A novel machine learning-based healthy ageing scale

Katarina Gašperlin Stepančič[1], Ana Ramovš, Jože Ramovš, Andrej Košir*

[a] *IBM Slovenija d.o.o., Ameriška ulica 8, Ljubljana*
[b] *Anton Trstenjak Institute of Gerontology and Intergenerational Relations, Resljeva 7, Ljubljana*
[c] *Anton Trstenjak Institute of Gerontology and Intergenerational Relations, Resljeva 7, Ljubljana*
[d] *Faculty of Electrical Engineering, University of Ljubljana, Tržaška 25, Ljubljana*

Abstract

Ageing is one of the most important challenges in our society. Evaluating how one is ageing is important for suggesting actions that could improve the ageing course and help in other domains, such as determining long-term care eligibility and accepting more informed policy decisions. This paper presents a novel scale for evaluating the healthy ageing of older adults in which a group of gerontology experts was involved in the scale development process. The main aspects of healthy ageing were factor-analyzed and presented via a web annotation application used by experts to provide healthy ageing expert ratings. This process was then modelled using machine learning. Data collected via personal field interviews as part of independent research, Ageing in Slovenia: Survey on the Needs, Abilities and Standpoints of the Slovene Population Aged 50 Years and Over, was used. The results show that the machine learning model could provide healthy ageing scores, thus reducing the cost, time and need for skilled assessors. Gerontology knowledge could be considered as integrated into the model by having experts present throughout the scale development. This helps move toward expert-based estimation, where a machine learning-based healthy ageing scale could be extended to decision support systems in clinical practice.

Keywords: healthy ageing, older adults, novel scale, machine learning, factor analysis, expert ratings

1. Introduction

According to the United Nations World Population Ageing 2020 Highlights report [1], the world continues to experience a sustained change in the population's age structure. People are living longer lives, and while both the share and the number of older people in the total population are increasing rapidly, the speed of population ageing will also likely continue to grow over the coming decades [2]. According to WHO [3], in 2020 the global population aged 60 years and over (older persons) was just over 1 billion people, representing 13.5% of the world's population of 7.8 billion. That number was 2.5 times greater than in 1980 (382 million) and is projected to reach nearly 2.1 billion by 2050. Population ageing has therefore been identified as one of the four global demographic megatrends [4], and good health with well-being at all ages was recognized as one of the goals in the 2030 Agenda for Sustainable Development [5].

Consequently, healthy ageing has been recently receiving considerable attention from governments, organizations and other stakeholders. WHO declared 2021-2030 a decade of healthy ageing, "the process of developing and maintaining the functional ability that enables well-being in older age" [3]. The report suggests that to enable healthy ageing monitoring across the life

*Corresponding author
Email address: andrej.kosir@fe.uni-lj.si (Andrej Košir)
URL: https://www.ibm.com/planetwide/si/ (Katarina Gašperlin Stepančič), https://www.inst-antonatrstenjaka.si/eng/ (Ana Ramovš), https://www.inst-antonatrstenjaka.si/eng/ (Jože Ramovš), https://www.lucami.org/ (Andrej Košir)

course, several requirements need to be met, such as more data standardisation for measuring healthy ageing and more innovation in collecting, analysing and using the information.

Many measurement tools and scales are available to measure healthy ageing domains. Therefore, some argue that a single scale might not be necessary. However, we see the development of a unified, healthy ageing scale with an emphasis on the involvement of gerontology experts in the overall process as the critical enabler that will not only help to ensure the validity of such a scale, but also provide a robust scale for use in healthy ageing-related applications, such as clinical decision systems and recommendation systems, where long and tedious procedures are not acceptable in terms of experts time and participant engagement.

This paper aims to propose and evaluate a novel domain-specific healthy ageing scale and its use as a target variable in developing a healthy ageing machine learning model. A combination of gerontology experts' opinions and a data-driven approach was used throughout the process of developing the scale, where the dataset used in the study consisted of answers to questions obtained via personal field interviews. The proposed scale is applicable for a single point in time healthy ageing estimation for adults aged 50+. However, the questions asked during the interview incorporate the relevant participant history into the estimation.

1.1. Research questions

The research questions that are addressed in this paper are the following:

(i) Can the domain-specific scale be developed for healthy ageing concerning the validity and reliability of psychometric properties?

(ii) Can the domain-specific healthy ageing scale be developed using an annotation process in which domain-specific experts provide their ratings via a purpose-developed web annotation application and where the scale is obtained by combining ratings using the ground truth procedure?

(iii) Can explanatory factor analysis be applied to our data to find relevant constructs that define healthy ageing for the annotation process?

(iv) Which are the main aspects influencing healthy ageing of adults aged 50+?

(v) Can we model the obtained healthy ageing scale using a machine learning approach and evaluate it using standard evaluation metrics?

(vi) Could we also create a machine learning model for the healthy ageing scale by using raw data from the survey, or should we factor analyse it first?

1.2. Paper's contributions

To summarize, the main contributions of this paper are the following:

(i) The construction of a novel domain-specific healthy ageing scale using an existing dataset with quantitative and qualitative health-related survey results of the population aged 50+ collected by the Anton Trstenjak Institute of Gerontology and Intergenerational Relations;

(ii) Gerontology experts-based selection of a set of constructs and items that best represent healthy ageing and, for the rating process, development of the web annotation application for visualizing this set of data;

(iii) Development of a machine learning model with a novel healthy ageing scale as the target variable, while testing how the model's performance depends on the input data. Three combinations of input variables were used: raw data (all items that were input into the explanatory factor analysis process), a subset of raw data (only items used directly in the web annotation application and items influencing constructs that were used

in the application) and finally, items and constructs used in the application.

2. Related work

When investigating the currently available healthy ageing measurement tools, one may find it hard to get a unified definition of healthy ageing and decide on the right tools to measure it. Many definitions and instruments are available covering different aspects of healthy ageing. Related to the interdisciplinary approach to research on older adults and healthy ageing measurement tools, machine learning has been highlighted as one of the important technological enablers. Additionally, to validate the developed constructs, the need for the early involvement of domain experts and stakeholders has been seen as crucial.

Today, according to our knowledge, there is no unidimensional machine learning-based healthy ageing scale, where the scale development process would also involve close cooperation with experts in the gerontology domain, which is one of the goals of this study.

This section presents the most relevant healthy ageing definitions, measurement tools, techniques for human-annotated data acquisition, and machine learning usage in the field of older adults.

2.1. Healthy ageing and its domains

Healthy ageing has been a central topic in many countries and research studies in recent years. It leads to an improved quality of life, decreased health care consumption, and contributes to the labour supply, decreasing the likelihood of early retirement [6]. Definitions of healthy ageing vary. The basic definition of healthy ageing is described as the general condition of the ageing of a person's mind and body, usually meaning freedom from illness, injury, or pain [7]. [8] describes a definition of healthy ageing, obtained via semi-structured interviews, as the ability to go and do a meaningful activity. [6] defines healthy ageing as optimising opportunities for physical, social and mental health to enable older people to take an active part in society without discrimination and enjoy independence and good quality of life. The World Report on ageing and health by WHO [9] defines healthy ageing as the process of developing and maintaining the functional ability that enables well-being in older age, where well-being is considered in the broadest sense and includes domains such as happiness, satisfaction, and fulfilment. Healthy ageing is also used interchangeably with terms such as "active", "successful", or "productive" ageing [10]. Successful ageing is a term often used in the gerontological literature to cover ageing processes throughout the life span [11]. It describes critical ideas such as life satisfaction, longevity, freedom from disability, mastery and growth, active engagement with life, and independence. Rowe and Kahn's definition of successful ageing is based on three standards which are "low probability of disease and disease-related disability and related risk factors", "high cognitive and physical functional capacity", and "active engagement with life" [12]. In 2015, Rowe and Kahn suggested adding societal-level principles to evaluate healthy ageing, which includes more opportunities for employment and volunteering, thus creating new roles and responsibilities for older adults, as well as inclusion of older people in some other forms of civic engagement to use wisdom and talent based on ability rather than chronological age [13]. According to a review of healthy ageing definitions and measures [14], a comprehensive health outcome should measure how well a human can function and adapt to environmental challenges in domains assessing physical, mental and social well-being.

Based on lifestyle-based intervention studies, five fundamental domains of a "healthy ageing phenotype" were developed by [15] in order to help guide research in the area of healthy ageing: physiological and metabolic health, physical capability, cognitive function, social wellbeing, and psychological well-being. Additional areas such as general health status, security, and health behaviours are proposed based on the healthy ageing literature review [10].

2.2. Healthy ageing measurement tools

Ageing is a complex process that depends on many factors; therefore, no unified measure of healthy ageing exists; and, as per [11], there is no reference criterion for assessing healthy ageing. However, in its review, [14] states that a standard for defining and quantifying the concept of healthy ageing is needed. Most studies on healthy ageing are multidimensional, and many standardized instruments exist that combine information from tests, the ability to perform tasks, biomarkers, and subjective assessment [7]. The current efforts to assess the health of older adults are mostly using items drawn from 4 categories [16]: on fulfilling or performing functions, activities, or roles (basic activities of daily living, instrumental activities of daily living, advanced activities of daily living); items reflecting the WHO definition of health and well-being (describing physical, social and mental aspects of health); symptom-oriented; and those concerned with adaptation or coping with non-fatal health conditions or limitations. The most often used instruments to measure the basic activities of daily living are the Katz Index [17], which evaluates six functions (bathing, dressing, toileting, transferring, continence, feeding), and the Barthel Index [18], which evaluates ten functions. Lawton's scale is one of the assessment instruments for instrumental activities of daily living that describes one's ability to keep an independent household. In contrast, open-ended questions are often asked for advanced activities of daily living. In the review of healthy ageing measurements used in epidemiological studies[10], many apply Rowe and Kahn's three standards to assess healthy ageing. In contrast, some use the WHO's active ageing model or Kul's theory of healthy biological ageing. Furthermore, sourcing from [10], the most often used measurements to evaluate physical capabilities were instrumental activities of daily living. To evaluate cognitive function, Mini-Mental State Examination was used most frequently.

2.3. Human-annotated data acquisition process

Obtaining an annotated or labelled training dataset can be one of the most time-consuming parts of applying machine learning but, on the other hand, also an important factor in its success. Various strategies for collecting labels can be applied depending on the field, from using domain expert human raters to involve people from the general public (crowdsourcing) [19] or using data programming frameworks [20].

In the case of multiple human raters, the ground truthing process is required to obtain a single ground truth value. An important aspect of it is to assess reliability and inter-rater agreement. Reliability is influenced by factors such as the type of rating procedure and rater-specific distortions such as level of expertise and domain knowledge, personality and perceptiveness [21]. Various reduction concepts exist, such as majority reduction, reduction based on probability, or three-way decision theory[19]. The design principles of software applications for annotation purposes as well as the amount and presentation of content are also influenced by the cognitive load theory (CLT) as well as human-computer interaction (HCI) principles [22], which both share basic assumptions of the human cognitive system and a need to reduce irrelevant load. Also, according to HCI principles, an application's usability level requires an in-depth understanding of its target users and the specific tasks they need to accomplish. This study used the annotation approach of multiple human raters with domain expertise.

2.4. Machine learning use in older adults domain

The literature review shows that the use of machine learning in healthy ageing is a current topic. Machine learning has been widely used in research focusing on older adults and, related to healthy ageing, [23] mentions explicitly a need for a new, more holistic and interdisciplinary approach where data science and machine learning are highlighted as helpful enablers.

[24] created the unidimensional metric of healthy ageing comprised of 45 items on self-reported health,

where factor analysis and Bayesian multilevel Item Response Theory were used. [25] used six machine learning algorithms, including ensemble, to develop predictive models for successful ageing, where features were defined based on Rowe and Kahn's theory. In both, machine learning was used to find relationships between calculated health scores and the items. [26] developed a machine learning-based clinical decision support system that predicts the quality of life considering the physical, psychiatric, and social factors. Machine learning has also been used to estimate the biological age of the organism using biomarkers as input features [27], where blood-based and brain-based biological ages show the best performance in terms of accuracy and predicting mortality risk. Machine learning also tackled predicting specific age-related conditions such as dementia [28] and Alzheimer's disease [29]. It has also been widely used in developing ambient-assisted living systems: applications such as anomaly detection of daily activities, changes in behavioural patterns, and mild cognitive impairment detection, where data from wearable, ambient, or IoT sensors was used [30].

Reviews on the use of machine learning throughout various domains also bring recommendations for further research. Specifically, the study of machine learning use in the mental health domain [31] suggests that for more effective and implementable machine learning systems, more research would be needed to (i) test the validity of the developed constructs and (ii) ensure that machine learning outputs are robust enough to be used in practice (reliability). It also presents the need to involve target users and key stakeholders early to reach system acceptance. It emphasizes that domain experts can provide critical insights into construct validity, ground truth and biases assessments, and important contextual information that can help interpret data findings, improve rigour, and manage deployment risks and tradeoffs.

3. Materials and methods

3.1. Dataset

The dataset used in this research was obtained by Anton Trstenjak Institute of Gerontology and Intergenerational Relations (further referred to as the institute), a Slovenian national scientific, research, expert, and end-user institution within the gerontology and good intergenerational relations field in Slovenia. Data were collected through personal field interviews conducted using specially trained interviewers. The institute developed the extensive questionnaire in the scope of the independent research "Ageing in Slovenia: Survey on the Needs, Abilities and Standpoints of the Slovene Population Aged 50 Years and Over" [32]. The National Medical Ethics Committee of the Republic of Slovenia considered the questionnaire and the research concept, and an opinion was issued that the research was ethically impeccable. Ethical consent (nr. 115/09/09) was issued for its implementation [33]. Special methodological attention was paid to the respondent's motivation for the selected sample and the training and monitoring of interviewers and data entry into the database.

The dataset captures information about the standpoints, needs, and potentials of the Slovenian population aged 50+. It involves quantitative and qualitative data and covers topics of physical health, health strengthening, taking drugs, public health, everyday chores and mobility, accommodation adjustment, interpersonal relations and long-term care, mental health and attitudes, intergenerational solidarity, local community and living, employment and retirement, family, demography. It holds information on 1047 participants of the survey, who are a representative sample of Slovenians aged 50+, out of which 41.3% is women and 58.7% is men. The average age of the participants was 66.03 years. The youngest participant was 50 years old and the oldest was 98 years old [32].

The targeted population for this paper's proposed metrics is people aged 50+ with demographic charac-

teristics that meet the dataset characteristics in terms of age, gender, and education.

3.2. The development of the healthy ageing scale

The dataset acted as the basis for developing the healthy ageing scale. The most relevant items, each representing a question from the survey, were selected by gerontology experts and put into multiple categories and sub-categories. The process is graphically described in Fig. 1.

3.3. Explanatory factor analysis

As part of the scale development process, explanatory factor analysis was conducted to identify factors and find underlying relationships between groups of items [34].

3.3.1. Selection of items entering explanatory factor analysis

The original dataset captures information on various aspects of older adults. Therefore, before conducting explanatory factor analysis, we asked gerontology experts to compile a list of categories and subcategories of items that, based on their domain knowledge, would reflect how well a person is ageing.

3.3.2. Explanatory factor analysis

Explanatory factor analysis was performed for each category or, instead, sub-category if the category had one. Once the correlation matrix was constructed, PCA was performed to extract factors. The number of factors retained was determined using the parallel analysis method[35] [36]. Explanatory factor analysis was done using standard R packages corrplot and psych.

3.3.3. Selection of healthy ageing scale constructs

After obtaining category or sub-category factor matrices, a detailed discussion was held with gerontology experts to find and define relevant constructs and items (Tab. 2) which should be part of the healthy ageing scale.

3.4. Annotation of how well the person is ageing

A special web annotation application was developed to capture gerontology expertise in defining a healthy ageing scale. The web application was developed using the Django framework and Python programming language. Data was stored in the SQLite database, a default database used with Django applications. The purpose of the application was to provide a user-friendly interface for raters who used it to rank how each person in the dataset is ageing. The application included three main screens: the registration screen, the login screen, and the annotation screen. The annotation screen is presented in Fig. 2. It visualizes a set of constructs and items chosen by gerontology experts as ones that best describe how well the person is ageing.

During the construction of the web annotation application, a feasible cognitive load of raters was taken into account to include only the amount of information that the rater can work with during the annotation process. The amount of information and how information was visualized on the application were validated by four test raters prior to the rating.

Randomization was used so that each rater who annotated older adults had his or her order of persons to be annotated. The reason for using randomization was to eliminate cross-annotated elderly effects.

An initialisation process was used to prevent raters from calibrating their annotations based on the first annotations, during which each rater annotated thirty different records. Randomly selected records also included records with extreme values.

A training session as well as a web annotation application usage guide were prepared for raters prior to the rating. Four raters participated in the annotation process.

3.5. Ground truth for a healthy ageing scale

As a result of the annotation process, multiple healthy ageing ratings were obtained for each older

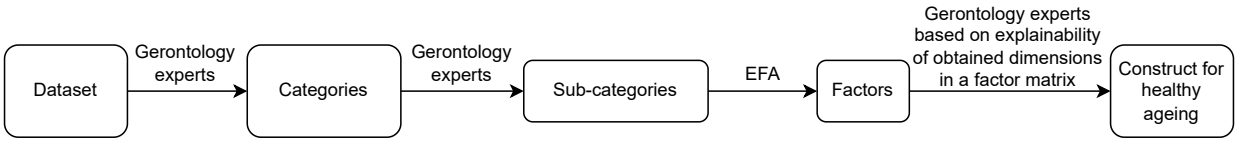

Figure 1: Diagram showing the general process and selection methods for obtaining healthy ageing constructs and items for placement onto the web annotation application. Items from the dataset were first organized into healthy ageing-related categories and sub-categories by gerontology experts. Second, explanatory factor analysis (EFA) was conducted on each sub-category. Third, gerontology experts reviewed EFA factors and selected a subset of factors and individual items that best describe how the person is ageing according to their domain knowledge and expertise. The set of chosen factors and items, a healthy ageing construct, was further used in the annotation process.

adult in the study. A ground truth determination procedure was used to get a one-dimensional healthy ageing scale from multiple ratings. It is used when human annotations provide the most reliable means of obtaining ground truth and there is no direct empirical evidence of the observed construct. This procedure reduces rater bias and maximizes inter-rater agreement, as described in [21]. Annotator bias removal procedure from [21] was applied. The inter-rater agreement was also measured using Krippendorff's alpha [37], a reliability coefficient that measures the agreement among multiple raters. A value of Krippendorff's alpha can be between zero and one, where zero means perfect disagreement (raters agree as if chance had produced the results) and one means perfect agreement.

3.6. Machine learning for healthy ageing scale modelling

This section describes how machine learning was used to create a classification model which predicts how healthy the person is ageing based on his or her health data. This step aims to show that a one-dimensional healthy ageing scale obtained via ground truth procedure from annotations can be successfully modelled using machine learning techniques. The model could then, in the future, be used to rank unseen older adults' data instead of using gerontology experts for time-consuming manual ranking.

3.6.1. Input variables

Three sets of input variables were used while developing the machine learning model for the healthy ageing scale. Using multiple sets, we wanted to evaluate which set of input variables has the best predictive power in terms of area under the curve evaluation metric and consequently determine which one should be used for collecting and processing information about new subjects for which healthy ageing scores would need to be generated.

(i) **Set 1**. Input variables were eight information units presented on the web annotation application screen (factors, individual items and calculation obtained by a sum of the items).

(ii) **Set 2.** Input variables were eight information units presented on the web annotation application screen. However, the items determined to influence those factors during the explanatory factor analysis process were used instead of using the factors themselves.

(iii) **Set 3.** Input variables were all items that were the input into the explanatory factor analysis.

3.6.2. Target variable

The target variable of machine learning modelling was the healthy ageing scale. The healthy ageing scale was calculated as described in 3.5.

3.6.3. The selected classifier: XGBoost

The classifier used for building a machine learning model was XGBoost, a scalable machine learning system for tree boosting. XGBoost open-source library in Python was used. XGBoost (Extreme Gradient Boosting) provides a reliable and efficient implementation of

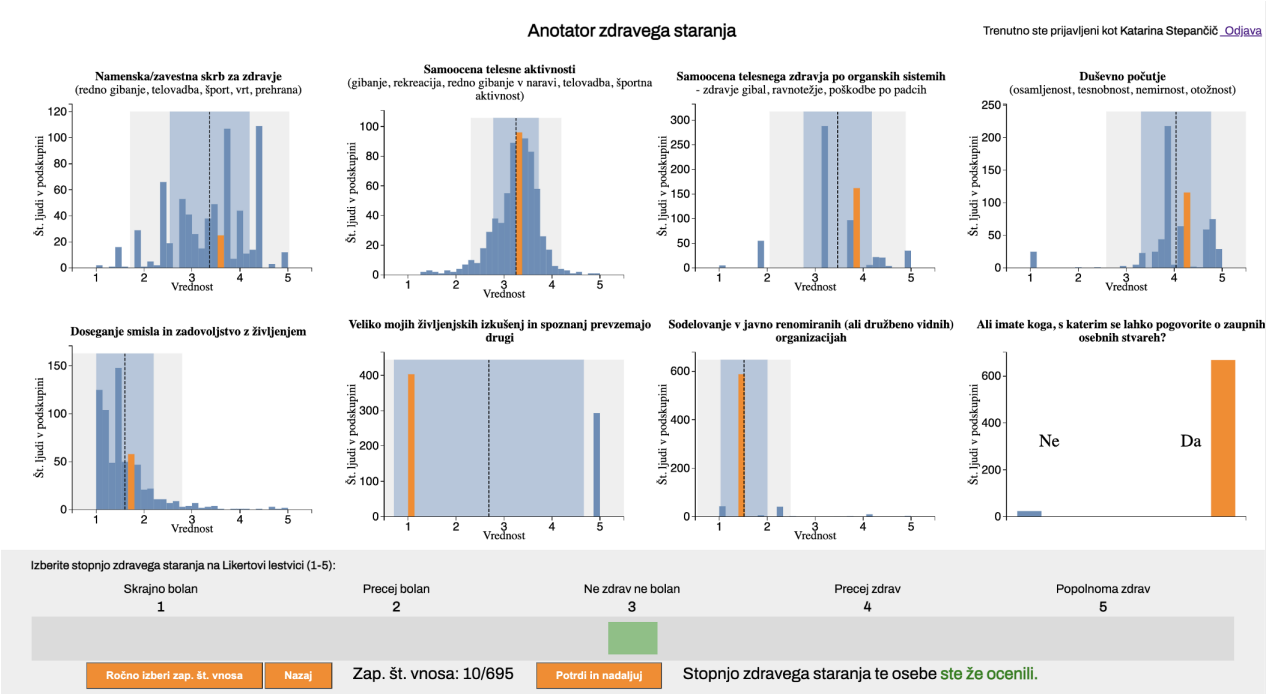

Figure 2: The annotation screen holds graphical information about eight healthy ageing constructs of a person (from left to right, starting in the upper left corner): one's conscious care for health, one's self-evaluation of physical activity, one's self-evaluation of body health according to organic systems, mental well-being, achieving purpose and life satisfaction, "Many of my life experiences are summarized by others?", one's participation in publicly renowned and socially visible organizations, and "Do you have someone with whom you can talk about private and personal topics?". Graphs contain mean values for each construct and coloured intervals of three and five standard deviations to identify outliers and extreme values. A Likert scale from 1 (very ill) to 5 (very healthy) was used by raters to determine the level of healthy ageing for each participant in the study. Values for the person being rated are coloured in orange.

the gradient boosting algorithm and is often used as the component in many winning solutions in machine learning competitions [38].

XGBoost is a decision tree ensemble machine learning algorithm based on gradient boosting and is designed to be highly scalable [39]. It aims to accurately predict a target variable by combining a set of smaller, simpler, and weaker learners into a strong learner in an iterative way. In order to control the overfitting, the regularized objective (minimization) function L consists of two parts.

$$L(\phi) = \sum_{n=1}^{N} l(y_i, F(x_i)) + \sum_{m=1}^{M} \Omega(f_m)$$

where

$$\Omega(f_m) = \gamma T + \frac{1}{2}\lambda ||\omega||^2$$

$l(y_i, F(x_i))$ is the differentiable convex loss function that measures the difference between the prediction y_i and the target $F(x_i)$. The regularization term Ω penalizes the complexity of the model, where T is the number of leaves in the tree and ω are the output scores of the leaves. The value of γ controls the minimum loss reduction gain needed to split an internal node. Higher values of γ result in simpler trees. As the XGBoost algorithm can suffer from over-fitting if the iterative process is not properly regularized, there are various other parameters we can configure to prevent it. Regularization can be achieved by applying a shrinkage (learning rate) to reduce each gradient descent step. Additional regularization can be applied to reduce the complexity of the trees by limiting the tree depth and by using randomization techniques such as random subsampling (without replacement) to create individual trees and column subsampling at the tree and tree node level.

The following hyperparameters were tuned for XG-

Boost in our machine-learning process:

- The learning rate (learning_rate) or shrinkage.
- The maximum depth of the tree (max_depth).
- The number of estimators.
- The sampling rate (subsample) for the size of the random samples (training instances). Subsampling will occur once in every boosting iteration.
- The sampling ratio of columns when constructing each tree (colsample_bytree). Subsampling occurs once for every tree constructed.
- The minimum sum of instance weight needed in a child (min_child_weight). The larger min_child_weight is, the more conservative the algorithm will be.
- The minimum loss reduction required to make a further partition on a leaf node of the tree (γ). The larger gamma is, the more conservative the algorithm will be meaning the shallower the trees.

Selected values are summarized in Sec. 4.5.2.

3.6.4. The machine learning process configuration

After choosing input and target variables for a machine learning process, a stratified 10-fold cross-validation was performed [40] to split data in train and test sets. A stratified k-fold was used to preserve the percentage of samples for each class in the target variables.

Additionally, values for hyperparameters were selected. Some hyperparameters (learning_rate, number of estimators) were set to fixed values. In contrast, values for others (gamma, subsample, colsample_bytree, min_child_weight, max_depth) were determined via a grid search procedure which was used to find an optimal combination of hyperparameters for creating the best model.

3.6.5. Evaluation metric

Model performance was evaluated using the standard metrics: the area under the receiver operating characteristic curve (AUC) evaluation metric [41], F1 score, precision, and recall. Values of AUC can range from 0.5 (no predictive ability) to 1 (perfect predictive ability). Due to the multi-class classification problem, both One-vs-one (OvO) and One-vs-rest (OvR, also referred to as One-vs-all or OvA) strategies were used when calculating the area under the curve to select the best strategy [42]. The OvO approach splits the multi-classification problem for each class versus every other, so one classifier is learned to discriminate between each pair. Then the outputs of these base classifiers are combined to predict the output class. OvR splits the multi-classification problem into learning a classifier for each class, so the base classifiers giving a positive answer indicate the output class. For aggregated evaluation across three categories, we used the macro-average value, which calculates AUC independently for each category and then creates an average. The macro-average was chosen over the micro-average due to class imbalance in our data where the macro-average is less sensitive and considers each category equally[43]. Similarly, the F1, precision and recall score are common measures that rate a classifier's success. F1 score aggregates precision and recall measures under the concept of harmonic mean. Their value can range from 1 (best) to 0 (worst). An averaging method can access a single F1 score, precision and recall for easier comparison in a multi-classification problem. Macro-average was selected [**?**].

4. Results

4.1. Selection of participants for the study

The Anton Trstenjak Institute of Gerontology and Intergenerational Relations dataset captures information about 1047 adults aged 50+. According to the targeted population, we removed all participants who did not meet the requirements. Requirements were that for

all participants in the study, complete data should be available. At the same time, the characteristics of the subsample population in terms of demographics (age, gender, and education) should be preserved.

The procedure resulted in a subsample of 696 participants. Fig. 3 compares age histograms across all participants in the dataset and a subset of participants used in our study. A two-sample nonparametric Kolmogorov-Smirnov test was performed to compare the selected sample's age distribution with the original-sized dataset's age distribution. $p - value > 0.05$ confirmed the two distributions come from identical populations. The subsample includes 41.5 % of women and 58.5 % of men. The mean value of education level in a subsample is 3.17, while in the overall dataset is 3.13.

4.2. Selection of categories and sub-categories by gerontology experts

During the operationalization process, the most important categories and sub-categories that define healthy ageing were selected by gerontology experts. The categories were selected based on their domain knowledge and the analysis they performed during their study of the survey results.

A summary of selected sub-categories and their descriptions are provided in Tab. 1. Selected domains match those mentioned as common among the healthy ageing studies review: physical, social and mental [14].

4.3. Explanatory factor analysis results

Explanatory factor analysis was performed for each category or sub-category, depending on whether the category had sub-categories. The PCA method was used in explanatory factor analysis, and multiple combinations of factoring methods (WLS, Minres) and rotations (no rotation, Varimax, Quartimax, Promax) were tested.

Results were discussed with gerontology experts who provided feedback on factor interpretations, a selection of factors, and additional items for placement in the web annotation applications. Five factors originating from EFA results of five different sub-categories along with two individual items from the dataset and one value calculated from multiple items were selected for the web annotation application. A summary of the selected information, along with the information type, is summarized in Tab. 2. Additionally, the factoring method and rotation method are provided for factors.

A list of items with their corresponding factor loadings for each construct is given in Tab. 3.

Kaiser-Meyer-Olkin (KMO) test and Bartlett's test of sphericity were performed to measure the suitability of the data for EFA. KMO values are given in Tab. 4, indicating good sampling adequacy. Bartlett's test yielded a low p-value of $p < 0.01$ for all models, indicating that the data are suitable for dimensionality reduction such as EFA, see Tab. 4.

4.3.1. Psychometric characteristics

Validity

The data used in this research was collected via a questionnaire that gerontology experts designed. The validity of the healthy ageing scale development was obtained via the construction process, where a focus group with four gerontology domain experts was used to establish the validity of the findings. The focus group was involved consistently throughout the process by determining the relevant sub-categories for healthy ageing, defining constructs, confirming the web annotation application design, and acting as raters.

Reliability

To select a proper measurement model, we applied the Chi-square difference test, eliminated the more restricted measurement models (e.g., parallel, tau-equivalent), and chose a unidimensional, congeneric measurement model to assess the reliability of the proposed models [44]. All obtained p values of the Chi-square test were $p < 0.01$.

To verify the variability of the proposed models, we estimated congeneric reliability ρ_C (reliability coefficient of a congeneric model), McDonald's ω (the

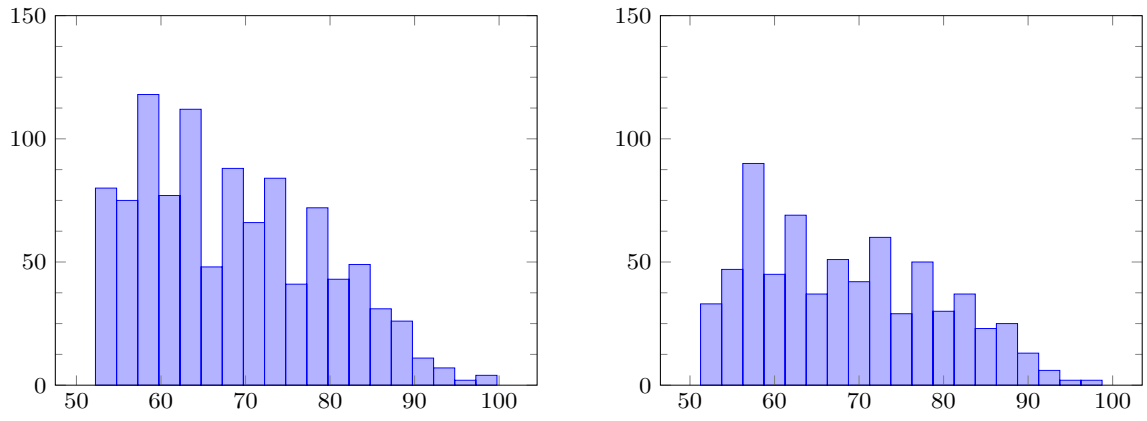

Figure 3: Histograms showing age across all participants (left) and a subset of participants used in our study (right) are similar. The Kolmogorov-Smirnov test confirmed that the two distributions come from identical populations.

Category: Sub-category	**Description**
Physical health: Basic physical health	Vital human body functions
Physical health: Advanced physical health	Person's lifestyle
Social health: Family	Person's relationship with his/her family
Social health: Society	Person's involvement in society (job, organizations)
Mental health: Basic mental health	Person's well-being, loneliness, and memory
Mental health: Advanced mental health	Is a person reaching their purpose and happy with his/her life?
Activities: Physical activities	Is the person physically active?
Independent living: Independent Living	Can a person take care of his/her daily activities like feeding and walking?

Table 1: Short compact descriptions of categories and sub-categories that were, in the process of operationalization, selected by gerontology experts as important for the definition of healthy ageing.

proportion of variability extracted by the model), and reliability coefficients Cronbach's alpha.

The psychometrics characteristics of five explanatory factor analyses are given in Tab. 5. Note that the reliability coefficient Cronbach's α does not meet assumptions of the congeneric measurement model, but we still list it for better comparability to other studies.

4.4. Selection of annotation screen and annotation procedure

Eight information units were determined to be presented on the web annotation application for each person as specified in Tab. 6. Possible raters' cognitive overload was considered by including only the amount of information an annotator can work with during the annotation process. The selection of information for the screen was done in close cooperation with gerontology experts who selected the information that would help them to most accurately evaluate how the person is ageing.

Information was presented in a graphical way using histograms and distribution graphs with descriptions as presented in Fig. 2. Each histogram visualized the distribution of values for all the people being annotated and highlighted the bar where the value for the person currently annotated is located. The histogram also included statistical information such as mean value, one standard deviation span, and two standard deviation spans.

Information displayed on the annotation screen	Information type	Factoring method	Rotation
Dedicated/conscious health care	Factor	WLS	Quartimax
Self-assessment of physical activity	Factor	WLS	Quartimax
Self-assessment of physical health by organ systems	Factor	Minres	Varimax
Mental well-being	Factor	Minres	Quartimax
Achieving meaning and satisfaction with life	Factor	WLS	Varimax
Many of my life experiences and insights are taken over by others	Item	/	/
Participation in organizations according to the type of organization	Calculation	/	/
Do you have someone to talk to about confidential, personal matters?	Item	/	/

Table 2: Table summarizes five factors, two items and one calculation (calculation was obtained using the sum of the values from multiple items) selected for placement on the web annotation application. It also provides the information type, and in case the information was obtained via EFA, the factoring method and rotation are provided as well.

The scale used in the annotation process to determine the level of healthy ageing was the Likert scale. A 5-point Likert scale was chosen. Values had the following meaning: 1 - very ill, 2 - mostly ill, 3 - neither ill nor healthy, 4 - mostly healthy, and 5 - healthy. Visualization of information, as well as the selected Likert scale, were both confirmed by gerontology experts.

Four raters with gerontology expertise participated in the annotation process during which each of them provided a Likert value (healthy ageing) for every person included in the study. Annotators were also able to annotate a specific person multiple times. In this case, the person's last result was valid. Before the annotation process began, the initialization process was completed as described in Sec. 3.4. Krippendorff's alpha that measures inter-rater agreement was 0.59.

4.5. Healthy ageing scale machine learning model

4.5.1. Target variable preparation

The target variable of the machine learning modelling was the healthy ageing scale created from the annotation results using the ground truth procedure as described in Sec. 3.5. The obtained ground truth was the categorical variable with values ranging from 1 to 5 increasing by 0.5 (span from 1 to 5 was due to a 5-point Likert scale). Reclassification was applied to reduce the number of categories in the target variable. New classes were defined by reclassifying the original values into three more meaningful categories representing poor, medium, and good healthy ageing categories. The resulting proportions of the target variable's poor, moderate, and good healthy ageing categories are shown in Tab. 7. Due to an unbalanced dataset, the synthetic minority oversampling technique (SMOTE) was used [45]. SMOTE is a method in which the minority class is over-sampled by creating synthetic data points that are moderately different from the original.

Class	%
Poor healthy ageing	0.160
Moderate healthy ageing	0.239
Good healthy ageing	0.521

Table 7: Target variable class representation

4.5.2. Machine learning configuration settings

There were multiple settings that we configured in the process. The objective chosen for machine learning was multi:softprob. As a result, it returns the predicted probability of each data point belonging to each class. The value for the learning rate was set to 0.1, and the number of estimators was set to 600. A grid search procedure determined val-

Construct	Items	Factor loadings
Conscious health care	How do you strengthen your health and maintain your physical strength? I pay attention to a suitable diet.	0.292
	How do you strengthen your health and maintain your physical strength? I regularly exercise in nature (walking, running).	0.496
	How do you strengthen your health and maintain your physical strength? I regularly exercise.	0.242
	How do you strengthen your health and maintain your physical strength? I regularly do sports.	0.193
	How do you strengthen your health and maintain your physical strength? I am gardening.	0.337
	How do you strengthen your health and maintain your physical strength? Regardless of whether I do the above, I don't consciously care for my health.	-0.428
Self-assessment of physical activity	How many hours did you spend yesterday (a normal working day is meant - if yesterday was a holiday, keep in mind one of the previous working days) sleeping and resting?	-0.399
	How many hours did you spend yesterday (a normal working day is meant - if yesterday was a holiday, keep in mind one of the previous working days) on movement, recreation, and entertainment?	0.749
	How many hours did you spend sleeping and resting last Sunday?	-0.426
	How many hours did you spend on movement, recreation, and entertainment last Sunday?	0.771
	How do you strengthen your health and maintain your physical strength?: I regularly exercise in nature (walking, running).	0.295
	How do you strengthen your health and maintain your physical strength?: I exercise regularly.	0.178
	How do you strengthen your health and maintain your physical strength?: I do sports regularly.	0.274
Self-assessment of physical activity by organ systems	On the scale, rate your health for the domain of movement. Consider last year as overall and not only current status.	0.464
	On the scale, rate your health for the domain of balance. Consider last year as overall and not only current status.	0.693
	Have you ever injured yourself in a fall that left you unable to do your work and regular activities for more than three days?	-0.149
Mental well-being	Rate how often it happens to you that you feel lonely.	0.630
	Rate how often it happens to you that you feel anxious.	0.805
	Rate how often it happens to you that you feel restless.	0.644
	Rate how often it happens to you that you feel saddened.	0.755
	On the scale, rate your health for the mental health domain. Consider last year as overall and not only current status.	0.478
Achieving meaning and satisfaction with life	Today, it is often heard that man also has spiritual needs and spiritual abilities. What is your opinion on this? I believe that man also has spiritual needs and abilities.	0.959
	Today it is often heard that man also has spiritual needs and spiritual abilities. What is your opinion on this? I do not deal with whether a person also has spiritual needs and abilities.	-0.941

Table 3: A summary of items with corresponding factor loadings for constructs selected for the healthy ageing scale.

ues of other parameters (gamma, subsample, colsample_bytree, min_child_weight, max_depth). The ranges being explored in the grid search are presented in Tab. 8.

Hyperparameter	Values range
min_child_weight	[4, 5, 6, 8, 10, 12]
gamma	[0.1, 0.2, 0.3, 0.4, 0.5]
subsample	[0.6, 0.7, 0.8, 0.9, 1.0]
colsample_bytree	[0.6, 0.7, 0.8, 0.9, 1.0]
max_depth	[4, 5, 6]

Table 8: List of parameters and their range for grid search procedure

Tab. 9 shows the selected values for the hyperparameters for all three machine-learning input variable sets.

4.5.3. Evaluation of the machine learning model

Performance results for all three models built are presented in Tab. 10.

5. Conclusion and discussion

This paper presents the creation of a domain-specific healthy ageing scale. Unique to the scale is that gerontology domain experts were involved throughout the whole process of scale development, which also

Topic	Bartlett	KMO	Factoring method	Rotation method	Dimension
Physical activities	< 0.01	0.73	WLS	Quartimax	2
Advanced physical health	< 0.01	0.61	WLS	Quartimax	4
Basic physical health	< 0.01	0.85	Minres	Varimax	5
Basic mental health	< 0.01	0.82	Minres	Quartimax	2
Advanced mental health	< 0.01	0.75	WLS	Varimax	3

Table 4: Factorisability (Bartlett test, Kaiser-Meyer-Olkin (KMO) Test for Sampling Adequacy), factorisation method and rotation method applied for each of the selected topics of individual healthy ageing. One EFA was performed for each of the listed topics. Estimated number of dimensions are also added.

Category/sub-category	Cronb. α	Congen. ρ_C	McDon. ω
Physical activities	0.81	0.84	0.52
Advanced physical health	0.71	0.76	0.58
Basic physical health	0.77	0.82	0.71
Basic mental health	0.82	0.88	0.63
Advanced mental health	0.74	0.79	0.69

Table 5: Psychometric characteristics of five Explanatory factor analyses (EFA) applied. Selected factors of these EFAs were used to present the older adult to the raters of his healthy ageing.

provides validity to the overall scale development approach. The scale comprises five healthy ageing domains that gerontology experts selected as necessary. These domains are physical health, social health, mental health, physical activities and independent living. Explanatory factor analysis was used on these domains to find relevant constructs for visualization in the web annotation application, used for the healthy ageing rating. The unidimensional, congeneric measurement model was used to assess the reliability of the constructs, and Chi-square tests were applied. Eight information units were identified as key for placement on the application. These were a combination of constructs, individual items, and calculations derived from multiple items. The design of the application was confirmed through a discussion with gerontology experts. The application visualized data for each older adult participating in the study in relation to the overall study target population data. Multiple raters with gerontology backgrounds used the application to rate how well one is ageing on a Likert scale from 1 to 5. The ground truth procedure was applied to get the single value per older adult from multiple ratings. The obtained ground truth, summarized into three categories, served as a target variable for machine learning modelling. The process of creating the healthy ageing multi-class classification machine learning model included datasets with three different combinations of input variables. By choosing multiple sets of input variables, it was tested if using our approach in the study brings improved performance to the machine learning model compared with simply using raw data or a subset of raw data. The results indeed confirm superior performance in the machine learning results when carefully selected constructs, items or calculations relevant to healthy ageing are used.

During the study, potential limitations were noted. The data for the scale development captures information on older adults at a single time when the interview was conducted, and data includes information on self-reported health. Therefore, in the future, there is room to add a broader set of information, from the perspective of both time and content. The dataset also stores information on people aged 50 or older, termed

Information displayed on the annotation screen
Dedicated/conscious healthcare (regular exercise, exercise, sports, gardening, nutrition)
Self-assessment of physical activity (movement, recreation, regular exercise in nature, exercise, sports activity)
Self-assessment of physical health by organ systems - health of movements, balance, injuries after falls
Mental well-being (loneliness, anxiety, restlessness, sadness)
Achieving meaning and satisfaction with life
Many of my life experiences and insights are taken over by others
Participation in organizations according to the type of organization
Do you have someone to talk to about confidential, personal matters?

Table 6: Description of the information which was placed on the annotation screen.

Input features	min_child_weight	gamma	subsample	colsample_bytree	max_depth
Set 1	6	0.2	0.6	0.7	6
Set 2	6	0.2	0.7	0.6	5
Set 3	12	0.2	0.8	0.6	6

Table 9: Selected parameter values for each of the three sets of input variables.

"early old age". However, some definitions of healthy ageing define older people as people aged 60 or older [3] [14]. Therefore, our healthy ageing scale might apply to the younger generation of older adults without many chronic diseases and conditions. The explanatory factor analysis was used to develop constructs for the rating process, and only records without missing data were kept for the analysis. Further analysis would be required to investigate if groups of older adults with specific health conditions were omitted by omitting incomplete records.

The ageing population in Slovenia, where the development data comes from, is considered quite typical of the ageing population in European and developed countries [32], so results are applicable in this sense. The development data comes from a carefully designed, implemented and controlled large-scale study conducted by the Anton Trstenjak Institute of Gerontology and Intergenerational Relations in 2010 and represents a reliable source of objective data.

The proposed healthy ageing scale could also be applied in actual practice as a time-efficient method for obtaining the ground truth values of healthy ageing, where long and tedious procedures for capturing healthy ageing are not acceptable due to limitations in expert time and participant engagement. By incorporating gerontology expertise, we embraced an extensive range of aspects and integrated them into a uni-dimensional scale. It could also be used as an accompanying tool to develop intelligent home-based and artificial intelligence-based automated healthy ageing applications. In light of the shift of focus from a disease-centred to a person-centred approach [46], the proposed metrics could also be a valuable tool to provide a regular assessment of an older person's health in the scope of developed personalized health plans or healthy ageing-related activities recommendation systems, thus providing a timely trigger to react and adapt to a person's changing health.

6. Future research

Several aspects could be explored in future research. By using additional data, the accuracy of the scale could be enhanced. Such data could comprise infor-

Input features	AUC OvO	AuC OvR	F1	Precision	Recall
Set 1	0.92	0.91	0.72	0.75	0.69
Set 2	0.66	0.64	0.47	0.51	0.46
Set 3	0.73	0.71	0.52	0.59	0.50

Table 10: Evaluation of XGBboost classifier performance for each set of input variables using the macro-average of the model performance metrics AUC OvO, AUC OvR, F1, precision and recall.

mation captured via longitudinal studies and standardized tests (e.g. walking tests). Behaviour data could be captured via intelligent devices. Age and culture-specific information could be added to help identify those adults whose functions are in the upper range of physical, mental and social well-being domains [14]. In this study, the classification method was used to create a model that outputs categorical healthy ageing scores and mimics the discrete rating of a domain expert user. However, machine learning could also be used to develop a continuous model for healthy ageing score production, thus mimicking the continuous nature of the ageing process. Another exciting aspect for research would be the development of a healthy ageing scale where input data would comprise only information that could be directly influenced by older adults' change of habits or corrective actions, thus creating a scale that could serve older adults as the indicator or their efforts to improve the ageing process.

Acknowledgements

We thank Mrs Ajda Svetelšek for providing gerontology expertise and insight. We also thank all participating gerontologists for their time spent rating the healthy ageing of older adults. We are also very grateful to the Anton Trstenjak Institute of Gerontology and Intergenerational Relations for providing data for the analysis.

Data availability statement

The participants of this study did not give written consent for their data to be shared publicly, so due to the sensitive nature of the research, supporting data is unavailable.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

[1] U. N. D. of Economic, S. Affairs, World Population Ageing 2020: Highlights, United Nations, 2021. doi:https://doi.org/10.18356/9789210051934.

[2] W. Lutz, W. Sanderson, S. Scherbov, The coming acceleration of global population ageing, Nature 451 (7179) (2008) 716–719. doi:10.1038/nature06516.

[3] WHO, et al., Decade of healthy ageing: baseline report (2020) 2.

[4] U. Secretary-General, Review and appraisal of the programme of action of the international conference on population and development and its contribution to the follow-up and review of the 2030 agenda for sustainable development: report of the secretary-general.

[5] S. Weiland, T. Hickmann, M. Lederer, J. Marquardt, S. Schwindenhammer, The 2030 agenda for sustainable development: transformative change through the sustainable development goals?, Politics and Governance 9 (1) (2021) 90–95. doi:https://doi.org/10.17645/pag.v9i1.4191.

[6] S. N. I. of Public Health, Healthy ageing: a challenge for Europe, Swedish National Institute of Public Health, 2007.

[7] J.-P. Michel, R. Sadana, "healthy aging" concepts and measures, Journal of the American Medical Directors Association 18 (6) (2017) 460–464. doi:10.1016/j.jamda.2017.03.008.

[8] L. L. Bryant, K. K. Corbett, J. S. Kutner, In their own words: a model of healthy aging, Social science & medicine 53 (7) (2001) 927–941. doi:10.1016/s0277-9536(00)00392-0.

[9] W. H. Organization, World report on ageing and health, World Health Organization, 2015.

[10] W. Lu, H. Pikhart, A. Sacker, Domains and measurements of healthy aging in epidemiological studies: A revieww, The Gerontologist 59 (4) (2019) e294–e310. doi:https://doi.org/10.1093/geront/gny029.

[11] P. Martin, N. Kelly, B. Kahana, E. Kahana, B. J. Willcox, D. C. Willcox, L. W. Poon, Defining successful aging: A tangible or elusive concept?, The Gerontologist 55 (1) (2015) 14–25. doi:https://doi.org/10.1093/geront/gnu044.

[12] J. W. Rowe, R. L. Kahn, Successful aging, The gerontologist 37 (4) (1997) 433–440. doi:https://doi.org/10.1093/geront/37.4.433.

[13] J. W. Rowe, R. L. Kahn, Successful aging 2.0: Conceptual expansions for the 21st century, The Journals of Gerontology: Series B 70 (4) (2015) 593–596. doi:https://doi.org/10.1093/geronb/gbv025.

[14] N. Peel, H. Bartlett, R. McClure, Healthy ageing: how is it defined and measured?, Australasian Journal on Ageing 23 (3) (2004) 115–119. doi:https://doi.org/10.1111/j.1741-6612.2004.00035.x.

[15] J. Lara, A. Godfrey, E. Evans, B. Heaven, L. J. Brown, E. Barron, L. Rochester, T. D. Meyer, J. C. Mathers, Towards measurement of the healthy ageing phenotype in lifestyle-based intervention studies, Maturitas 76 (2) (2013) 189–199. doi:10.1016/j.maturitas.2013.07.007.

[16] R. Sadana, Development of standardized health state descriptions, Geneva: World Health Organization, 2002, Ch. 7.1, pp. 315–328.

[17] M. Shelkey, M. Wallace, Katz index of independence in activities of daily living, Home Healthcare Now 19 (5) (2001) 323–324.

[18] F. Mahoney, D. Barthel, Functional evaluation: the Barthel index, Maryland state medical journal 14 (1965) 61–65. doi:10.1016/S0140-6736(10)62108-3.

[19] A. Campagner, D. Ciucci, C.-M. Svensson, M. T. Figge, F. Cabitza, Ground truthing from multi-rater labeling with three-way decision and possibility theory, Information Sciences 545 (2021) 771–790. doi:https://doi.org/10.1016/j.ins.2020.09.049.

[20] A. J. Ratner, C. M. De Sa, S. Wu, D. Selsam, C. Ré, Data programming: Creating large training sets, quickly, Advances in neural information processing systems 29. doi:https://doi.org/10.48550/arXiv.1605.07723.

[21] A. Košir, G. Strle, M. Meža, Weak ground truth determination of continuous human-rated data, IEEE Access 9 (2020) 4594–4606. doi:10.1109/ACCESS.2020.3046293.

[22] R. E. Mayer, R. Moreno, Nine ways to reduce cognitive load in multimedia learning, Educational psychologist 38 (1) (2003) 43–52. doi:https://doi.org/10.1207/S15326985EP3801_6.

[23] R. Y. Wong, A new strategic approach to successful aging and healthy aging (2018). doi:10.3390/geriatrics3040086.

[24] F. F. Caballero, G. Soulis, W. Engchuan, A. Sánchez-Niubó, H. Arndt, J. L. Ayuso-Mateos, J. M. Haro, S. Chatterji, D. B. Panagiotakos, Advanced analytical methodologies for measuring healthy ageing and its determinants, using factor analysis and machine learning techniques: the athlos project, Scientific reports 7 (1) (2017) 1–13. doi:10.1038/srep43955.

[25] Z. Asghari Varzaneh, M. Shanbehzadeh, H. Kazemi-Arpanahi, Prediction of successful aging using ensemble machine learning algorithms, BMC Medical Informatics and Decision Making 22 (1) (2022) 258. doi:https://doi.org/10.1186/s12911-022-02001-6.

[26] M. Ahmadi, R. Nopour, Clinical decision support system for quality of life among the elderly: an approach using artificial neural network, BMC Medical Informatics and Decision Making 22 (1) (2022) 293. doi:https://doi.org/10.1186/s12911-022-02044-9.

[27] A. Gialluisi, A. Di Castelnuovo, M. B. Donati, G. De Gaetano, L. Iacoviello, M. sani Study Investigators, Machine learning approaches for the estimation of biological aging: the road ahead for population studies, Frontiers in medicine 6 (2019) 146. doi:https://doi.org/10.3389/fmed.2019.00146.

[28] S.-Y. Chien, S.-F. Chao, Y. Kang, C. Hsu, M.-H. Yu, C.-T. Ku, Understanding predictive factors of dementia for older adults: A machine learning approach for modeling dementia influencers, International Journal of Human-Computer Studies 165 (2022) 102834. doi:https://doi.org/10.1016/j.ijhcs.2022.102834.

[29] S. Adhikari, S. Thapa, U. Naseem, P. Singh, H. Huo, G. Bharathy, M. Prasad, Exploiting linguistic information from nepali transcripts for early detection of alzheimer's disease using natural language processing and machine learning techniques, International Journal of Human-Computer Studies 160 (2022) 102761. doi:10.1016/j.ijhcs.2021.102761.

[30] G. Cicirelli, R. Marani, A. Petitti, A. Milella, T. D'Orazio, Ambient assisted living: A review of technologies, methodologies and future perspectives for healthy aging of population, Sensors 21 (10) (2021) 3549. doi:https://doi.org/10.3390/s21103549.

[31] A. Thieme, D. Belgrave, G. Doherty, Machine learning in mental health: A systematic review of the hci literature to support the development of effective and implementable ml systems, ACM Transactions on Computer-Human Interaction (TOCHI) 27 (5) (2020) 1–53.

[32] J. Ramovš, Staranje v Sloveniji: raziskava o potrebah, zmožnostih in stališčih nad 50 let starih prebivalcev Slovenije, Inštitut Antona Trstenjaka, 2013.

[33] J. Ramovš, Potrebe, zmožnosti in stališča starejših ljudi v sloveniji, Kakovostna starost 14 (2) (2011) 3–21.

[34] J. C. Hayton, D. G. Allen, V. Scarpello, Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis, Organizational research methods 7 (2) (2004) 191–205. doi:https://doi.org/10.1177/1094428104263675.

[35] W. R. Zwick, W. F. Velicer, Comparison of five rules for determining the number of components to retain., Psychological bulletin 99 (3) (1986) 432. doi:https://doi.org/10.1037/0033-2909.99.3.432.

[36] W. F. Velicer, C. A. Eaton, J. L. Fava, Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components, Problems and solutions

in human assessment (2000) 41–71doi:10.1007/978-1-4615-4397-8_3.

[37] K. Krippendorff, Computing krippendorff's alpha-reliability, Retrieved from https://repository.upenn.edu/asc_papers/43.

[38] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794. doi:https://doi.org/10.48550/arXiv.1603.02754.

[39] C. Bentéjac, A. Csörgő, G. Martínez-Muñoz, A comparative analysis of gradient boosting algorithms, Artificial Intelligence Review 54 (3) (2021) 1937–1967. doi:https://doi.org/10.1007/s10462-020-09896-5.

[40] H. Dalianis, H. Dalianis, Evaluation metrics and evaluation, Clinical text mining: secondary use of electronic patient records (2018) 45–53doi:https://doi.org/10.1007/978-3-319-78503-5_6.

[41] R. Tsopra, X. Fernandez, C. Luchinat, L. Alberghina, H. Lehrach, M. Vanoni, F. Dreher, O. U. Sezerman, M. Cuggia, M. de Tayrac, et al., A framework for validating ai in precision medicine: considerations from the european itfoc consortium, BMC Medical Informatics and Decision Making 21 (1) (2021) 1–14. doi:https://doi.org/10.1186/s12911-021-01634-3.

[42] M. Galar, A. Fernández, E. Barrenechea, H. Bustince, F. Herrera, An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes, Pattern Recognition 44 (8) (2011) 1761–1776. doi:https://doi.org/10.1016/j.patcog.2011.01.017.

[43] F. Liu, P. Zhou, S. J. Baccei, M. J. Masciocchi, N. Amornsiripanitch, C. I. Kiefe, M. P. Rosen, Qualifying certainty in radiology reports through deep learning–based natural language processing, American Journal of Neuroradiology 42 (10) (2021) 1755–1761. doi:10.3174/ajnr.A7241.

[44] E. Cho, Making reliability reliable: A systematic approach to reliability coefficients, Organizational Research Methods 19. doi:10.1177/1094428116656239.

[45] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: synthetic minority oversampling technique, Journal of artificial intelligence research 16 (2002) 321–357. doi:https://doi.org/10.1613/jair.953.

[46] M. Cesari, Y. Sumi, Z. A. Han, M. Perracini, H. Jang, A. Briggs, J. A. Thiyagarajan, R. Sadana, A. Banerjee, Implementing care for healthy ageing, BMJ Global Health 7 (2) (2022) e007778. doi:10.1136/bmjgh-2021-007778.