11-11.

A. Abbott. Sequence analysis: New methods for old ideas. *Annual review of sociology*, 21 (1):93–113, 1995.

S. Aisenbrey and A. E. Fasang. New life for old ideas: The" second wave" of sequence analysis bringing the" course" back into the life course. *Sociological methods & research*, 38 (3):420–462, 2010.

B. Arpino, J. Gumà, and A. Julià. Early-life conditions and health at older ages: The mediating role of educational attainment, family and employment trajectories. *PloS one*, 13 (4):e0195320, 2018.

S. Athey. The impact of machine learning on economics. In *The economics of artificial intelligence*, pages 507–552. University of Chicago Press, 2019.

R. Atkins, A. J. Turner, T. Chandola, and M. Sutton. Going beyond the mean in examining relationships of adolescent non-cognitive skills with health-related quality of life and biomarkers in later-life. *Economics & Human Biology*, 39:100923, 2020.

N. Z. Bakkeli. Predicting psychological distress during the covid-19 pandemic: Do socioeconomic factors matter? *Social Science Computer Review*, page 08944393211069622, 2022.

R. Berk. *Criminal justice forecasts of risk: A machine learning approach*. Springer Science & Business Media, 2012.

F. C. Billari, J. Fürnkranz, and A. Prskawetz. Timing, sequencing, and quantum of life course events: A machine learning approach. *European Journal of Population/Revue Européenne de Démographie*, 22(1):37–65, 2006.

D. Blazer, L. K. George, R. Landerman, M. Pennybacker, M. L. Melville, M. Woodbury, K. G. Manton, and K. Jordan. Psychiatric disorders: a rural/urban comparison. *Archives of general psychiatry*, 42(7):651–656, 1985.

A. Bogomolov, B. Lepri, M. Ferron, F. Pianesi, and A. Pentland. Daily stress recognition from mobile phone data, weather conditions and individual traits. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 477–486, 2014.

D. Bolano and M. Studer. The link between previous life trajectories and a later life outcome: a feature selection approach. Working paper 82, Swiss National Competence Center in Research, 2020.

M. H. Bornstein. Sensitive periods in development: structural characteristics and causal interpretations. *Psychological bulletin*, 105(2):179, 1989.

A. Börsch-Supan. Survey of Health, Ageing and Retirement in Europe (SHARE), wave 7. *Release version 7.1.0*, 7(0), 2019.

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

P. Brunori and G. Neidhöfer. The evolution of inequality of opportunity in germany: A machine learning approach. *Review of Income and Wealth*, 67(4):900–927, 2021.

M. Charrad, N. Ghazzali, V. Boiteau, and A. Niknafs. Nbclust: an R package for determining the relevant number of clusters in a data set. *Journal of statistical software*, 61:1–36, 2014.

A. A. Choudhury, M. R. H. Khan, N. Z. Nahim, S. R. Tulon, S. Islam, and A. Chakrabarty. Predicting depression in bangladeshi undergraduates using machine learning. In *2019 IEEE Region 10 Symposium (TENSYMP)*, pages 789–794, 2019. doi: 10.1109/TENSYMP46218.2019.8971369.

E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. Van Calster. A systematic review shows no performance benefit of machine learning over logistic

regression for clinical prediction models. *Journal of clinical epidemiology*, 110:12–22, 2019.

A. E. Clark and T. Lee. Early-life correlates of later-life well-being: Evidence from the Wisconsin longitudinal study. *Journal of Economic Behavior & Organization*, 181:360–368, 2021.

I. Colman and A. Ataullahjan. Life course perspectives on the epidemiology of depression. *The Canadian Journal of Psychiatry*, 55(10):622–632, 2010.

D. J. Conti and W. N. Burton. The economic impact of depression in a workplace. *Journal of Occupational Medicine*, 36(9):983–988, 1994.

J. Currie and D. Almond. Human capital development before age five. In D. Card and O. Ashenfelter, editors, *Handbook of Labor Economics*, volume 4, pages 1315–1486. Elsevier, 2011.

E. Diener, R. Inglehart, and L. Tay. Theory and validity of life satisfaction scales. *Social Indicators Research*, 112(3):497–527, 2013.

R. N. Engstrom, J. Hersh, and D. Newhouse. Poverty in HD : What does high resolution satellite imagery reveal about economic welfare ? 2016.

J. Falkingham, M. Evandrou, M. Qin, and A. Vlachantoni. Accumulated lifecourse adversities and depressive symptoms in later life among older men and women in England: a longitudinal study. *Ageing and Society*, 40(10):2079–2105, 2020. doi: 10.1017/S0144686X19000461.

S. Flèche, W. N. Lekfuangfu, and A. E. Clark. The long-lasting effects of family and childhood on adult wellbeing: Evidence from British cohort data. *Journal of Economic Behavior & Organization*, 181:290–311, 2021. ISSN 0167-2681. doi: https://doi.org/10.1016/j.jebo.2018.09.018. URL https://www.sciencedirect.com/science/article/pii/S016726811830266X.

J. Friedman, T. Hastie, R. Tibshirani, et al. *The elements of statistical learning*. Springer series in statistics New York, 2001.

J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

A. Gabadinho, G. Ritschard, N. S. Mueller, and M. Studer. Analyzing and visualizing state sequences in R with TraMineR. *Journal of statistical software*, 40(4):1–37, 2011.

R. Garriga, J. Mas, S. Abraha, J. Nolan, O. Harrison, G. Tadros, and A. Matic. Machine learning model to predict mental health crises from electronic health records. *Nature Medicine*, 28:1240–1248, 2022.

C. Haslam, S. A. Haslam, J. Jetten, T. Cruwys, and N. K. Steffens. Life change, social identity, and health. *Annual Review of Psychology*, 72:635–661, 2021.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

E. Havari and F. Mazzonna. Can we trust older people's statements on their childhood circumstances? Evidence from SHARELIFE. *European Journal of Population*, 31(3): 233–257, 2015.

T. L. G. Health. Mental health matters. *The Lancet. Global Health*, 8(11):e1352, 2020.

N. Jaques, S. Taylor, A. Sano, and R. Picard. Multi-task, multi-kernel learning for estimating individual wellbeing. In *Proc. NIPS Workshop on Multimodal Machine Learning, Montreal, Quebec*, 2015.

L. Kaufman and P. J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.

S. Kisely. No mental health without oral health. *The Canadian Journal of Psychiatry*, 61 (5):277–282, 2016.

R. Layard, A. E. Clark, F. Cornaglia, N. Powdthavee, and J. Vernoit. What predicts a successful life? A life-course model of well-being. *The Economic Journal*, 124(580):720–738, 2014.

A. K. Leist, M. Klee, J. H. Kim, D. H. Rehkopf, S. P. Bordas, G. Muniz-Terrera, and S. Wade. Mapping of machine learning approaches for description, prediction, and causal inference in the social and health sciences. *Science Advances*, 8(42), 2022.

L. Lesnard. Setting cost in optimal matching to uncover contemporaneous socio-temporal patterns. *Sociological methods & research*, 38(3):389–419, 2010.

D. F. Levinson. The genetics of depression: a review. *Biological psychiatry*, 60(2):84–92, 2006.

T. F. Liao, D. Bolano, C. Brzinsky-Fay, B. Cornwell, A. E. Fasang, S. Helske, R. Piccarreta, M. Raab, G. Ritschard, E. Struffolino, and M. Studer. Sequence analysis: Its past, present, and future. *Social Science Research*, 107:102772, 2022. ISSN 0049-089X. doi: https://doi.org/10.1016/j.ssresearch.2022.102772. URL https://www.sciencedirect.com/science/article/pii/S0049089X22000783.

D. Librenza-Garcia, I. C. Passos, J. G. Feiten, P. A. Lotufo, A. C. Goulart, I. de Souza Santos, M. C. Viana, I. M. Benseñor, and A. R. Brunoni. Prediction of depression cases, incidence, and chronicity in a large occupational cohort using machine learning techniques: an analysis of the ELSA-Brasil study. *Psychological Medicine*, 51(16):2895–2903, 2021. doi: 10.1017/S0033291720001579.

N. Lin and W. M. Ensel. Life stress and health: stressors and resources. *American Sociological Review*, 54(3):382–399, 1989.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30:4768–4777, 2017.

S. M. Lundberg, G. G. Erion, and S.-I. Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018.

S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67, 2020.

L. McBride and A. Nichols. Retooling poverty targeting using out-of-sample validation and machine learning. *The World Bank Economic Review*, 32(3):531–550, 2018.

C. Molnar. *Interpretable machine learning*. Lulu. com, 2020.

K. N. Mossakowski. The influence of past unemployment duration on symptoms of depression among young women and men in the united states. *American Journal of Public Health*, 99(10):1826–1832, 2009.

M. D. Nemesure, M. V. Heinz, R. Huang, and N. C. Jacobson. Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence. *Scientific reports*, 11(1):1–9, 2021.

E. Pakpahan, R. Hoffmann, and H. Kröger. The long arm of childhood circumstances on health in old age: Evidence from SHARELIFE. *Advances in Life Course Research*, 31: 1–10, 2017.

M. J. Prince, F. Reischies, A. T. Beekman, R. Fuhrer, C. Jonker, S.-L. Kivela, B. A. Lawlor, A. Lobo, H. Magnusson, M. Fichter, et al. Development of the EURO–D scale–a European Union initiative to compare symptoms of depression in 14 European centres. *The British Journal of Psychiatry*, 174(4):330–338, 1999.

R. H. Rahe. Epidemiological studies of life change and illness. *The International Journal of Psychiatry in Medicine*, 6(1-2):133–146, 1975.

M. Sajjadian, R. W. Lam, R. Milev, S. Rotzinger, B. N. Frey, C. N. Soares, S. V. Parikh, J. A. Foster, G. Turecki, D. J. Müller, et al. Machine learning in the prediction of depression treatment outcomes: a systematic review and meta-analysis. *Psychological Medicine*, 51(16):2742–2751, 2021. doi: 10.1017/S0033291721003871.

D. Sansone. Beyond early warning indicators: high school dropout and machine learning. *Oxford bulletin of economics and statistics*, 81(2):456–485, 2019.

L. S. Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2:307–317, 1953.

P. Sobocki, B. Jönsson, J. Angst, and C. Rehnberg. Cost of depression in europe. *Journal of Mental Health Policy and Economics*, 2006.

D. J. Stekhoven and P. Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

E. W. Steyerberg, A. J. Vickers, N. R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. J. Pencina, and M. W. Kattan. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*, 21(1):128–138, 2010.

M. Studer and G. Ritschard. What matters in differences between life trajectories: A comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 179(2):481–511, 2016.

United Nations, Department of Economic and Social Affairs. World population ageing 2019. Technical report, United Nations, 2019. URL https://www.un.org/en/development/desa/population/publications/pdf/ageing/WorldPopulationAgeing2019-Highlights.pdf.

S. Van de Velde, P. Bracke, and K. Levecque. Gender differences in depression in 23 european countries. Cross-national variation in the gender gap in depression. *Social science & medicine*, 71(2):305–313, 2010.

B. Vanhoutte, M. Wahrendorf, and J. Nazroo. Duration, timing and order: How housing histories relate to later life wellbeing. *Longitudinal and Life Course Studies*, 8(3):227–243, 2017. doi: 10.14301/llcs.v8i3.445.

M. Wahrendorf, D. Blane, M. Bartley, N. Dragano, and J. Siegrist. Working conditions in

mid-life and mental health in older ages. *Advances in Life Course research*, 18(1):16–25, 2013.

S. S. Walker and U. Schimmack. Validity of a happiness implicit association test as a measure of subjective well-being. *Journal of Research in Personality*, 42(2):490–497, 2008.

E. WHO. Mental health: Facing the challenges, building solutions. In *Report from the WHO European Ministerial Conference*. WHO, 2005.

T. D. Wilson and D. T. Gilbert. Affective forecasting: Knowing what to want. *Current Directions in Psychological Science*, 14(3):131–134, 2005.

X. Zheng, S. Shangguan, Z. Fang, and X. Fang. Early-life exposure to parental mental distress and adulthood depression among middle-aged and elderly chinese. *Economics & Human Biology*, 41:100994, 2021.

A Depression

TABLE 1: Depression prevalence within countries and across genders

	Female		Male	
Country	N	%	N	%
Austria	1939	24	1351	11
Germany	2393	27	2177	15
Sweden	2393	22	1788	10
the Netherlands	1165	13	976	7
Spain	2706	33	2187	12
Italy	2703	32	2276	16
France	2415	31	1795	14
Denmark	2083	14	1819	6
Greece	2183	24	1682	11
Switzerland	1530	12	1279	6
Belgium	3196	26	2644	13
Czech Republic	2832	20	1888	9
Poland	1201	46	945	27
Luxembourg	603	31	503	14
Hungary	868	41	537	22
Portugal	607	46	449	18
Slovenia	1927	22	1312	10
Estonia	2858	32	1618	17
Croatia	1065	36	833	19
Total	35957	25	28016	12

Note: The depression prevalence (%) indicates respondents' percentages with at least one depression measurement in the longitudinal measurements. We define Depression as a binary indicator that takes the value of one when the respondent has more than three EURO-D symptoms. EURO-D symptoms are: sadness, pessimism, suicidality, guilt, sleep, interest, irritability, appetite, fatigue, concentration, enjoyment, and tearfulness

B Descriptive statistics

TABLE 2: Static predictors data set. Descriptive statistic by gender. Imputed sample

	Female			Male		
Variables and categories	N.imputed	Mean or Frequency	SD or %	N.imputed	Mean or Frequency	SD or %
Age: 50-89	0	64.93	9.22	0	65.37	8.85
Cohort:	0					
< 1930		1924	0.05		1411	0.05
1930-1939		6896	0.19		5607	0.20
1940-1949		11677	0.32		9719	0.35
1950-1959		12219	0.34		9265	0.33
1960-1969		3241	0.09		2014	0.07
Migration:	11			0		
no migration		32776	0.91		25684	0.92
before 18		1139	0.03		865	0.03
after 18		2042	0.06		1467	0.05
Educational level (ISCED):	0			0		
High		7128	0.20		6755	0.24
Medium		12898	0.36		11000	0.39
Low		14022	0.39		9137	0.33
No education		1909	0.05		1124	0.04
Childhood conditions						
Occupation parents (ISCO) :	6313			5190		
Very High		4442	0.12		3497	0.12
High		4292	0.12		3351	0.12
Low		22869	0.64		17986	0.64
Very Low		4354	0.12		3182	0.11
Rooms per capita	3347	0.73	0.58	2871	0.76	0.66
Number of books:	2684			2374		
1 (0-10 books)		14030	0.39		11307	0.40
2 (11-25 books)		8337	0.23		6461	0.23
3 (26-100 books)		8248	0.23		6407	0.23
4 (101-200 books)		2702	0.08		1846	0.07
5 (more than 200 books)		2640	0.07		1995	0.07
Basic facilities	0	25814	0.72	0	20458	0.73
Ever in hospital in	63	2252	0.06		1843	0.07
Ever missed school	344	4014	0.11	192	2925	0.10
No infectious disease	0	5530	0.15		5563	0.20
No neoplastic disease	0	33418	0.93	0	26124	0.93
No neuro disease		33347	0.93		26720	0.95
Childhood self health rated:	32			23		
1 ("Fair, poor, spontaneous")		4128	0.11		2601	0.09
2 ("Good")		9484	0.26		6786	0.24
3 ("Very good")		11431	0.32		8847	0.32
4 ("Excellent")		10914	0.30		9782	0.35
Live with biological father	2610	31996	0.89	2357	25217	0.90
Live with biological brother	2610	30311	0.84	2357	23518	0.84
Adulthood controls						
Ever blood pressure		24205	0.67		19128	0.68
Regular dentist	41	27596	0.77	25	19547	0.70
Children:	162			272		
after 30 years old		4165	0.12		7058	0.25
between 25-30 years old		9699	0.27		10755	0.38
before 25 years old		18370	0.51		6485	0.23
no children		3723	0.10		3718	0.13
SES between 20 and 30	0					
very high		1052	0.03		755	0.03
high		2276	0.06		2358	0.08
low		15806	0.44		11055	0.39
very low		4858	0.14		4759	0.17
not employed		11965	0.33		9089	0.32
Job change	387			158		
never in paid employment		3797	0.11		297	0.01
No change		10208	0.28		7248	0.26
One change		8200	0.23		7182	0.26
Two or more change		13752	0.38		13289	0.47
Age own home	11			1		
after 30		2264	0.06		3332	0.12
before 25		27718	0.77		15667	0.56
between 25 and 30		5975	0.17		9017	0.32
Cohabitation	0			0		
Cohabitation only before/at 25		2878	0.08		912	0.03
Cohabitation between 25 and 50		24246	0.67		14036	0.50
Cohabitation after 50		6344	0.18		10761	0.38
Never cohabitation		2489	0.07		2307	0.08

C Sequence states

We construct life sequences for six variables: work status, housing arrangement, family, health, residence location, and general life events. We draw three sequences (work status, health, and housing) from the Gateway to Global Aging portal (gwd). All the variables come from the SHARELIFE questionnaire.

To construct the work history, SHARELIFE asked respondents to report when they finished full-time education and question specific job spells. We used details on the start and end of respective job spells, and we determined if the gap was because of being unemployed (both searching and not searching for a job), home or family work, retirement, or a remaining group of others. The other category includes being sick or disabled, voluntary work, military services, and traveling.

To construct health history variables, SHARELIFE asked respondents how many periods of poor health or disability (lasting more than a year) they had in their life from age 16 onwards. If the number of periods of poor health or disability was more than three, people were automatically classified as "Ill most of their life" throughout their history. In contrast, if the respondent answered three or fewer periods, respondents were additionally asked to report when the respective periods were. SHARELIFE respondents reported the precise years when each period started and ended.

The housing arrangement histories combine details regarding the respondent's housing spells, including the reported year they left their parent's home and reported the year they established their household, if applicable. In SHARELIFE, respondents could report up to 28 housing spells that lasted six months or longer. We classify as non-private those types of residences that are hard to classify, such as rent-free or non-private residences (e.g., boarding schools, hospitals, or prisons). To classify residences as "abroad," we used information on whether the residence was in the country. We did not distinguish between types of residences abroad because the number of people who lived abroad was too small.

The family histories combine the children's, cohabitation, and partner's histories. The

children's histories contain information on the age at which the respondent had or adopted their child and the number of children at each respondent' age. We include information on death in the case that a child dies. The cohabitation history contains information on cohabitation spells. SHARELIFE asked respondents about when they started living with a partner (beginning of spell) and, if they stopped living with the same partner, the age at which they stopped living with them (end of a spell). The end of cohabitation could be because a partner died, a relationship broke up, a partner moved into nursing or cared home, or other reasons. The partner history distinguishes between married or non-married partnership and the alone status.

The residence location histories inform about the location where respondent report they had their accommodations. SHARELIFE asked respondents about housing spells and whether it was in a big city, rural area, large town, or small village for each period.

Finally, the general life history combines the period of stress, financial stress, happiness, and hunger. SHARELIFE asked respondents to reflect on their past life and report whether there was a distinct period during which they were happier, under more stress, with financial hardship, or suffering from hunger. In an affirmative answer, the respondents must report the starting year and the stopping year or whether the period was still ongoing. In a negative response, we classified the history as "no events."

TABLE 3: Sequences alphabet

Work status	Health	Location	Family	Event	Housing
Employed (E)	Not ill	Big city (BC)	Alone (A)	Financial stress (FS)	Owner (O)
Self-Employed (SE)	Ill	Large town (LT)	Alone with children (AC)	Happy (H)	Tenant (T)
Unemployed (U)	Ill mostly	Rural area (RA)	Married (M)	Hunger (Hu)	Non-private (NP)
Family work (FW)		Small town (ST)	Married with children (MC)	No events (NE)	Abroad (Ab)
Retired (R)		Suburbs (Sub)	With partner (P)	Stress (S)	Parental home (Par)
In education (FE)		Missing (NA)	With partner and children (PC)	Happy and Stress (H+S)	
Other job (Oj)				Happy and fin. stress (H+FS)	
Missing (NA)				Other events (Ot)	
				Stress and fin.stress (S+FS)	

D Data pre-processing

We perform five main pre-processing steps on the three: imputation of missing data, one-hot encoding of categorical variables, removal of zero-variance and near-zero variance (sd < 0.015) items, drop perfectly collinear variable and highly collinear variable (correlation > 0.8) and normalizing predictors through the min-max normalization.

Although we could have performed more data pre-processing operations, e.g., eliminating variance inflation factor and variables with very low (but not zero) variance, we decided to perform only these five basic operations for this first empirical exploration. Indeed, automatic, uncontrolled elimination of variables could deprive the model of help-

ful information for learning that would affect the resulting predictions.

E Clusters

TABLE 4: Prevalence of cluster solutions and measures of homogeneity for six the variables analyzed. Females sample.

	Work			**House**			**Family**		
Cluster	N	%	Homogen.	N	%	Homogen.	N	%	Homogen.
1	26760	67	0.503	7923	20	0.67	2032	5	0.64
2	8017	20	0.634	27880	70	0.45	32178	80	0.62
3	2663	7	0.443	3986	10	0.47	2262	6	0.64
4	2349	6	0.185				3317	8	0.17
Pseudo R^2	0.52			0.57			0.58		
	General Life			**Health**			**Location**		
Cluster	N	%	Homogen.	N	%	Homogen.	N	%	Homogen.
1	27026	68	0.62	37669	95	0.98	7196	18	0.7
2	5978	15	-0.17	1187	3	0.40	13339	34	0.7
3	6785	17	0.61	933	2	0.99	8527	21	0.60
4							7367	19	0.53
5							13360	8	0.51
Pseudo R^2	0.50			0.93			0.81		

Note: We measure homogeneity within clusters through the "Average Silhouette Width." Comparing the average distance of an observation from the other members of its cluster and its average weighted distance from the closest group. Low values indicate low cluster homogeneity. The pseudo R^2 informs to what extent the cluster solution allows explaining sequences variability

TABLE 5: Prevalence of cluster solutions and measures of homogeneity for six the variables analyzed, males sample

	Work			**House**			**Family**		
Cluster	N	%	Homogen.	N	%	Homogen.	N	%	Homogen.
1	26575	86	0.71	8790	28	0.38	24316	79	0.59
2	4394	14	0.55	17473	57	0.57	2289	7	0.69
3				2511	8	0.29	2213	7	0.46
4				2195	7	0.60	2151	7	0.01
Pseudo R^2	0.58			0.62			0.58		
	General Life			**Health**			**Location**		
Cluster	N	%	Homogen.	N	%	Homogen.	N	%	Homogen.
1	27026	72	0.73	29602	95	0.98	5500	18	0.63
2	5978	27	0.17	861	3	0.40	10923	34	0.66
3				933	2	0.99	6753	21	0.58
4							5182	19	0.59
5							2611	8	0.48
Pseudo R^2	0.44			0.92			0.79		

Note: see 4

F ML Algorithms

When dealing with classification problems with a binary outcome, logistic regression models have been largely applied and have been proven to achieve high predictive performance, especially when compared to other simple probabilistic classification methods such as linear and quadratic discriminant analysis.

The logistic regression function looks like:

$$\hat{Y} = \frac{exp^{(\hat{\beta}_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j)}}{1 + exp^{(\hat{\beta}_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j)}} \tag{F.1}$$

Where $\hat{Y}$ is the probability that the target y is positive.

In the logistic regression model, the optimization criterion (loss function) is the log likelihood:

$$\hat{\beta} = \text{argmin}_{\beta \in R^p} \left\{ \sum_{i=1}^{N} y_i \ln(\hat{y}_i) + (1 - y_i)(1 - \ln(\hat{y}_i) \right\} \tag{F.2}$$

The logistic regression model has several advantages, as demonstrated by its wide use in real-world applications. Firstly, the logistic model allows an interpretation of the regression coefficients in terms of increasing the probability of a positive outcome. Furthermore, it is very efficient from a computational point of view.

However, when the number of predictors increases above a certain threshold, multicollinearity issues and the course of dimensionality limit the possibility of using logistic regressions. In this case, shrinkage methods such as Ridge, Lasso, and Elastic net can come to the aid.

Shrinkage methods act similar to subset selection methods because they reduce the number of initial predictors to a subset that has the highest predictive power while shrinking or setting all the other coefficients to zero. They are preferred to subset selection methods because they provide more robust results and being continuous shrinkage meth-

ods, suffer from less variability.

Shrinkage methods control for over-fitting by adding a penalization term $E_\beta(\beta)$ the loss function $\mathscr{L}$. The optimization criteria take the form of:

$$\hat{\beta} = \text{argmin}_{\beta \in R^p} \left\{ \mathscr{L}(\beta) + \lambda E_\beta(\beta) \right\} \tag{F.3}$$

where λ is the regularization coefficient that controls the importance of regularization; this parameter must be estimated through cross-validation (Friedman et al., 2001).

The form of the regularization term $E_\beta(\beta)$ determines the regularized models. The Ridge regression imposes the l_2-penalty to the coefficients such that $E_\beta(\beta) = || \beta ||_2 = \sum_{j=1}^{p} \beta_j^2$; the Lasso regression imposes the l_1 norm such that $E_\beta(\beta) = || \beta ||_1 = \sum_{j=1}^{p} | \beta_j |$. These two methods have been intensively used, and they differ essentially in the shrinkage effects they have on the parameters: the ridge regression shrink less important parameter towards zero while never setting them exactly to zero, and the lasso method, instead, allows the parameter to be exactly equal to zero thus implementing real variable selection.

Although these methods have shown success in many situations, they have some limitations. In the case $p > n$, e.g., the number of predictors is higher than the number of observations, the Lasso selects at most n variables before it saturates. Moreover, if a group of highly correlated variables exists, Lasso would choose only one variable from the group while neglecting the others. These limitations have been partly addressed by a new regularizer proposed by Zou H. and Hastie T. (2005) under the name of Elastic net.

The Elastic net penalty is a convex combination of the lasso and ridge penalties and takes the following form:

$$(1 - \alpha) || \beta ||_1 + \alpha || \beta ||_2 .$$

The elastic net solves the inner problem of parameter dimensionality of Lasso and Ridge but requires higher computational power. When $\alpha = 0$ the elastic net is equal to the ridge regression. When $\alpha = 1$, the net Elastic net penalty reduces to the lasso penalty.

Tree-based methods were popularized by Breiman, Freidman, Olshen, and Stone around

1984, and they have gained increasing attention in recent years.

Decision tree algorithms are non-parametric models that recursively segment the prediction space X into non-overlapping regions. This procedure gives rise to the "tree" structure that gives the algorithm its name. The partitioning approach divides the data into smaller subsets until the algorithm determines that the data within the subsets are sufficiently homogeneous.

The model equation for decision trees is:

$$\hat{y} = \sum_{i=1}^{M} c_m \cdot 1_{(X \in R_m)};$$

where $R_1, \ldots, R_M$ are disjoint partitions of the predictor space as resulting in the terminal nodes and c_m is a constant. The optimization criterion, in this case, is the average log-likelihood.

Various methods can calculate homogeneity within subsets. The Gini entropy for classification problems is the most widely used for classification problems and the sum of square errors for regression problems. Moreover, the algorithm requires tuning other hyper-parameters, namely the maximum depth, the minimum number of samples at the leaf nodes, and the maximum number of features to consider for splitting.

Three-based methods often yield good predictions on the training set but are likely to overfit the data, resulting in dire out-of-sample predictions. One way to solve this problem is by relying on ensemble learning. Ensemble learning is a learning paradigm that, instead of trying to learn one super-accurate model, focuses on training a large number of low-accuracy models and then combining the predictions given by those weak models to obtain a high-accuracy meta-model.

The Random Forest is an ensemble method built upon decision trees. It was originally developed by Leo Breiman in 2001 (Breiman, 2001).

The Random Forest optimization procedures first require randomly selecting an independent subsample (bootstrap) of the training sample. For each random sub-sample $b = 1, \ldots, B$, it builds a depth tree and estimates a prediction of the test sample. Finally, it averages over B to obtain a low-variance statistical learning model. The functional form

looks as follows:

$$\hat{y}_{avg} = \frac{1}{B}\sum_{b=1}^{B}\hat{y}^b$$

So far, the Random Forest procedure follows the "bagging" procedure. In addition to bagging, the Random Forest algorithm not only selects a random sub-sample of the training set but also performs a random selection within the predictor matrix, choosing at each iteration a subset of the predictor set, $\bar{X} \subseteq X$, of size m. In the presence of a strong predictor, this procedure allows other less significant predictors to be selected, reducing the correlation among trees and the variance of the learning algorithm. The parameter m is a hyper-parameter that we tuned with combined cross-validation and grid search.

Like the Random Forest, Gradient Boosting is an ensemble of decision trees (for an in-depth explanation, see Friedman, 2001). However, contrarily to Random Forest, which builds each tree independently, the Gradient Boosting procedure builds the tree sequentially. Each new tree helps correct the error from the previous by modifying the weight of the misclassified observations. The AdaBoost algorithm gives the simple version of boosting algorithm. The AdaBoost starts by training a simple decision tree where each observation has an equal weight. After evaluating the first prediction error, the algorithm increases the importance of the problematic observations and lowers the weight of the easy ones. Thus, the second tree is grown on the weighted data. The idea is then to learn and improve from the predictions of the previous tree. This procedure runs for a specified number of iterations. The final prediction is then a weighted average of all the predictions of this successive iterations. Extreme Gradient Boosting modifies this procedure by calculating gradients in the loss function, thus it can handle any differentiable loss function.

Artificial neural networks (ANN) belong to the algorithmic class of the so-called "black box" methods. An ANN models the relationship between the set of predictors and the output in a way that mirrors the process of reaction of the biological brain to external sensory input. Like the human brain, the ANN structure involves a network of interconnected artificial neurons that transform the initial input signal into an output signal.

ANNs have shown outstanding performance in image recognition and detection tasks. They can adapt to classification or numeric prediction problems. Their flexible structure allows for modeling more complex patterns than nearly any algorithm. Despite these significant advantages, ANN applications are scarce in social sciences, mainly due to the impossibility of interpreting the parameters of the optimized structure. In addition, they require substantial computational and data requirements.

The typical optimization procedure of ANNs is that of backpropagation. In its more general form, the backpropagation algorithm iterates several times in two sequential processes. The completion of this cycle is called an epoch. Each epoch consists of a forward phase and a backward phase. In the forward phase, the input features transmit through the network, transforming themselves through the combination of activation functions and weights until they reach the output layer, where a prediction or output signal is produced. All predictions are compared to the true target value in the training data to estimate a cost given a loss function. The backward phase consists in adjusting the connection weights by taking the derivative of the loss with respect to each connection weight; this technique is called gradient descent.

The complex training procedure of an ANN, over time will reduce the total error of the network but it is likely to overfit the data. Resulting in bad out of sample performance. To control overfitting in ANN various methods have been proposed. In this analysis we use the skip connections residual connection. The skip connection allows to construct regularized deep networks by skipping one layer in the network and feeds the output of one layer as the input to the next layers.

G Optimal hyper-parameters

Optimal hyper-parameters used in the different training data-set, for each gender.

TABLE 6: Optimal hyper-parameters selected through 10-fold cross validation. Female sample

panel A: Ridge				
Hyper-parameter	Baseline	Cluster	Features	Unstructured
λ	0.31	0.011	0.06	0.002
α	0	0	0	0
threshold	0.31	0.34	0.356	0.38
panel B: Lasso				
λ	0.31	0.013	0.183	0.147
α	1	1	1	1
threshold	0.31	0.32	0.356	0.39
panel C: Elastic Net				
lambda	0.67	0.004	0.431	0.025
alpha	0.5	0.46	0.46	1
threshold	0.33	0.34	0.35	0.38
panel D: Gradient Boosting				
Max depth	11	8	7	8
Min child weight	11	20	22	24
Max delta step	1	1	6	1
N estimators	50	85	67	72
Learning Rate	0.01	0.1	0.1	0.1
Threshold	0.47	0.47	0.345	0.371
panel E: Neural Network				
N. Epochs	50	3	3	4
Learning rate	0.01	0.01	0.01	0.01
N. Neurons	20	20-30	100-250	100-200
Batch size	1024	1024	1024	1024
Activ.function	"sigmoid"	"sigmoid"	"sigmoid"	"sigmoid"
N. Hidden Layers	3	4	5	5
N. Skip connection	1	2	2	2
Threshold	0.24	0.356	0.418	0.391

TABLE 7: Optimal hyper-parameters selected through 10-fold cross validation. Male sample

panel A: Ridge				
Hyper-parameter	Baseline	Cluster	Features	Unstructured
λ	0.014	0.013	0.019	0.026
α	0	0	0	0
threshold	0.27	0.249	0.259	0.268
panel B: Lasso				
λ	0.025	0.023	0.023	0.683
α	1	1	1	1
threshold	0.27	0.278	0.259	0.285
panel C: Elastic Net				
lambda	0.012	0.135	0.008	0.05
alpha	0.55	0.37	0.1	0.19
threshold	0.27	0.262	0.259	0.267
panel D: Gradient Boosting				
Max depth	17	6	6	10
Min child weight	9	23	19	14
Max delta step	7	4	4	5
N estimators	60	84	55	65
Learning Rate	0.1	0.3	0.15	0.01
Threshold	0.297	0.293	0.303	0.287
panel E: Neural Network				
N. Epochs	50	5	3	5
Learning rate	0.01	0.01	0.01	0.01
N. Neurons	20	20-30	200-250	150-300
Batch size	1024	1024	1024	2048
Activ.function	"sigmoid"	"sigmoid"	"sigmoid"	"sigmoid"
N. Hidden Layers	3	5	4	5
N. Skip connection	1	2	2	2
Threshold	0.24	0.273	0.300	0.265

H Predictive performances

In this section we report additional predictive performance metrics for the sequence features' predictor set.

TABLE 8: Predictive performance metrics for the features predictor set. Female sample

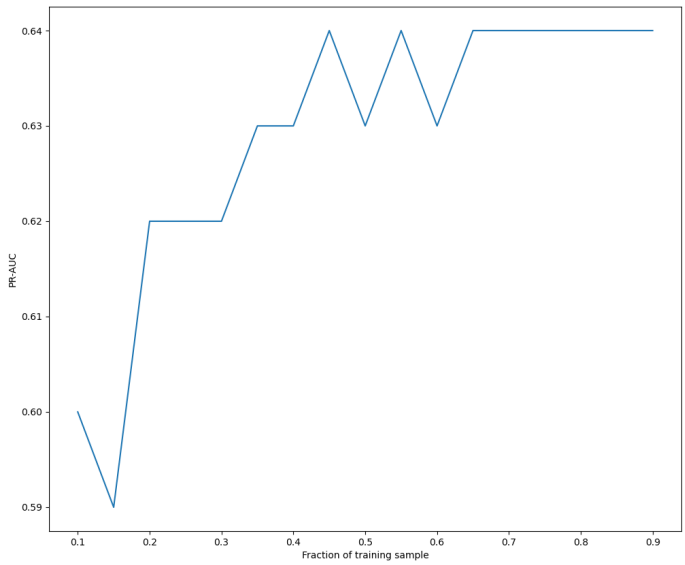
model	recall	accuracy	roc-auc	pr-auc	precision	threshold
logistic	0.651	0.639	0.685	0.673	0.627	0.355
ridge	0.659	0.637	0.684	0.672	0.622	0.356
lasso	0.661	0.639	0.685	0.673	0.625	0.386
elnet	0.658	0.640	0.685	0.673	0.626	0.352
XGBoost	0.655	0.640	0.695	0.682	0.641	0.385
ANN	0.707	0.638	0.695	0.675	0.597	0.418

TABLE 9: Predictive performance metrics for the features predictor set. Male sample

model	recall	accuracy	roc-auc	pr-auc	precision	threshold
logistic	0.250	0.711	0.666	0.453	0.519	0.249
ridge	0.282	0.713	0.668	0.457	0.524	0.259
lasso	0.282	0.713	0.668	0.457	0.525	0.258
elnet	0.283	0.713	0.668	0.457	0.524	0.259
XGBoost	0.213	0.716	0.668	0.460	0.544	0.303
ANN	0.245	0.717	0.672	0.451	0.540	0.300

I Sample size independence test

FIGURE 9: Test PR-AUC and training data dimensionality



J Subsamples

FIGURE 10: PR-AUC in the test sample for increasing EURO-D depression discrimination thresholds