Gašperlin Stepančič *et al.*
BMC Medical Informatics and Decision Making (2024) 24:317
https://doi.org/10.1186/s12911-024-02714-w

BMC Medical Informatics and Decision Making

RESEARCH

Open Access

A novel explainable machine learning-based healthy ageing scale

Katarina Gašperlin Stepančič[1], Ana Ramovš[2†], Jože Ramovš[2†] and Andrej Košir[3*†]

Abstract

Background Ageing is one of the most important challenges in our society. Evaluating how one is ageing is important in many aspects, from giving personalized recommendations to providing insight for long-term care eligibility. Machine learning can be utilized for that purpose, however, user reservations towards "black-box" predictions call for increased transparency and explainability of results. This study aimed to explore the potential of developing a machine learning-based healthy ageing scale that provides explainable results that could be trusted and understood by informal carers.

Methods In this study, we used data from 696 older adults collected via personal field interviews as part of independent research. Explanatory factor analysis was used to find candidate healthy ageing aspects. For visualization of key aspects, a web annotation application was developed. Key aspects were selected by gerontologists who later used web annotation applications to evaluate healthy ageing for each older adult on a Likert scale. Logistic Regression, Decision Tree Classifier, Random Forest, KNN, SVM and XGBoost were used for multi-classification machine learning. AUC OvO, AUC OvR, F1, Precision and Recall were used for evaluation. Finally, SHAP was applied to best model predictions to make them explainable.

Results The experimental results show that human annotations of healthy ageing could be modelled using machine learning where among several algorithms XGBoost showed superior performance. The use of XGBoost resulted in 0.92 macro-averaged AuC OvO and 0.76 macro-averaged F1. SHAP was applied to generate local explanations for predictions and shows how each feature is influencing the prediction.

Conclusion The resulting explainable predictions make a step toward practical scale implementation into decision support systems. The development of such a decision support system that would incorporate an explainable model could reduce user reluctance towards the utilization of AI in healthcare and provide explainable and trusted insights to informal carers or healthcare providers as a basis to shape tangible actions for improving ageing. Furthermore, the cooperation with gerontology specialists throughout the process also indicates expert knowledge as integrated into the model.

Keywords Healthy ageing, Older adults, Novel scale, Machine learning, Factor analysis, Expert ratings, Explainability

[†]Ana Ramovš, Jože Ramovš and Andrej Košir contributed equally to this work.

*Correspondence:
Andrej Košir
andrej.kosir@fe.uni-lj.si
Full list of author information is available at the end of the article

Background

The world continues to experience a change in the population's age structure [1]. People are living longer lives causing the share of older people in the total population to increase rapidly and this trend will likely continue [2]. While in 1980 the global population aged 60 years and over was 382 million, that number was already over 1 billion people in 2020 and is projected to reach nearly 2.1 billion by 2050 [3]. 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 recently received considerable attention from governments, organizations and other stakeholders. World Health Organization (WHO) also declared 2021-2030 a decade of healthy ageing [3].

Healthy ageing definitions vary. Among others it has been described as the ability to go and do a meaningful activity [6]; as a general condition of the ageing of a person's mind and body, usually meaning freedom from illness, injury, or pain [7]; and 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 [8]. According to a review of healthy ageing definitions and measures [9], a comprehensive health outcome should measure how well a human can function in domains assessing physical, mental and social well-being. Healthy ageing is also used interchangeably with terms such as "active", "successful", or "productive" ageing [10].

The evaluation of how a person is ageing and the derivation of potential actions for ageing course improvements is important in many aspects. Healthy ageing leads to an improved quality of life, decreased health care consumption, and contributes to the labour supply, decreasing the likelihood of early retirement [11]. It could also be important in determining long-term care eligibility. As ageing is a complex process that depends on many factors, no unified measure of healthy ageing exists. The efforts to assess the health of older adults are mostly using items drawn from 4 categories [12]: (i) fulfilling or performing functions, activities, or roles (basic activities of daily living, instrumental activities of daily living, advanced activities of daily living); (ii) items reflecting the WHO definition of health and well-being (describing physical, social and mental aspects of health); (iii) symptom-oriented; and (iv) those concerned with adaptation or coping with non-fatal health conditions or limitations.

Recently, machine learning has been widely used in research focusing on older adults and has been highlighted as a helpful enabler for the more holistic and interdisciplinary approach towards healthy ageing evaluation [13]. Multiple research reports on the topic can be found in the literature. Caballero et al. [14] created the unidimensional multi-class metric of healthy ageing comprised of 45 items on self-reported health, utilizing factor analysis and Bayesian multilevel Item Response Theory. Asghari et al. [15] used six machine learning algorithms, including ensemble, to develop a binary class model for successful ageing, where features were defined based on Rowe and Kahn's theory. Yazdani et al. [16] uses the adaptive network-based fuzzy system for the prediction of successful ageing while [17] 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 in other areas such as estimating the biological age of the organism [18], predicting specific age-related conditions such as dementia [19] and Alzheimer's disease [20], and developing ambient-assisted living systems [21].

Increased use of machine learning also brings recommendations for further research. Specifically, the study of machine learning use in the mental health domain [22] suggests that for more implementable machine learning systems, more research would be needed to (i) test the validity of the developed constructs and (ii) ensure the robustness of the outputs for practical use (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.

As artificial intelligence-based (AI-based) systems are becoming increasingly important for decision-making in organizations, another topic on the table is their black-box nature which is limiting their use to its full potential. Explainability, besides the early involvement of domain experts, is, therefore, the crucial element for establishing transparency and trust in machine learning model results as it enables communicating the reasons for decisions to target users and stakeholders and improving human/AI collaboration. It is one of the frequently debated topics in highly-regulated industries such as healthcare [23], finance [24], insurance [25] and public services [26]. In healthcare, the lack of explainability can cause hesitation by medical professionals to use these models in real-world scenarios as the high accuracy of machine learning is insufficient and a single number does not provide the information on how the result has arrived. The reasons behind model predictions should be known so clinicians can make informed decisions about treatment and care [27]. Several approaches exist in the field of explainable AI for healthcare. Two popular approaches used are Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP). In the field of Alzheimer's disease prediction, nearly 70 % of studies are utilizing these two techniques [27]. Among others, they are also used for explanations for diabetes prediction [28], deep-learning-based medical imaging applications [29] and retinoblastoma diagnosis [30].

To some extent, the use of explainable AI methods can also already be found in specialized applications within the ageing domain such as predicting brain age based on morphological features [31], fall predictions for older adults [32] and prediction of comorbidity [33]. To the best of our knowledge, there is no machine-learning-based healthy ageing scale as of today that would on one side involve process-wide active cooperation with gerontology experts to capture their critical insights and build trust and on the other side utilize explainability frameworks to provide reasons for model decisions.

In this paper, we present a novel machine learning-based healthy ageing scale which provides explainable results and would be easy to understand by domain experts. The study comprised several stages designed and completed in close cooperation with gerontology experts which leads towards closer integration and inclusion of gerontology knowledge in the scale itself as well as the use of the SHAP interpretation technique [34] for explaining individual machine learning predictions.

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) [35] or using data programming frameworks [36]. In our study, we asked multiple gerontology experts to collaborate on the design of healthy ageing constructs as well as to provide annotations. We consider that an important differentiation from other studies. To obtain a healthy ageing scale a selection of relevant healthy categories and individual variables was chosen from data on adults aged 50+, gathered via in-person field interviews in the independent study. Explanatory factor analysis (EFA) was applied to our data to find and select relevant constructs that describe healthy ageing. As opposed to other studies we conducted explanatory factor analysis individually for every identified health category and gerontology domain experts then selected the most relevant healthy ageing aspects out of derived factors. A web annotation application was designed and developed as a basis for visualising those aspects for each older adult from the study and was successfully used for capturing annotations from gerontology experts. The design principles of software applications for annotation purposes as well as the amount and presentation of content are influenced by the cognitive load theory (CLT) as well as human-computer interaction (HCI) principles [37], which both share basic assumptions of the human cognitive system and a need to reduce irrelevant load. Both aspects were taken into account while developing the application. The ground truth obtained from the annotations was calculated and reliability and inter-rater agreement were assessed [38]. Ground truth was than used as a target in a machine learning process within which we tested six different algorithms with extreme gradient boosting (XGBoost) being selected as the best performer. To reduce reservations towards black-box model predictions at the end we applied SHAP interpretation techniques for explaining individual predictions. During the research process, we also addressed the question if a healthy ageing scale be developed based on combining multivariate statistics and domain expert annotations concerning the validity and reliability of psychometric properties. The described approach increases the scale potential and robustness to be used in practice in healthy ageing-related applications, such as context-aware explainable recommendation systems and clinical decision systems, where long and tedious evaluation procedures are not acceptable in terms of domain experts' time and participant engagement.

Methods

Dataset

The dataset used in this research was obtained by Anton Trstenjak Institute of Gerontology and Intergenerational Relations (further referred also as the institute), a Slovenian national scientific, research, expert, and end-user institution within the gerontology and good intergenerational relations field in Slovenia. Data collection was part of a separate, independent study and the research presented in this paper uses the resulting data collected there. For the purposes of that study, the institute developed an extensive questionnaire on ageing that was used for conducting in-person interviews. Results of this study are published in "Ageing in Slovenia: Survey on the Needs, Abilities and Standpoints of the Slovene Population Aged 50 Years and Over" [39]. The questionnaire used during the interviews is, however, not publicly available. The National Medical Ethics Committee of the Republic of Slovenia considered the questionnaire as well as the research concept of the source research on this data about the ageing in Slovenia [39], and an opinion was issued that the research was ethically impeccable. Ethical consent (nr. 115/09/09) was issued for its implementation [40] and informed consent was obtained from all participants included in the study.

Those not providing the consent, were not interviewed. The research reported here in this paper was completely aligned with the aims of the data was collected for and no additional ethics-related issues were opened. During the interview process, special methodological attention was paid to the respondent's motivation for the selected sample. The training and monitoring of interviewers and data entry into the database was conducted as well.

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 [39].

The targeted population for this paper's proposed metrics is people aged 50+ with demographic characteristics that meet the dataset characteristics in terms of age, sex, and education.

Geronthology experts experience overview

In this research, we closely cooperated with gerontology experts from Anton Trstenjak Institute and Intergenerational Relations.

The gerontology expert profiles are associate professor, Doctor of Philosophy (PhD) in the field of anthropology, and a social worker. He specialized in Frankl logotherapy (European Diploma in Psychotherapy) and partner communication. He has 35+ experience in the domain with research and pedagogical focus on co-existence in solidarity; communication among young, middle and third generation; personal preparation for quality ageing and preparation of the society for a large proportion of the older population; addictions and intoxication. In theory, he focuses on the holistic image of man in his physical, mental, spiritual, social, developmental and living dimensions. He develops programs for quality life and coexistence between people based on everyday resources (anthropohygiene). His bibliography includes over a thousand items (scientific, professional and popular books, articles, contributions at congresses, radio, television and online, mentoring for diplomas, masters and doctorates); a medical doctor with a research focus in the fields of healthy ageing, preventive medicine, public health, geriatrics, ethics, telemedicine and telecare). Participates in the coordination of national and international projects related to health aspects of gerontology and long-term care; psychologist and a professional worker in the field of social welfare whose main work fields are social programs development, gerontechnology and data processing. The focus of her research are quality ageing, encompassing positive psychology and health psychology.

The healthy ageing scale development process

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 based on their experience and put into identified health categories and sub-categories.

The development of the healthy ageing scale comprised several steps which are summarized in Table 1.

Table 1 Summary of the healthy ageing scale development steps

Step number	Step description
Step 1	Selection of health categories and where applicable, subcategories.[a]
Step 2	Selection of dataset items that fall under each health category/subcategory.[a,b]
Step 3	Explanatory factor analysis performed for each health category/subcategory.[b]
Step 4	Selection of factors and items most relevant for describing healthy ageing of a person.[a]
Step 5	Design and development of web annotation application.[a]
Step 6	Annotation of healthy ageing for each person from this study.[c]
Step 7	Calculation of ground truth from healthy ageing annotation results.
Step 8	Machine learning model development using 6 different classification algorithms.
Step 9	Selection of best-performing model.
Step 10	Application of SHAP interpretation technique to individual model predictions.

[a] Tight cooperation with gerontology domain experts

[b] If the health category had one or multiple subcategories, the step was completed for each subcategory

[c] Annotations were performed by gerontology domain experts

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 in the category/subcategory [41]. Explanatory factor analysis was performed for each category or, instead, sub-category if the category had one. Once the correlation matrix was constructed, principal component analysis was performed to extract factors. Determining the number of factors to extract is an important decision in exploratory factor analysis. For determining the number of components to retain multiple methods are available (Horn's parallel analysis, Velicer's minimum average partial [MAP], Cattell's scree test, Bartlett's chi-square test, and Kaiser's eigenvalue greater than 1.0 rule). According to multiple studies, Horn's parallel analysis and Velicer's minimum average partial have consistently emerged as best performance options [42, 43] and for this study, we decided to use parallel analysis. Explanatory factor analysis was done using standard R packages corrplot and psych. The obtained factor matrices were used for a detailed discussion with gerontology domain experts to find and define relevant constructs and items which should be part of the context for evaluating a person's healthy ageing.

Annotation of how well the person is ageing

A custom web annotation application was developed to capture gerontology expertise in defining a healthy ageing scale. The web application was developed using the Django framework [44] 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. 1. It visualizes information about eight healthy ageing constructs of a person: one's conscious care for health, one's self-assessment of physical activity, one's self-assessment of body health according to organ systems, mental well-being, achieving meaning and life satisfaction, perception of how one's own life experiences are summarized by others, one's participation in publicly renowned and socially visible organizations, and information if a person has someone with whom it can talk about private and personal topics. Graphs contain mean values for each construct (vertical black line) and coloured intervals of three (light blue) and five (light grey) standard deviations to identify outliers and extreme

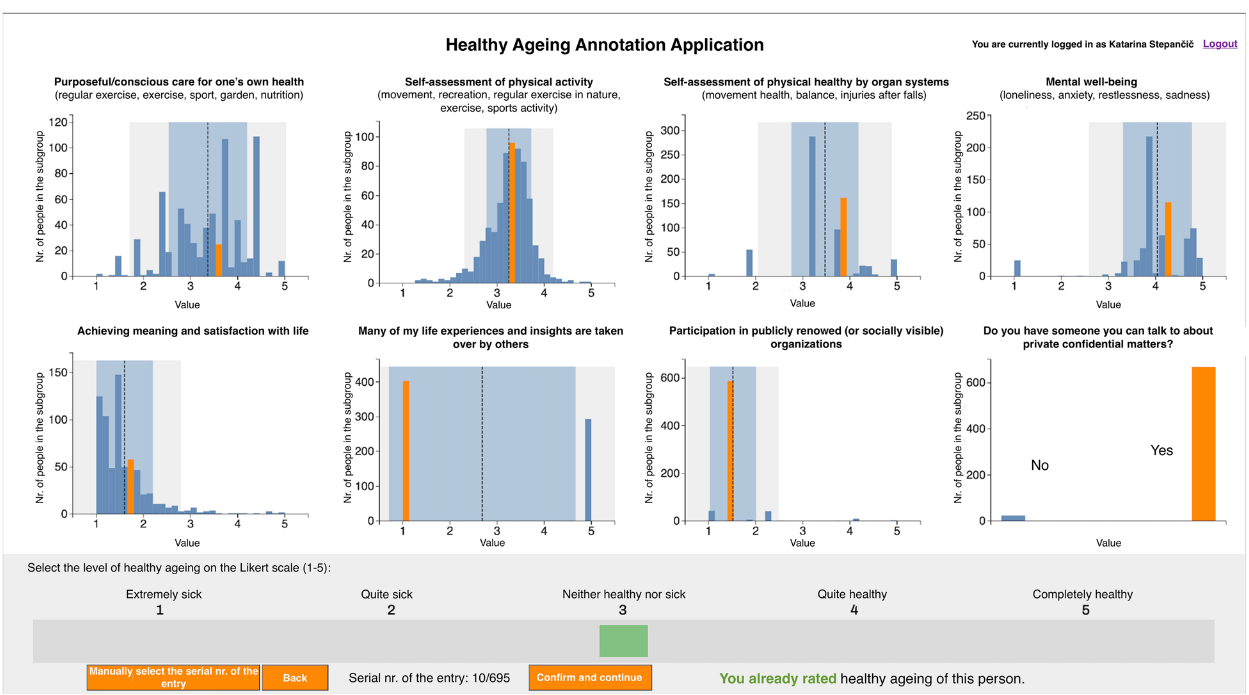

Fig. 1 The web application annotation screen for healthy ageing annotation procedure

values. Values for all participants are shown and the value of a person being rated is highlighted in orange. A Likert scale from 1 (extremely sick) to 5 (completely healthy) was used by raters to determine the level of healthy ageing for each participant in the study.

Before the annotation application was developed the discussion and specification of annotated items was carried out with the gerontology experts. Furthermore, 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 before the rating. Randomization was used during the rating process so that each rater who annotated older adults had its own order of cases. The reason for using randomization was to eliminate cross-annotated elderly effects. In the beginning, 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. The web annotation application allowed raters to return to previous, already-rated cases and rate them again. In case when multiple ratings were provided for the same user, the latest rating counted. A training session as well as a web annotation application usage guide were prepared for raters before the rating. Four raters participated in the annotation process.

Ground truth for a healthy ageing scale

As a result of the annotation process, four healthy ageing ratings were obtained for each older 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 [38]. Annotator bias removal procedure from [38] was applied. The inter-rater agreement was also measured using Krippendorff's alpha [45], 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.

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/her needs, abilities and attitudes 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. Six machine learning algorithms were used during the process to select the best-performing classifier for modelling healthy ageing on the available data. Those were logistic regression, decision tree classifier, random forest, k-nearest neighbours (KNN), support vector machines (SVM) and extreme gradient boosting (XGBoost). The grid search procedure was used to find an optimal combination of hyperparameters for each classifier and stratified 10-fold cross-validation was used [46] to split data into train and test sets. A stratified k-fold was used to preserve the percentage of samples for each class in the target variables.

Data

In total 11 input variables were used for machine learning comprising data used on the annotation screen (5 factors, 2 individual items and 1 calculation) with the addition of age, education and gender. Two other sets of input features were considered for machine learning. First option was the usage of all 82 initial items available in the dataset, without any data preprocessing. The second option was to select only items that influenced the 5 factors placed on the annotation screen together with the remaining 2 individual features and 1 calculation from the screen. However, the initial performance of the machine learning utilizing raw data was not satisfactory. When using XGBoost we got an area under the curve one-versus-one (AUC OvO) Macro of 0.73 and F1 Macro of 0.53 for the first option; AUC OvO Macro of 0.67 and F1 Macro of 0.65 for the second option. Therefore additional tests were not performed and reported in this study. This also indicates the importance of the data preparation process (in our case dimensionality reduction using EFA) for obtaining quality machine learning results which in our case was deeply connected with domain experts' involvement.

The target variable used for the machine learning process was the ground truth value. The ground truth value was obtained by calculating the weighted truncated mean of the four ratings gathered from gerontology domain experts via the annotation procedure. More details on the procedure are available in "Ground truth for a healthy ageing scale" section.

Overview of best-performing 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 provides a reliable and efficient implementation of the gradient boosting algorithm and is often used as the component in many winning solutions in machine learning competitions [47].

XGBoost is a decision tree ensemble machine learning algorithm based on gradient boosting and is designed to be highly scalable [48]. 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. To control the overfitting, the regularized objective (minimization) function L consists of two parts.

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

where

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

$l(y_i, F(x_i))$ 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 XGBoost 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.

Evaluation metrics

Model performance was evaluated using the standard metrics: accuracy, the area under the receiver operating characteristic curve (AUC) evaluation metric [49], 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-versus-one (OvO) and One-versus-rest (OvR, also referred to as One-versus-all or OvA) strategies were used when calculating the area under the curve to select the best strategy [50]. 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 [51]. 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 [52].

Explainability

Two popular approaches used for machine learning model explainability are LIME and SHAP. Some other techniques applied in healthcare are partial dependence plots, individual conditional explanation, accumulated local effects and permutation feature importance [53].

LIME is a technique that offers localized interpretability (explaining a single prediction) by generating a new dataset using perturbed samples from the surrounding region and creating accompanying predictions using the black-box model. It then fits a new, interpretable model (e.g. a linear model) on this new set of data, measured by the distance between the sampled occurrences and the instance of interest. SHAP is a game theoretic approach that provides global and local interpretability insights where the weight is assigned to each feature to measure its contribution to the prediction. Both framework approaches are open-source, and model-agnostic and can be used for classification and regression.

In this paper, the decision to apply SHAP was made due to several advantages over LIME as reported in explainable AI-related work. The comparison in [27, 28] states that the advantages of SHAP over LIME are stability and consistency; fair distribution of contribution for each of the variables, ensured by Shapley value; options for entire model explanation and not only local explanations; no challenges with explanations for more complex models; no assumptions about the model linearity; ability to generate contrastive explanations and better visualization options. It also states that due to its theoretical guarantees and simplicity, SHAP is more widely used. On the other side, LIME is faster and simpler to use and has more stability on the traits with high relevance scores. LIME is more stable for top-ranked features while SHAP is more stable when the majority of features are present. Additionally, LIME requires less computing time. However, for tree-based models (which we also have in our use case), SHAP offers a fast implementation option that proved crucial for its acceptance [53].

Results

Selection of participants for the study

The Anton Trstenjak Institute of Gerontology and Intergenerational Relations dataset captures information about 1047 adults aged 50+. Before further analysis rows with missing values were dropped which resulted in a subsample of 696 participants. At the same time, the characteristics of the subsample population in terms of demographics (age, sex, and education) were preserved. Figure 2 compares age histograms across all participants in the dataset and a subsample 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 (in initial dataset 41.3 % are women and 58.7 % are men). The mean value of education level in a subsample is 3.17 (in the initial dataset is 3.13).

Selection of health categories and sub-categories

The most important categories and their sub-categories that define healthy ageing were selected. The categories were selected based on gerontology domain experts' experience and the findings they performed during their study of the independent study survey results. A summary of selected sub-categories and their descriptions are provided in Table 2. Selected domains match those mentioned as common among the healthy ageing studies review: physical, social and mental [9].

Each category and accompanying sub-category consisted of and was defined by several items from the original dataset. A total of 82 items were chosen from the original dataset. All items and accompanying answer choices together with categories and sub-categories to which they belong and were the input into explanatory factor analysis are provided in additional file (see Additional file 1).

Explanatory factor analysis results

Explanatory factor analysis was performed for each category or sub-category, depending on whether the category had sub-categories. The principal component analysis method was used in explanatory factor analysis, and multiple combinations of factoring methods (weighted least squares (WLS), minimum residual (Minres)) and

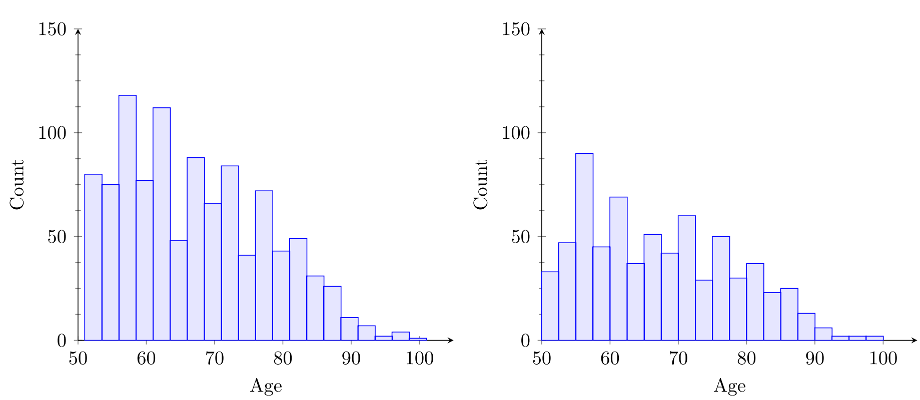

Fig. 2 Age histogram across all participants (left) and a subset used in our study (right)

Table 2 Short descriptions of categories and accompanying sub-categories that were chosen by gerontology experts

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

rotations (no rotation, Varimax, Quartimax, Promax) were tested. Results were discussed with gerontology experts who provided feedback on factor interpretations. Five factors resulting from explanatory factor analysis were selected as relevant for inclusion in the web annotation application. Along with five factors two individual items from the dataset (individual questions) and one value calculated from multiple items were selected for the web annotation application as well. A summary of the selected information, along with the information type, is summarized in Table 3. 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 Table 4.

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 Table 5, 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.

Psychometric characteristics: validity and reliability

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.

To assess the reliability of the proposed models [54] and 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. 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 proportion of variability extracted by the model), and reliability coefficient Cronbach's alpha. The psychometrics characteristics of five explanatory factor analyses are given in Table 6. 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. It is also a lower bound of the adequate reliability coefficient.

Table 3 Factorisability, factorisation method and rotation method applied for each of the selected constructs

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 4 Items summary with corresponding factor loadings for constructs selected for the healthy ageing scale

Construct	Item	Loading
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 5 Factorisability, factorisation method and rotation method applied for each of the selected constructs

Construct	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 6 Psychometric characteristics of five explanatory factor analyses applied

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

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 Table 7. 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. Additionally, all information units descriptions were given and coordinated with them as well.

Information was presented in a graphical way using histograms and distribution graphs with descriptions as presented in Fig. 1. Each histogram visualized the distribution of values for all the people being annotated and highlighted the bar (orange) where the value for the person currently annotated is located. 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 - extremely sick, 2 - quite sick, 3 - neither sick nor healthy, 4 - quite healthy, and 5 - completely 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 "Annotation of how well the person is ageing" section. Krippendorff's alpha that measures inter-rater agreement was 0.59. As estimated agreements of annotators were satisfactory, that showed their interpretation of the data was similar and therefore no post-interviews and results interpretation was carried out after the annotation procedure was completed.

Healthy ageing scale machine learning model

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 "Ground truth for a healthy ageing scale" section. The obtained ground truth was the categorical variable with values ranging from 1.5 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. Originally, the plan was to reclassify those values back to 5 categories. However, the bottom (extremely sick) and top classes (completely healthy) were represented with a small number of instances, 5 and 20 respectively, which would limit the success of a machine learning effort. Therefore decision was made to reclassify the original values into three more meaningful and representative categories representing poor, moderate, and good healthy ageing categories. The resulting proportions of the target variable's poor, moderate, and good healthy ageing categories are shown in Table 8. Due to an unbalanced dataset, the synthetic minority oversampling technique (SMOTE) [55] was applied to the training dataset to make the ratio of classes in the dataset equal. The test dataset used to evaluate the classifiers' performance consisted of real samples only. SMOTE is a method in which the minority class is over-sampled by

Table 7 Description of the information which was placed on the annotation screen

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 8 Target variable class representation

Class	%
Poor healthy ageing	0.174
Moderate healthy ageing	0.566
Good healthy ageing	0.260

creating synthetic data points that are moderately different from the original.

Machine learning configuration settings

Six machine learning algorithms were applied in the machine learning process with the purpose to identify the best performing model for the given dataset. Those were logistic regression, decision tree classifier, random forest, k-nearest neighbors, support vector machines and XGBoost. Grid search procedure was used to determine the most optimal hyperparameter values for each training procedure and model was refitted with the selected hyperparameters values. Tested hpyerparameters, ranges and final selected values are summarized in Table 9.

Evaluation of the machine learning models

Each of the models was evaluated using the accuracy, macro-averaged area under the curve one-versus-one strategy (AUC OvO), area under the curve one-versus-rest strategy (AUC OvR), F1, precision and recall. Performance results for all three models built are presented in Table 10. The best performing algorithm in terms of macro-averaged F1 and AUC OvO was XGBoost. XGBoost learns the target function additively which means that during the process it creates an ensemble of weak learners (decision trees) that in the iterative way minimizes the objective function. A new tree is added in each iteration, and the objective function is optimized. The learning objective selected for the training was multi:softprob which as a result, returns the predicted probability of each data point belonging to each class.

Explainability of XGBoost results

For the interpretation of why XGBoost makes a certain prediction, SHAP [34], a framework for interpreting predictions, was used. As XGBoost is not interpretable by itself, having the tools to help understand why a model makes a certain prediction is crucial for results to be useful in practice and applications. SHAP assumes each feature represents a "contributor" to the predictions of a model [56] and assigns each feature a SHAP value. SHAP value quantifies each feature's contribution to the prediction. SHAP provides global and local interpretation methods based on aggregations of Shapley values. It can be applied to any machine learning model as a post

Table 9 A summary of algorithms and accompanying hyperparameters with ranges tested within grid search procedure

Algorithm	Parameter	Range	Value
Logistic Regression	Penalty	['l1', 'l2']	'l1'
	C	[1.0, 0.5, 0.1]	1.0
	Solver	['liblinear']	'liblinear'
Decision Tree Classifier	Criterion	['giny', 'entropy']	'entropy'
	min_samples_leaf	[1, 2, 3, 4, 5, 6]	4
	max_depth	[1, 2, 3, 4, 5, 6]	6
	min_samples_split	[2, 3, 4, 5, 6]	2
Random Forest	min_samples_leaf	[1, 2, 3, 4, 5, 6]	1
	max_depth	[1, 2, 3, 4, 5, 6]	6
	min_samples_split	[2, 3, 4, 5, 6]	5
K-Nearest Neighbours	n_neighbors	[1, 2, 3, 4, 5, 6]	1
	weights	['uniform', 'distance']	'uniform'
	metric	['euclidean', 'manhattan']	'manhattan'
SVM	kernel	['linear', 'rbf']	'rbf'
	C	[1, 2, 3, 4, 5, 6]	6
XGBoost	learning rate	[0.1, 0.2, 0.3]	0.3
	max_depth	[4, 5, 6]	4
	min_child_weight	[1, 2, 3, 4]	1
	subsample	[1.0, 0.5, 0.1]	0.5
	n_estimators	[50, 100, 150]	100
	gamma	[0.1, 0.2, 0.3, 0.4, 0.5]	0.2
	colsample_bytree	[0.6, 0.7, 0.8, 0.9, 1]	1

Table 10 Evaluation of classifiers performance

Performance metric	Logistic regression	Decision tree	Random forest	KNN	SVM	XGBoost
Accuracy	0.72	0.63	0.77	0.57	0.75	0.78
AUC OvO (Macro)	0.92	0.81	0.92	0.64	0.90	0.92
AUC OvR (Macro)	0.90	0.79	0.91	0.63	0.88	0.91
F1 (Macro)	0.72	0.61	0.75	0.53	0.73	0.76
Precision (Macro)	0.70	0.61	0.77	0.53	0.72	0.76
Recall (Macro)	0.79	0.63	0.73	0.53	0.73	0.75

hoc interpretation technique, is agnostic towards the algorithm itself and is particularly efficient in providing explainability for algorithms such as random forests and gradient-boosted trees [57]. For a better presentation effect, SHAP offers many options for visualization of XGBoost predictions. A global feature importance plot takes each feature's mean absolute SHAP value over all the given samples to demonstrate the magnitude of feature importance. In multiclass classification, as shown in Fig. 3, such a plot is given for each class separately. In the case of multiclass classification (our XGBoost objective function was multi:softprob) the SHAP values are given in log odds that can make SHAP plots interpretation a bit more difficult. However, log odds values can be converted to probability values and for easier interpretation, Table 11 shows converted values of mean absolute SHAP from log odds to probabilities.

SHAP can also explain individual instances. It is important to note that while SHAP values tell us how each model feature has contributed to a prediction, they can not be used for causal inference. A waterfall plot was selected to display explanations for individual predictions in Fig. 4.

From top to bottom, this figure visualizes how and to what extent each feature positively (red colour) or negatively (blue colour) influenced each of the potential classes: poor, moderate or good ageing. The predicted class by the model for the presented sample was that this person has moderate ageing (highest SHAP value for f(x)). The bottom of each subplot starts as the expected value of the model output and each row above shows how the positive (red colour) or negative (blue colour) contribution of each feature moves the value from the expected model output to the model

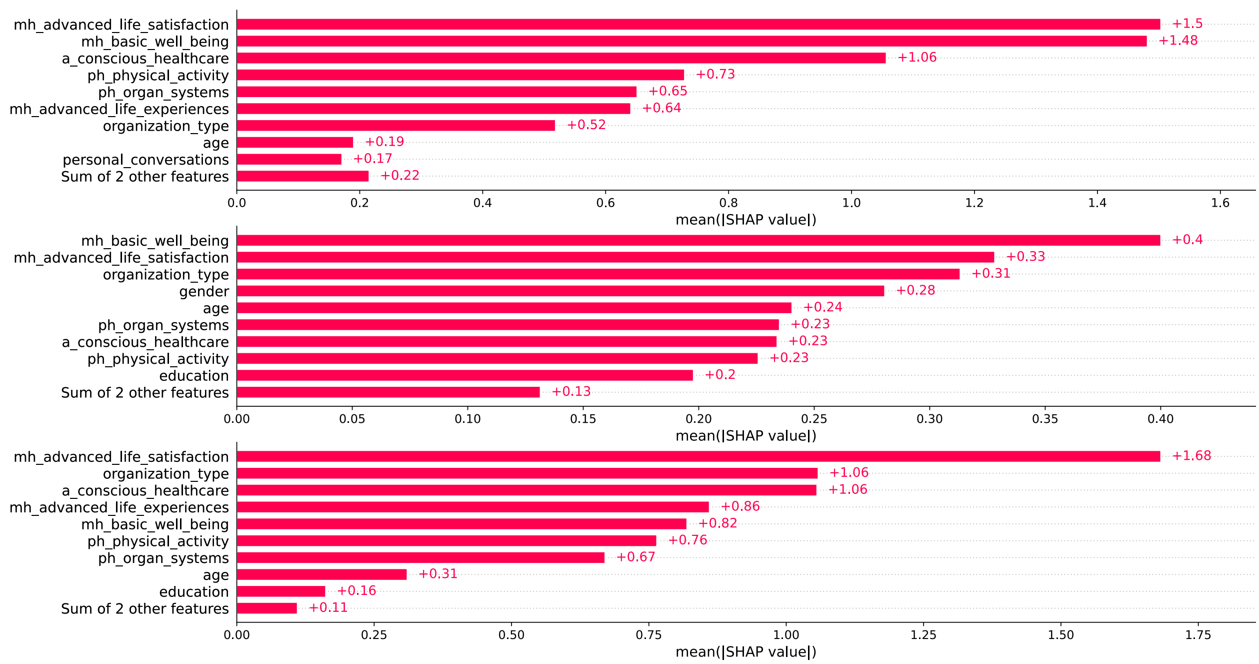

Fig. 3 The global feature importance plot. From top to bottom: poor ageing, moderate ageing, good ageing

Table 11 Mean absolute SHAP values converted to probabilities for all three classes in the test dataset

Feature name	Poor ageing	Moderate ageing	Good ageing
a_conscious_healthcare	0.12	0.09	0.11
mh_basic_well_being	0.18	0.11	0.09
mh_advanced_life_satisfaction	0.19	0.10	0.22
mh_advanced_life_experiences	0.08	0.08	0.10
ph_physical_activity	0.09	0.09	0.09
ph_organ_system	0.08	0.09	0.08
personal_conversations	0.05	0.07	0.04
organization_type	0.07	0.10	0.12
education	0.05	0.09	0.05
gender	0.05	0.09	0.04
age	0.05	0.09	0.06

output for this prediction. The ordinal axis displays all features and their accompanying values. The horizontal displays SHAP values for each feature given as log odds. For example, the f(x) = 1.371 can be converted using the softmax function to the probability of 0.48 that this person is ageing moderately.

Beeswarm plot for the predicted class is given in Fig. 5 to illustrate how features influence all test samples for the predicted class in magnitude and direction.

Discussion

This paper presents the novel domain-specific healthy ageing scale with an emphasis on embedding the elements in the design that could significantly increase the scale trust and understanding that are required by end-users to accept and use the scale. The first such element is the active involvement of gerontology domain experts throughout the whole process, which also provides validity to the overall scale development approach. Gerontology experts were present at stages of identifying the relevant healthy ageing domains, healthy ageing constructs creation, annotation application design and providing the annotation scores. Once the annotations were used for a machine learning-based scale development, the second unique element was the application of the SHAP explainability framework to the healthy ageing model predictions. This brings information on how predictors are influencing the model decision and in which direction.

The data used to develop the scale comprises five healthy ageing domains that gerontology experts selected as necessary. These domains were physical health, social health, mental health, physical activities and independent living. This is aligned with the previous research, which also utilises self-assessment health data on physical, functional, mental and social domains [14, 16, 58]. Some studies additionally use results of measured tests such as tests for measuring cognitive functions or physical abilities.

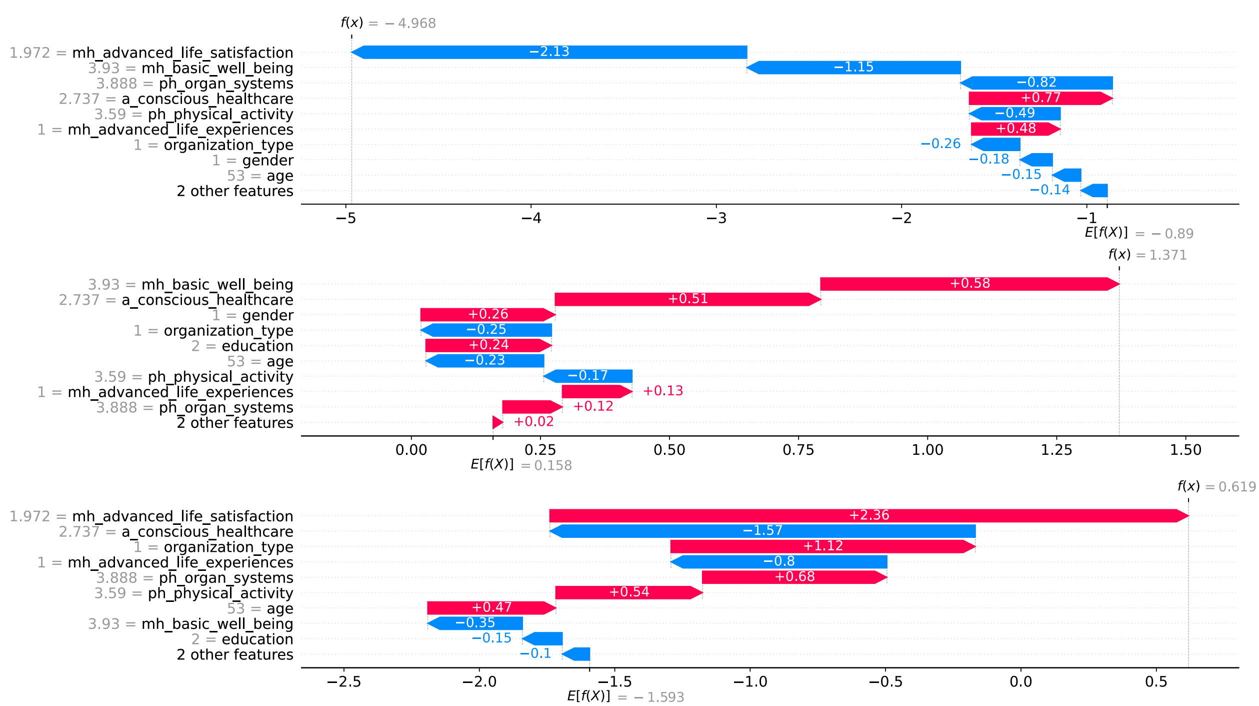

Fig. 4 The waterfall plots for each ageing class of a selected test example

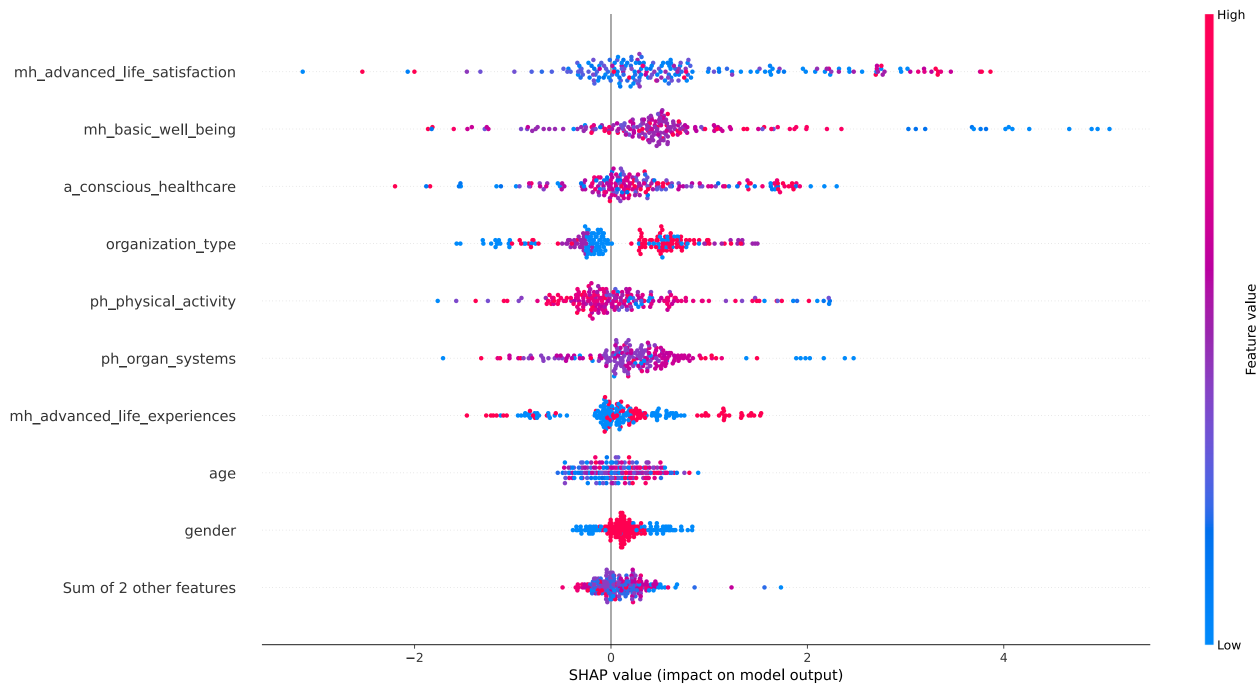

Fig. 5 Influence of features to the predicted class in magnitude and direction for all test instances

The ageing population in Slovenia, where the development data comes from, is considered quite typical of the ageing population in European and developed countries [39], 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 data. Also, consistent with the previous literature is that the similar two-phase approach first employing explanatory factor analysis/principal components analysis for dimensionality reduction and second using ML to predict healthy ageing has also been used in multiple studies [14, 16]. However, the approach to obtaining a unidimensional healthy ageing metric (target variable) differs from study to study. While [16] used a dataset with existing binary target feature that indicated if a person is ageing successfully or not, the [14] used Bayesian multilevel IRT approach to create a healthy ageing metric from 0 to 100 which was further categorized into 4 groups. On the other hand, this study used multiple expert human annotators to determine the healthy ageing of older adults and the resulting ground truth value was categorized into 3 groups. Additionally, this study utilizes a different approach in the dimensionality reduction phase. While other studies applied dimensionality reduction techniques directly on the full set of items, this study first divided items into health domains and applied EFA separately on each. We also had a richer set of initial items than other studies did: 82 items as opposed to 45 items [14] and 28 items [16]. EFA was used on health domains to find relevant constructs for visualization in the web annotation application used for the healthy ageing rating. The psychometric properties were also assessed during the study to address whether a healthy ageing scale can be developed based on combining multivariate statistics and domain expert annotations. The unidimensional, congeneric measurement model was used to assess the reliability of the constructs, and Chi-square tests were applied. Selected information was placed on the annotation application where the design of the application itself was confirmed through a discussion with gerontology experts. The application visually compared the data for each older adult participating in the study 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. Randomization and initialization processes were implemented to eliminate cross-annotated elderly effects and prevent raters from calibrating their annotations based on the first annotations. The ground truth procedure was applied to get the single value per older adult from multiple ratings. The obtained ground truth, categorized into 3 groups, served as a target variable for machine learning modelling.

Regarding related work on the machine learning approach, multiple machine learning classifiers were tested in most of the studies where the ones in common were usually random forest, support vector machines and decision trees. In our study, XGBoost performed best for multiclass classification and was followed by random forest. Study [14] that also performed multiclass classification reports on random forest having the best performance in terms of accuracy. Other studies are using machine learning for binary classification of successful ageing where in [59, 60] random forest behaved best and was followed by XGBoost. Study [16] reports on an adaptive network-based fuzzy inference system being a superior method and study [15] reports on the KNN-based ensemble method being the best. By reviewing the literature we can conclude there is no specific, commonly used dataset on older adults that would be used for performance benchmarking of different approaches and machine learning methods to predict healthy ageing. Several studies exist but each uses different datasets size and features obtained in various territories such as England [14], India [58] and Iran [16].

Explainability results in this study show that social and mental health components such as achieving meaning and life satisfaction, participation in publicly renowned or socially visible organizations, awareness that one own's life experiences are passed on to others and mental well-being are dominating in its contribution to healthy ageing. These results are aligned with the study [59] that also reports on life satisfaction, quality of life and official social relationships being the best factors affecting successful ageing. Similarly, study [60] also reports that factors such as social functional, social interpersonal relationship, depression and hypertension are important for predicting successful ageing.

In terms of applicability, we see the potential of the proposed healthy ageing scale to 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 unidimensional 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 [61], the proposed scale 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.

Potential limitations were noted during the study. First, 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. While data captured at a single time was used in the healthy ageing literature before [16], several studies use longitudinal datasets [14, 58]. Multiple participants whose data is captured in the dataset used in this study consented to a follow-up interview. Therefore, in the future, there is room to add a broader set of information, from the perspective of both time (longitudinal aspect could be introduced) and content (for example measured tests could be added). Second, the dataset used in this study is of moderate size with 696 cases. While we found a dataset of similar size was also used elsewhere in the research [15–17], several studies utilize a larger dataset [14, 58]. We might attribute this to larger countries having more resources for conducting such interviews than Slovenia and having a larger population; therefore, the available sample is also bigger. Third, the dataset used in this study stores information on people aged 50 or older, termed "early old age". While this is consistent with previous literature [14, 58], some definitions of healthy ageing define older people as people aged 60 or older [3, 9]. Therefore, our healthy ageing scale might apply to the younger generation of older adults without many chronic diseases and conditions. Next, 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. Furthermore, the classifier that performed best was XGBoost, which is considered a black box technique. As trust in the results can only be driven by end-user understanding of given model predictions, we tackled this challenge by utilizing the SHAP explainability framework.

Conclusion

Throughout this study, we investigated the feasibility of building a healthy ageing scale utilizing machine learning techniques fed by human-based annotations and demographics, health data (physical, social, mental) and activities. During the process, we closely cooperated with gerontology experts to identify the most relevant input variables/predictors that influence healthy ageing. We tested multiple classifiers with XGBoost performing best in terms of macro-averaged AUC and F1. Due to the black-box nature of the algorithm, we applied the SHAP framework for interpreting predictions. To our knowledge, this is the first study that uses a combination of active involvement of gerontology domain experts, machine learning and prediction explainability techniques to create a healthy ageing score that has the potential to be trusted and understood by informal carers.

Future work may include the implementation of a model and explainability application programming interface (API) endpoint which could be embedded into end-user applications like a decision support system for healthy ageing improvement. Further collaboration with gerontology experts would be applied to validating model results interpretation and development and evaluation of such recommendation system. Furthermore, the use of additional data to enhance the accuracy of the scale could be applied. Such data could comprise information captured via longitudinal studies and standardized tests (e.g. walking tests). Behaviour data could be captured via intelligent devices. Older adults could be split by age, gender or other categories and individual machine learning models could be developed for each. In terms of governance besides predictions explainability techniques already used in this study, additional aspects of governance could be explored such as identifying and mitigating potential model bias that can arise from the data.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-024-02714-w.

Additional file 1. Contains information on all itemsand possible answer choices used in this study. Items are categorized into health categories and sub-categories as chosen by gerontology experts.

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.

Authors' contributions

K.G.S and A.K. designed the study. A.R., J.R. and A.K. curated the data. K.G.S. preprocessed the data, performed the analyses and produced the results. A.R., J.R. and A.K. validated the results. A.R. and J.R. provided domain knowledge. A.K. supervised the study. K.G.S., A.R, J.R. and A.K. analysed the results. K.G.S. drafted the manuscript. All authors read, revised, and approved the final manuscript.

Funding

This research was partly supported by the project P2-0246 ICT4QoL - Information and Communications Technologies for Quality of Life.

Data availability

The datasets used and analyzed during the study are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate

Data collection was not part of the research presented in this paper. Data used in this study was obtained through the independent research "Ageing in Slovenia: Survey on the Needs, Abilities and Standpoints of the Slovene Population Aged 50 Years and Over" [39] within which a questionnaire was developed and data was collected via in-person interviews. The questionnaire is not publicly available. The National Medical Ethics Committee of the Republic of Slovenia considered the questionnaire and the research concept for study in [39], and an opinion was issued that the research was ethically impeccable. Ethical consent (nr. 115/09/09) was issued for its implementation [40] and informed consent was obtained from all participants included in the study. The research reported here in this paper was completely aligned with the aims of the data collected and no additional ethics-related issues were opened. During data collection, 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 analyzed in this study was anonymized and free of any personally identifiable information (PII).

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]IBM Slovenija d.o.o., Ameriška ulica 8, 1000 Ljubljana, Slovenia. [2]Anton Trstenjak Institute of Gerontology and Intergenerational Relations, Resljeva cesta 7, 1000 Ljubljana, Slovenia. [3]Laboratory for user-adapted communications and ambient intelligence, Faculty of Electrical Engineering, Tržaška cesta 25, 1000 Ljubljana, Slovenia.

Received: 25 June 2024 Accepted: 8 October 2024
Published online: 29 October 2024

References

1. Economic UND, Affairs S. World Population Ageing 2020: Highlights. United Nations Secretariat New York: United Nations; 2021. https://doi.org/10.18356/9789210051934.
2. Lutz W, Sanderson W, Scherbov S. The coming acceleration of global population ageing. Nature. 2008;451(7179):716–9. https://doi.org/10.1038/nature06516.
3. World Health Organization. Decade of healthy ageing: baseline report. 2020. https://www.who.int/publications/i/item/9789240017900. Accessed 20 Dec 2023.
4. Secretary General UN. Review and appraisal of the Programme of Action of the International Conference on Population and Development and it's contribution to the follow-up and review of the 2030 Agenda for Sustainable Development: report of the Secretary-General. 2019. https://www.un.org/development/desa/pd/content/review-and-appraisal-programme-action-international-conference-population-and-development. Accessed 20 Dec 2023.
5. Weiland S, Hickmann T, Lederer M, Marquardt J, Schwindenhammer S. The 2030 agenda for sustainable development: transformative change through the sustainable development goals? Polit Gov. 2021;9(1):90–95. https://doi.org/10.17645/pag.v9i1.4191.
6. Bryant LL, Corbett KK, Kutner JS. In their own words: a model of healthy aging. Soc Sci Med. 2001;53(7):927–41. https://doi.org/10.1016/s0277-9536(00)00392-0.
7. Michel JP, Sadana R. "Healthy aging" concepts and measures. J Am Med Dir Assoc. 2017;18(6):460–4. https://doi.org/10.1016/j.jamda.2017.03.008.
8. Organization WH, editor. World report on ageing and health. Switzerland: World Health Organization; 2015.
9. Peel N, Bartlett H, McClure R. Healthy ageing: how is it defined and measured? Australas J Ageing. 2004;23(3):115–9. https://doi.org/10.1111/j.1741-6612.2004.00035.x.
10. Lu W, Pikhart H, Sacker A. Domains and Measurements of Healthy Aging in Epidemiological Studies: A Review. Gerontologist. 2019;59(4):294–310. https://doi.org/10.1093/geront/gny029.
11. Ågren GBK. Healthy ageing: a challenge for Europe. Stockholm: Swedish National Institute of Public Health; 2007.
12. Sadana R. Development of standardized health state descriptions. Geneva: World Health Organization; 2002. pp. 315–328. Chap. 7.1
13. Wong RY. A new strategic approach to successful aging and healthy aging. 2018. https://doi.org/10.3390/geriatrics3040086.

14. Caballero FF, Soulis G, Engchuan W, Sánchez-Niubó A, Arndt H, Ayuso-Mateos JL, et al. Advanced analytical methodologies for measuring healthy ageing and its determinants, using factor analysis and machine learning techniques: the ATHLOS project. Sci Rep. 2017;7(1):1–13. https://doi.org/10.1038/srep43955.
15. Asghari Varzaneh Z, Shanbehzadeh M, Kazemi-Arpanahi H. Prediction of successful aging using ensemble machine learning algorithms. BMC Med Inform Decis Making. 2022;22(1):258. https://doi.org/10.1186/s12911-022-02001-6.
16. Yazdani A, Shanbehzadeh M, Kazemi-Arpanahi H. Using an adaptive network-based fuzzy inference system for prediction of successful aging: a comparison with common machine learning algorithms. BMC Med Inform Decis Making. 2023;23(1):229. https://doi.org/10.1186/s12911-023-02335-9.
17. Ahmadi M, Nopour R. Clinical decision support system for quality of life among the elderly: an approach using artificial neural network. BMC Med Inform Decis Making. 2022;22(1):293. https://doi.org/10.1186/s12911-022-02044-9.
18. Gialluisi A, Di Castelnuovo A, Donati MB, De Gaetano G, Iacoviello L, sani Study Investigators M. Machine learning approaches for the estimation of biological aging: the road ahead for population studies. Front Med. 2019;6:146. https://doi.org/10.3389/fmed.2019.00146.
19. Chien SY, Chao SF, Kang Y, Hsu C, Yu MH, Ku CT. Understanding Predictive Factors of Dementia for Older Adults: A Machine Learning Approach for Modeling Dementia Influencers. Int J Hum-Comput Stud. 2022;165:102834. https://doi.org/10.1016/j.ijhcs.2022.102834.
20. Adhikari S, Thapa S, Naseem U, Singh P, Huo H, Bharathy G, et al. Exploiting linguistic information from Nepali transcripts for early detection of Alzheimer's disease using natural language processing and machine learning techniques. Int J Hum-Comput Stud. 2022;160:102761. https://doi.org/10.1016/j.ijhcs.2021.102761.
21. Cicirelli G, Marani R, Petitti A, Milella A, D'Orazio T. Ambient assisted living: A review of technologies, methodologies and future perspectives for healthy aging of population. Sensors. 2021;21(10):3549. https://doi.org/10.3390/s21103549.
22. Thieme A, Belgrave D, Doherty G. Machine learning in mental health: A systematic review of the HCI literature to support the development of effective and implementable ML systems. ACM Trans Comput-Hum Interact. 2020;27(5):1–53. https://doi.org/10.1145/3398069.
23. Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Consortium P. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Making. 2020;20:1–9. https://doi.org/10.1186/s12911-020-01332-6.
24. Weber P, Carl KV, Hinz O. Applications of Explainable Artificial Intelligence in Finance–a systematic review of Finance, Information Systems, and Computer Science literature. Manag Rev Q. 2023;7:1–41. https://doi.org/10.1007/s11301-023-00320-0.
25. Owens E, Sheehan B, Mullins M, Cunneen M, Ressel J. Explainable Artificial Intelligence (XAI) in Insurance: A Systematic Review. Risks. 2022;10(12):230. https://doi.org/10.3390/risks10120230.
26. Mehdiyev N, Houy C, Gutermuth O, Mayer L, Fettke P. Explainable artificial intelligence (XAI) supporting public administration processes–on the potential of XAI in tax audit processes. In: Innovation Through Information Systems: Volume I: A Collection of Latest Research on Domain Issues. vol. 1. Switzerland: Springer International Publishing; 2021. pp. 413–428. https://doi.org/10.1007/978-3-030-86790-4_28.
27. Vimbi V, Shaffi N, Mahmud M. Interpreting artificial intelligence models: a systematic review on the application of LIME and SHAP in Alzheimer's disease detection. Brain Inform. 2024;11(1):10. https://doi.org/10.1186/s40708-024-00222-1.
28. Ahmed S, Shamim Kaiser M, Hossain MS, Andersson K. A Comparative Analysis of LIME and SHAP Interpreters with Explainable ML-Based Diabetes Predictions. IEEE Access. 2024:1. https://doi.org/10.1109/ACCESS.2024.3422319.
29. Chaddad A, Peng J, Xu J, Bouridane A. Survey of explainable AI techniques in healthcare. Sensors. 2023;23(2):634. https://doi.org/10.3390/s23020634.
30. Aldughayfiq B, Ashfaq F, Jhanjhi NZ, Humayun M. Explainable AI for retinoblastoma diagnosis: interpreting deep learning models with LIME and SHAP. Diagnostics. 22023;13(11):1932. https://doi.org/10.3390/diagnostics13111932.
31. Lombardi A, Diacono D, Amoroso N, Monaco A, Tavares JMRS, Bellotti R, et al. Explainable deep learning for personalized age prediction with brain morphology. Front Neurosci. 2021;15:578. https://doi.org/10.3389/fnins.2021.674055.
32. Tang YT, Romero-Ortuno R. Using explainable AI (XAI) for the prediction of falls in the older population. Algorithms. 2022;15(10):353. https://doi.org/10.3390/a15100353.
33. Kim R, Kim CW, Park H, Lee KS. Explainable artificial intelligence on life satisfaction, diabetes mellitus and its comorbid condition. Sci Rep. 2023;13(1):11651. https://doi.org/10.1038/s41598-023-36285-z.
34. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: von Luxburg U, Guyon I, Bengio S, Wallach H, Fergus R, editors. Proceedings of the 31st International Conference on Neural Information Processing Systems. vol. 1. United States: Curran Associates Inc.; 2017. pp. 4768–4777.
35. Campagner A, Ciucci D, Svensson CM, Figge MT, Cabitza F. Ground truthing from multi-rater labeling with three-way decision and possibility theory. Inf Sci. 2021;545:771–90. https://doi.org/10.1016/j.ins.2020.09.049.
36. Ratner AJ, De Sa CM, Wu S, Selsam D, Ré C. Data programming: creating large training sets, quickly. In: Lee DD, von Luxburg U, Gernett R, Sugiyama M, Guyon I, editors. Proceedings of the 30th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc.; 2016. pp. 3574–3582.
37. Mayer RE, Moreno R. Nine ways to reduce cognitive load in multimedia learning. Educ Psychol. 2003;38(1):43–52. https://doi.org/10.1207/S15326985EP3801_6.
38. Košir A, Strle G, Meža M. Weak Ground Truth Determination of Continuous Human-Rated Data. IEEE Access. 2020;9:4594–606. https://doi.org/10.1109/ACCESS.2020.3046293.
39. Ramovš J. Staranje v Sloveniji: raziskava o potrebah, zmožnostih in stališčih nad 50 let starih prebivalcev Slovenije. Slovenija: Inštitut Antona Trstenjaka; 2013.
40. Ramovš J. Potrebe, zmožnosti in stališča starejših ljudi v Sloveniji. Kakovostna Starost. 2011;14(2):3–21.
41. Hayton JC, Allen DG, Scarpello V. Factor retention decisions in exploratory factor analysis: a tutorial on parallel analysis. Organ Res Methods. 2004;7(2):191–205. https://doi.org/10.1177/1094428104263675.
42. Zwick W, Velicer W. Comparison of Five Rules of Determining the Number of Components to Retain. Psychol Bull. 1986;99:432–42. https://doi.org/10.1037/0033-2909.99.3.432.
43. Watkins MW. Determining Parallel Analysis Criteria. J Mod Appl Stat Methods. 2006;5(2):344–6.
44. Django. Django software fundation. 2019. https://djangoproject.com. Accessed 20 Dec 2023.
45. Krippendorff K. Computing Krippendorff's alpha-reliability. 2011.
46. Dalianis H. Evaluation metrics and evaluation. In: Clinical text mining: secondary use of electronic patient records. Sweden: Springer; 2018. pp. 45–53. https://doi.org/10.1007/978-3-319-78503-5_6.
47. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. vol. 1. Association for Computing Machinery; 2016. pp. 785–794. https://doi.org/10.48550/arXiv.1603.02754.
48. Bentéjac C, Csörgő A, Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms. Artif Intell Rev. 2021;54(3):1937–67. https://doi.org/10.1007/s10462-020-09896-5.
49. Tsopra R, Fernandez X, Luchinat C, Alberghina L, Lehrach H, Vanoni M, et al. A framework for validating AI in precision medicine: considerations from the European ITFoC consortium. BMC Med Inform Decis Making. 2021;21(1):1–14. https://doi.org/10.1186/s12911-021-01634-3.
50. Galar M, Fernández A, Barrenechea E, Bustince H, Herrera F. An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. Pattern Recognit. 2011;44(8):1761–76. https://doi.org/10.1016/j.patcog.2011.01.017.
51. Liu F, Zhou P, Baccei SJ, Masciocchi MJ, Amornsiripanitch N, Kiefe CI, et al. Qualifying certainty in radiology reports through deep learning-based natural language processing. Am J Neuroradiol. 2021;42(10):1755–61. https://doi.org/10.3174/ajnr.A7241.
52. Grandini M, Bagli E, Visani G. Metrics for Multi-Class Classification: an Overview. 2020. arXiv preprint arXiv:2008.05756.
53. Mohanty A, Mishra S. A comprehensive study of explainable artificial intelligence in healthcare. In: Mishra S, Kumar Tripathy H, Mallick P,

Shaalan K, editors. Augmented intelligence in healthcare: a pragmatic and integrated analysis. Singapore: Springer Nature Singapore; 2022. pp. 475–502. https://doi.org/10.1007/978-981-19-1076-0_25.
54. Cho E. Making reliability reliable: a systematic approach to reliability coefficients. Organ Res Methods. 2016;19(4):651–82. https://doi.org/10.1177/1094428116656239.
55. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57. https://doi.org/10.1613/jair.953.
56. Yi F, Yang H, Chen D, Qin Y, Han H, Cui J, et al. XGBoost-SHAP-based interpretable diagnostic framework for alzheimer's disease. BMC Med Inform Decis Making. 2023;23(1):137. https://doi.org/10.1186/s12911-023-02238-9.
57. Lundberg SM, Erion G, Chen Hea. From local explanations to global understanding with explainable AI for trees. Nat Mach Intel. 2020;2:56–67. https://doi.org/10.1038/s42256-019-0138-9.
58. Das A, Dhillon P. Understanding healthy ageing in India: insights from multivariate regression trees. Aging Clin Exp Res. 2024;36(1):1–10. https://doi.org/10.1007/s40520-024-02815-6.
59. Ahmadi M, Nopour R, Nasiri S. Developing a prediction model for successful aging among the elderly using machine learning algorithms. Digit Health. 2023;9:1–22. https://doi.org/10.1177/20552076231178425.
60. Mirzaeian R, Nopour R, Asghari Varzaneh Z, Shafiee M, Shanbehzadeh M, Kazemi-Arpanahi H. Which are best for successful aging prediction? Bagging, boosting, or simple machine learning algorithms? Biomed Eng Online. 2023;22(1):85. https://doi.org/10.1186/s12938-023-01140-9.
61. M C, Y S, et al HZA. Implementing care for healthy ageing. BMJ Glob Health. 2022;7(2):e007778. https://doi.org/10.1136/bmjgh-2021-007778.

Publisher's Note