Health status prediction for the elderly based on machine learning

Fang-Yu Qin[11,a], Zhe-Qi Lv[1,b], Dan-Ni Wang[c], Bo Hu[d,*], Chao Wu[c,e,*]

[a] *Department of Software Engineering, Zhejiang University, Hangzhou, China*

[b] *Department of Marine Informatics, Zhejiang University, Hangzhou, China*

[c] *Department of Public Affairs, Zhejiang University, Hangzhou, China*

[d] *Care Policy and Evaluation Centre, Department of Health Policy, London School of Economics and Political Science, London, UK*

[e] *Center of Social Welfare and Governance, Zhejiang University, Hangzhou, China*

ARTICLE INFO

Keywords
Machine learning
Elderly
Health prediction
Social service
Data-driven

ABSTRACT

Health and social care services are crucial to old people. The provision of services to the elderly with care needs requires more accurate predictions of the health status of the elderly to rationalize the allocation of the limited social care resources. The traditional analytical methods have proved incapable of predicting the demands of today's society, compared to which machine learning methods can more accurately capture the nonlinear relationships between the variables. To ascertain visually the performance of these machine learning methods regarding the prediction of the elderly's care needs, we designed and verified the experiment.

[1] Both authors contributed equally to this work.

[*] Corresponding authors at: Zijingang Campus of Zhejiang University, Yuhangtang Road No.388, Zhejiang Province, China (Chao Wu); 8.01 Pankhurst House, Clement's Inn, London WC2A 2AE, UK (Bo Hu). E-mail address: chao.wu@zju.edu.cn (Chao Wu), b.hu@lse.ac.uk (Bo Hu).

Highlights

1. The machine learning methods help researchers select the predictors of health status in the older population efficiently.

2. The machine learning methods automatically capture the complicated relationships between the non-linear predictors and the health outcomes.

3. The artificial neural networks have the best prediction accuracy in relation to older people's self-reported health.

1. Introduction

When people reach old age, some experience a general decline in their health or gradually lose their functional capability to perform basic but highly valued daily activities, such as bathing, washing or eating. The past few decades have seen a rapid change in the population structure in many countries. Older people have been the fastest-growing group in the population. At the same time, a massive quantity of public resources is devoted to meet the care needs of older people each year.

Health and social care services help older people to recover from disease, alleviate the symptoms of their illness, and enable them to live independently in their own homes, all of which improve people's quality of life and overall wellbeing. However, the care resources are limited, so effectively allocating them and providing high-quality services for the elderly is a salient challenge facing both developing and developed countries alike. Therefore, it is vital to make accurate predictions regarding the health status of the elderly so that services can be provided according to their care needs.

To make an accurate prediction of the health status of the elderly, we must analyze the factors that are most strongly associated with the health of the elderly in the first place. Against this backdrop, a massive body of literature has investigated this issue, with the knowledge and expertise in this area piling up rapidly. In particular, the availability of large datasets in the information age provides researchers with opportunities to gain a clearer understanding of the complex relationships existing between the variables and enables them to conduct more rigorous research. However, as the amount of information accessible to researchers increases exponentially, the practical challenge arises of how to process this information effectively (namely identifying the most relevant predictors, and utilizing these to make accurate predictions). It takes a long time for researchers to familiarize themselves with these large amounts of data and identify predictors that are not found in the existing literature (manual feature engineering). By the time this task is finished, new data will be available, and the processed information will be already outdated.

The machine learning methods can be used to extract nonlinear and seemingly insignificant influential factors that were difficult to find using the conventional methods, thus enabling the feature selection process to be completed more accurately. Machine learning methods can also greatly reduce the time required to extract important factors from large data sets and ultimately improve prediction accuracy.

The structure of this paper is as follows: we first briefly introduce the topic to be studied and its background. In the second section, the existing literature on the determinants of the health of the elderly population is reviewed. In the third part, the description of the data source and the design of the experiment is presented. In the fourth part, the experimental results are outlined. In the fifth part, based on the research results, we discuss the methodology and theoretical significance of the research. In the sixth part, the main conclusion of this research is summarized. In the machine learning literature, input variables are normally termed as features, whereas in social science studies input variables go by the names of predictors or independent variables. These terms are used interchangeably in this paper.

2. Literature review

The World Health Organization (2015) defines health as 'a state of complete physical, mental and social well-being, and not merely the absence of disease or infirmity' (p.228). Health is a multi-dimensional construct. For older people, physical health, mental well-being, and cognitive capability are all closely linked to their quality of life. Regarding physical health, there are three commonly used indicators in the literature: the presence of illness, the level of functional capability and subjective evaluation of health (O'Donnell et al, 2008). This study focuses on the latter two indicators of physical health.

Many studies have examined the factors affecting the physical capability and self-reported health of older people. These factors can be divided into four categories: demographic, socio-economic, lifestyle and disease factors in older people.

The research in the field of demography has focused on age, gender and the lifestyle of older people. Age is considered the most important cause of functional disability in older people. Hebert conducted longitudinal research on a representative sample of individuals aged over 75 years old. The results show that. for every five years of age, the risk of functional decline doubles (Hebert, Brayne, & Spiegelhalter, 1997). At the same time, gender is also one of the factors to consider. Researchers found that older people's ability to use the toilet is impaired before they lose their ability to get dressed. The reverse pattern is often seen among old women, which indicates a gender difference regarding the progression of disability (Dunlop, Hughes, & Manheim, 1997). However, there are no significant gender differences in terms of the incidence of disability (Oman, Reed, & Ferrara, 1999). Moreover, older people who live alone are more likely to suffer from disabilities during the aging process because of the lack of care from their partner/children (Bonsang, 2009).

Self-reported health (SRH) is considered a valuable source of data on various aspects of general health (Idler, Benyamini, 1997; Nummela, Sulander, Heinonen, & Uutela, 2007). People in higher age groups generally report a decline in general health. (Li, Meng, Wang, Ma, Chen, & Liu, 2017). Meanwhile, SRH is worse among women than men (Ties, Ahmad, Emese, & Somnath, 2016).

According to previous research, the socio-economic environment in which older people live can significantly affect the extent of their functional capability and general health status. Parker et al. (1994) used the activities of daily living to explore the relationship between physical disability and social class among elderly people in Europe. Former white-collar workers have better functional capability than former blue-collar workers. A survey on functional capability for the elderly in Bangladesh showed that the place of residence has a large impact. Higher rates of mortality and disability have been observed in rural areas compared to urban areas (Shariful, Ismail, Nazrul, Ahbab, Hafiz, & Sharifa, 2017). Avlund et al. (2008) studied a representative sample of 75-year-old men and women living in a suburban area west of Copenhagen, Denmark and found that the impact of income was significantly higher than occupation or education. Studies have shown that poor SRH is more prevalent among poor and socially disadvantaged groups. Migrants are more likely to report their health as poor (Li, Meng, Wang, Ma, Chen, & Liu, 2017).

Similarly, different lifestyles often result in varying degrees of physical disability. For example, being overweight due to an unhealthy diet and smoking are associated with a greater risk of disability (Artaud, Dugravot, Sabia, et al., 2012), but the causality between drinking alcohol and the physical functioning of the elderly continues to be debatable. Artaud argued that drinking alcohol is unrelated to physical disability in the elderly, while

Benfante et al. (1995) analyzed the effect of alcohol intake and found that older people who did not drink alcohol were more likely to remain disease-free. Physical exercise is recognized as a way to delay the development of physical disability in later life. Tak et al. (2013) presented a meta-analysis of the association between PA and the incidence and progression of basic ADL disability. He divided the exercise regime into three levels (low, medium and high), calculated their effects on physical health, and found that physical medium and high-level exercise slowed the progress of functional disability in older people. However, the association between weight and SRH was not found to be significant among migrant laborers in China. The data indicated that alcohol consumption and SRH are related and some studies show that excessive alcohol consumption increased the likelihood of poor self-reported health (Szmitko, & Verma, 2005; Tsai, Ford, Li, Pearson, & Zhao, 2010; Jimenez, Chiuve, Glynn, Stampfer, Camargo Jr, Willett, et al., 2012).

Prevalence of illness and mental health is not the focus of discussion in this study, but their correlative relationships with people's functional capability and SRH cannot be ignored. Previous studies pointed out that chronic diseases (such as arthritis, diabetes, hypertension, cardiovascular disease) and functional aging (such as hearing loss and vision deterioration), that are gradually manifested with age result in the development of functional disability among older people (Boult, Kane, Louis, & McCaffrey, 1994). Although some chronic diseases do not directly make it more difficult for the elderly to complete their daily activities, they can predict future disability (Fried, Herdman, Kuhn, et al., 1991).0 Meanwhile, other researchers have reported that mental illness, especially depression, affects functional disability among older people. There is clear evidence of a positive correlation between functional disability and depression (Greenglass, Fiksenbaum, & Eaton, 2006). This is mainly because older people with depression are more likely to have negative and resistive emotions when facing arduous daily tasks, and their pessimistic mental state will cause them to participate less in social activities and receive less social support. Unsurprisingly, the presence of chronic illnesses such as diabetes, coronary heart disease, and elevated BMI is positively associated with SRH. A family history of breast cancer or coronary heart disease, or knowledge of an increased risk of genetic disease, can also affect a person's SRH (Krijger, Schoofs, Marchal, Van De Vijver, Borgermans, & Dervroey, 2014).

Most of the research discussed above used the traditional research methods in social science or simple machine learning, such as multiple linear regression, which are typically based on the manual selection of features. Using these methods, researchers were likely to select features suggested by the existing literature, and it was difficult to find nonlinear correlations between the variables. Such a feature selection approach is inevitably subjective.

Machine learning methods are considered a branch of artificial intelligence research. A series of algorithms enable computers to learn from data and use the established models and new inputs to make predictions without the need for explicit external indications (Casanova, Saldana, Lutz, Plassman, Kuchibhatla, & Hayden, 2017). Researchers have now begun to apply the machine learning methods to address tough research questions in gerontological and geriatric studies. Oscar et al. (2017) examined the stigma of Alzheimer's disease using the ML techniques. They modeled the stigmatization expressed in 31,150 tweets related to Alzheimer's disease. Facal et al. (2019) used a variety of machine learning algorithms (e.g., linear regression, support vector machines, Gaussian naive Bayes) to explore the role of cognitive reserve (CR) in the conversion from mild cognitive impairment (MCI) to dementia. In this study, machine learning methods are used to investigate the predictors of the health conditions of older people. The powerful ability of the machine learning methods to process massive amounts of data and capture nonlinear features in a short time makes our approach especially appealing to researchers who have rich information but face time constraints to process it. At the same time, because the whole process (from feature selection to modeling) is data-driven, the methods used in this research should be replicable to studies in other fields.

3. Experiment

This experiment is mainly divided into the following sections:

- Data preview. Obtaining the original data set, importing it into the database, and observing its data structure.
- Data merge. In the original data set, the questionnaires for each topic are separated. For ease of analysis, it is useful to merge these into a single data set.
- Feature selection. Extracting the factors that have the greatest impact on the health of the elderly through the application of machine learning algorithms.
- Data processing. Processing the data until the data can be used to train machine learning models.
- Training models. Training several machine learning models using the training set for comparative experiments.
- 5-fold cross-validation. Randomly dividing the entire dataset into 5 groups, designating one group as the validation set and the rest as the training set, and executing the validation five times.
- Results. Obtain the results and further summarize the conclusions.

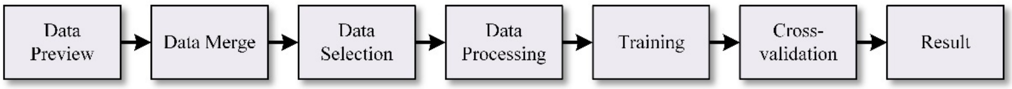

Figure 1 Workflow

3.1. Data Preview

Extracting data from the 2013 and 2015 China Health and Retirement Longitudinal Survey (CHARLS) is the first step of the entire experiment. Employing a multi-stage sampling, the 2013 wave of CHARLS covered 28 Chinese provinces and surveyed a sample of 18,333 Chinese people aged 45 and over in 10,803 households in 150 counties. The CHARLS 2015 was a follow-up study that re-contacted the previous respondents, among whom 2,290 people dropped out. With a refreshed sample of 4,802 people, a total of 20,845 respondents participated in the CHARLS 2015. We removed the observations with missing values in the self-reported health and functional limitations variables and those with missing values in more than 80% of the features, which resulted in a sample size of 29,477 for the combined dataset. The survey questionnaire contains basic personal information, family structure, financial support, health status, physical measurement, medical service utilization, medical insurance, work, retirement and pension, income, consumption, assets, and the basic community conditions. The attrition rate is 23%. The advantage of using the 2013 and 2015 wave is that the data for these two years were gathered using a similar questionnaire design and the sample size is relatively large.

Table 1
CHARLS Information

QUESTION NUMBER	CONTENT
B	DEMOGRAPHIC BACKGROUND
C	FAMILY
D	HEALTH STATUS AND FUNCTIONING
E	HEALTH CARE AND INSURANCE
F	WORK, RETIREMENT, AND PENSION
G&H	INCOME, EXPENDITURE, AND ASSETS
I	HOUSING CHARACTERISTICS

- DEMOGRAPHIC BACKGROUND: name, date of birth, address, etc.
- FAMILY: parent, childrearing and sibling information, time transfers, etc.
- HEALTH STATUS AND FUNCTIONING: health status, functional limitations, helpers, etc.
- HEALTH CARE AND INSURANCE: medical insurance, health care costs and utilization.
- WORK, RETIREMENT, AND PENSION: job status, fringe benefits, etc.
- INCOME, EXPENDITURE, AND ASSETS: household income and expenditure, household assets, etc.
- HOUSING CHARACTERISTICS: total housing land area, the year when the house was built, etc.

3.2. Data Merging

We entered all of the data onto the database, then each data table was added as a large table based on the respondents' ID. The database chosen here was MongoDB.

Unlike SQL Server, MySQL, etc., MongoDB is a NoSQL database. It can store data in documents, and the data are stored as key-value pairs. Keys are used uniquely to identify a document and are of a string type, while the values can be of various complex file types. This form of storage is called BSON. BSON is a JSON-like storage format in a binary form, which is also briefly referred to as Binary JSON. Such characteristics mean that MongoDB performed better compared to the SQL type database in this experiment because we mainly used indexless addition, deletion, modification and search operations.

3.3. Feature Selection

Each wave of CHARLS contains more than 3,500 features. We used the Maximal Information Coefficient (MIC) and Pearson Correlation Coefficient to calculate the most relevant factors. MIC is a novel correlation statistic that measures the strength of the linear or non-linear correlation between two variables, X and Y. The Pearson correlation coefficient (PCC) is a measure of the linear correlation between two variables. According to the Cauchy-Schwarz inequality, it has a value between +1 and −1, where +1 means a complete positive correlation, -1 means a complete negative correlation, and 0 means no linear correlation. It is widely used in science (Aldaz, Barza, Fujii, & Moslehian, 2015). We calculated the MIC (Maximal information coefficient) value and Pearson value for each feature, and then select factors with a high MIC value and low Pearson value to obtain the nonlinear correlation factors.

There are two concepts involved here: linear and non-linear correlation. A linear correlation exists when two quantities are proportional to one another. If we increase one of the quantities, the other quantity will either increase or decrease at a constant rate. For example, if you get paid $30 an hour, there is a linear relationship between your working hours and your salary. Working another hour always results in a $30 pay increase, regardless of how many hours you have already worked. Some non-linear correlations are monotonic, meaning they always either increase or decrease, but not both. Monotonic relationships differ from linear correlations because they do not increase or decrease at a constant rate. Non-linear features are used to predict health status here.

If researchers have access to linear features, a linear model such as logistic regression can be used to fit the data. If instead, the features are non-linear by nature, non-linear models, such as artificial neural networks, can be used. A nonlinear model captures the nonlinear relationships in the experimental data. Nonlinear models are specified by nonlinear equations and often are non-parametric. Non-parametric nonlinear models are common in machine learning methods.

3.4. Data Processing

The data for the 2013 and 2015 questionnaires were used primarily for the next steps.

Machine learning models such as linear regression and the Support Vector Machine (SVM) are highly sensitive to missing values in the data set. We imputed the missing values before feature selection. We experimented with four machine learning models including linear regression, k-Nearest Neighbors (kNN), Decision Tree, and XGBoost. We used the XBGoost algorithm since it showed the best performance in terms of imputation.

We examined two indicators of physical health in this study: self-reported health and the level of functional limitations. Health status was dichotomized, with 1 representing good health and 0 representing poor health. The CHARLS survey asked each respondent whether they could perform six activities of daily living (ADLs) and six instrumental activities of daily living (IADLs). For each item, people's functional capability was measured on a 4-point scale: 1 = I do not have difficulty, 2 = I have difficulty but can do it, 3 = I need help, and 4 = I cannot do it. We added up the scores of each item to create an indicator for functional capability, with a higher total score indicating a more severe functional limitation.

Data normalization was an important part of data processing. Some machine learning models are more vulnerable than others to highly variant scales in the dataset. To prevent larger values in the data from overwriting smaller ones, we rescaled all of the features so that the values fall into a uniform range between 0 and 1. This is done by subtracting the original values of the features by their mean values and dividing them by the variance.

3.5. Training

Once the above data processing had been completed, a table with dimensions of 29447×15 can be obtained. The 29744 rows represent the total number of observations. The 15 columns contain ID, the outcome variable (i.e., health status) and 13 features selected by the algorithms. Next, we predict health status by asking questions through the machine learning models.

Our research focuses on the performance of the nonlinear model in predicting the health status of older people, so the linear model is mainly used as an auxiliary comparative study. Here, we give a detailed description of the artificial neural networks given its central role in our experiment. Due to limited space, we only provide a brief overview of other machine learning algorithms.

For the artificial neural networks (ANN), learning is viewed as the process of updating the internal representation of the system in response to external stimuli so that it can perform a specific task (Zell, 2003). When we provide training data to the network, the ANN will learn about the data in an iterative manner, similar to how we learn from experience.

In this experiment, 13 features were obtained following the feature selection. We set 13 nodes in the input layer to receive 13 features. In the case of predicting self-reported health, the activation function was adjusted so that the ANN outputs a binary result (1 for good health and 0 for poor health). The number of hidden layers and the number of hidden nodes are hyperparameters researchers have to carefully choose. Networks with an insufficient number of hidden nodes will underfit the data and result in poor prediction accuracy. Conversely, if a network has too many hidden nodes, it will mistakenly treat random errors in the data as patterns. This will result in poor generalization to untrained data, which is also known as overfitting of the data (figure 2):

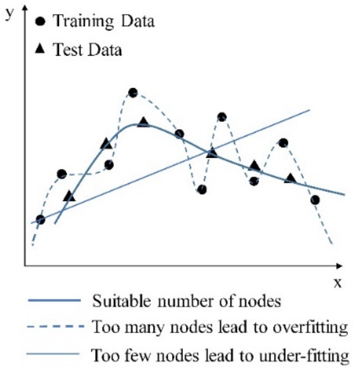

Figure 2 Overfitting example

The figure below presents a structural diagram of the neural network used in this experiment. In this neural network, 200 nodes were set in the hidden layer, and each layer was connected by an activation function. Neural network activation functions are a crucial component of deep learning. The activation function determines the output of the deep learning model, its accuracy and the computational efficiency of the training model. A Rectified Linear Unit (ReLU) was used in the training process and Softmax function was used in the output layer to obtain a binary result. During the training process, 50% of the nodes were randomly discarded to prevent overfitting, and the neural network was obtained after 200 rounds of training.

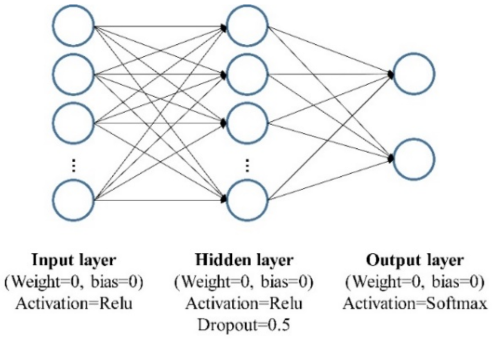

Figure 3 Neural network structure

Apart from the artificial neural networks, we also trained other commonly used machine learning models, such as Random forests, XGBoost, support vector machine, and logistic regression. The implementation of the random forest algorithm involves bootstrapping and de-correlating the decision tree models (Breiman, 2011). The decision tree models predict outcomes by dividing the predictor space into areas and collating the prediction results in each area. XGboost, like other boosting algorithms, constructs a powerful learner by combining a large number of weaker learners such as heavily pruned decision trees (Chen, & Carlos, 2016). Each learner is weak in the sense that it predicts poorly in the entire predictor space. However, they have good prediction performance in their respective local predictor space and thus can be regarded as 'specialists'. Support vector machines classify observations or make predictions by constructing a set of hyperplanes in a high dimensional space (Cortes, & Vapnik, 1995).

3.6. *Cross-validation*

The execution of the machine learning approach usually involves splitting the dataset into a training set and a test set (or validation set). A model built with the training set is tested on the test set to make sure that the model generalizes well as new data come in and that there is no over-fitting of the data. In this experiment, we used the 5-fold cross-validation as the verification approach. We randomly partitioned the sample into 5 equal-sized subsamples. Of the 5 subsamples, one subsample (namely 20% of the total sample) was retained as the validation data and the remaining subsamples were used as the training data. Then we repeated this procedure five times and took the average results of the 5 repetitions as our final estimate. The final verification results showed that the models did not appear overfitting, indicating that they were appropriate to be used to predict unknown data.

3.7. *Metrics of prediction performance*

In the machine learning literature, researchers often use Precision, Recall, Accuracy and F1-score as the indicators of modeling performance in classification experiments, and use Mean Squared Error (MSE), Mean Absolute Error (MAE) and R^2 score in regression experiments. In this experiment, we used accuracy to compare the performance of each model in the classification experiment while, in the regression experiment, we used the MSE as the basis of model comparison.

Accuracy measures how accurately all samples are classified. The higher the accuracy, the more accurate the classification, and the better the model performance. The confusion matrix, also called the error matrix, is a standard format for accuracy evaluation. It is expressed in a matrix with n rows and n columns. The confusion matrix and the corresponding formula used here are shown below:

Table 2
Confusion matrix

Algorithms	Predicted value = 1	Predicted value = 0
True value = 1	True Positive (TP)	False negative (FN)
True value = 0	False Positive (FP)	True negative (TN)

$$A \quad \frac{\mathrm{TP} \quad \mathrm{TN}}{\mathrm{TP} \quad \mathrm{FP} \quad \mathrm{TN} \quad \mathrm{FN}}$$

In the above formula, A represents accuracy.

The MSE is the expected value of the square of the difference between the estimated value and the true value of a parameter, which can be expressed as follows:

$$\mathrm{MSE}(y,\hat{y}) \quad \frac{1}{n}{}_{i\ 0}^{n\ 1}(y_i \quad \hat{y})^2$$

In the above formula, y represents the real value, $\hat{y}$ represents the predicted value and n represents the sample size. The MSE can evaluate the degree of change in the data. The smaller the MSE, the better the performance of the prediction model in describing the experimental data.

4. Experiment Results

13 predictors of health status were obtained in the feature selection step of the experiment: (1) weight, (2) medical expenditure, (3) residential area, (4) the current price of the house, (5) purchase, repair and spare parts costs for various vehicles (such as bicycles and electric bicycles, but excluding cars), (6) communication tools (such as telephones, mobile phones, etc.), (7) water and electricity expenses in the past month, (8) how much the family had spent on food (excluding eating out) in the past week, (9) whether there is a bath or shower in the house, (10) post, telecommunications, and communications expenses (including telephones, mobiles, the Internet, post, etc.), (11) education and training expenses (including tuition fees, training fees, etc.), (12) clothing consumption, expenditure on furniture, consumer durables and electrical appliances (including refrigerators, washing machines, televisions, computers and high-end instruments such as pianos), and (13) how much the family had spent on smoking and drinks in the past week.

The 13 predictors of the ADL and IADL limitations selected based on the MIC criteria are: (1) having health care insurance, (2) the amount of money received in the form of pension and social benefits, (3) housing tenure, (4) the amount of money received in the form of public housing fund, (5) whether or not in retirement, (6) the level of the out-of-pocket premium of private health care insurance, (7) marital status, (8) inheriting money, (9) inheriting housing properties, (10) type of social benefits, (11) amount of tax or social insurance paid by other household members, (12) number of hours worked per day on average in the last year, and (13) Total area of housing land.

It can be noted that many non-linear predictors of health conditions selected by the algorithms are in the category of socioeconomics status (SES). For self-reported health, important SES predictors often concern people's living expenses, whereas SES predictors of ADL and IADL limitations mainly relate to income and asset.

After obtaining these characteristics, Artificial Neural Network, Logistic Regression, Support Vector Machine, XGBoost Classifier, and Random Forest Classifier were used for the classification of self-reported health status and prediction of functional limitations. The results of the classification are shown in Table 3.

Table 3
Classification result (self-reported health status)

Algorithms	Accuracy
Artificial Neural Network	0.699
Logistic Regression	0.672
Support Vector Machine	0.672
XGBoost Classifier	0.635
Random Forest Classifier	0.606

Table 4
Regression result (ADL and IADL limitations)

Algorithms	**mean squared error, MSE**
Linear Support Vector Regression	0.0006
Linear Regression	0.0010
Random Forest Regressor	0.0109
XGBoost Regressor	0.0516
Artificial Neural Network	0.0673

Table 5
Regression result (self-reported health status)

Algorithms	**mean squared error, MSE**
Linear Regression	0.0630
Random Forest Regressor	0.0641
Linear Support Vector Regression	0.1009
XGBoost Regressor	0.1673
Artificial Neural Network	0.2049

When comparing the classification results, the ANN performed best. It has the highest accuracy of 69.9%. With an accuracy of 60.6%, the random forest classifier has the worst prediction performance. The logistic regression model has reached an accuracy of 67.2%. It is not the best model, but it is not substantially worse than the ANN classifier either.

Artificial Neural Network, Linear Regression, Linear Support Vector Regression, XGBoost Regressor, and Random Forest Regressor were used for the regression experiment of functional limitations and self-reported health. The regression results are shown in Tables 4 and 5.

For the regression results of different models, the MSE of the ANN is always the largest among these models, and the MSE of some of the linear models is smaller. For the prediction of functional limitations, the support vector regression with a linear kernel has the lowest MSE. It seems that the ANN is not always the most suitable model to capture the nonlinear relationships in the data despite its highly flexible modeling architecture.

5. Discussion

As the experiment showed, the newly established model is useful to predict the health status of older people, mainly due to the extraction of non-linear features and the establishment of a non-linear model in this experiment.

Traditional methods, like those based on Pearson measures, cannot capture the non-linear correlation between two variables. The traditional method of manually selecting features is based on a large amount of prior knowledge and domain experience and tends to select features with a linear correlation. Therefore, some latent and non-linear features are often ignored and excluded from the modeling. However, these non-linear features can play a crucial role in predicting the target variables. With MIC feature selection, features with a high linear correlation or non-linear correlation can be found more easily and the result shows that such a feature selection method found certain non-linear features (with low Pearson scores), which were hidden in the analysis of the linear features. Including these features in the modeling can improve the predictive model. With non-linear correlated features, it is necessary to adopt a non-linear model to capture the non-linear relationships between the features and the target variables. The deep neural network is a powerful non-linear machine learning model. The non-linearity is provided by the activation functions on a large number of neurons, while the depth of the network provides an elastic capability for capturing complex relations, compared with conventional methods like the random forest.

To build an accurate predictive model, it is important to consider the following issues: when a feature is selected, it should make a sound contribution in the prediction process, not necessarily because it has a causal relationship with the target variable. In other words, finding the correlation between the features and target variables is more important than causality here. Therefore, the result of this paper cannot be used for intervention: changing a feature will not necessarily have any effect on the target variables. In some cases, they have causality relationships but, in others, they are related to some hidden variables or related in a more complex way. Nevertheless, this does not mean that it is futile to use machine learning methods to solve these problems. In contrast, an accurate predictive model is useful, e.g., for planning resource allocation, etc. Also, these features selected provide a novel starting point for further analysis: why are they related, especially when this relationship is non-linear? Therefore, machine learning methods can help social scientists to discover previously ignored hidden factors and form a new hypothesis based on these, which is a new paradigm of data-driven social science research.

The application of our methods relies on its predictive capability. Some applicable scenarios include the prediction of various social indicators, resource pre-allocation, and promoting the development of other related industries.

The accurate prediction of disability can further be used to predict various social indicators, such as life expectancy, the expense of health insurance, health care costs, etc. Understanding the future requirements of health care services and pensions can help to pre-allocate the necessary resources. For example, according to the regional imbalances in the distribution of disability among the elderly, the construction of a pension infrastructure can be promoted more specifically. Analyzing the factors affecting the health of the elderly can promote the development of the elderly service industry, such as the types of services provided by caregivers, the distribution and size of nursing homes, etc., which can more effectively meet the needs of older people.

Based on the predictive model, with the further development of causality analysis (based on methods like the Bayesian network), we can even model an intervention model.

Our method has more powerful data processing capabilities than traditional methods. The most significant advantage of using machine learning methods is the capability quickly to process large, complex data sets. By 2020, it is estimated that, for every person on earth, 1.7MB of data will be created every second, which gives researchers a richer sample to mine the hidden relationships but also creates higher requirements regarding researchers' ability to process data. In this context, although the method adopted in this paper cannot fully deal with big data, it can quickly complete the preprocessing, feature extraction, training models, and predictions for a medium amount of data. Compared with the traditional methods, machine learning methods are faster, more accurate at prediction, and have better performance.

Another contribution of our method is its generalization capability. Since our feature selection approach does not require any prior knowledge, it can be applied to a wide range of domains. Moreover, the modeling method is not domain-specific: a deep neural network is a general-purpose model. Although it needs to adjust its hyper-parameters in different application environments, most of the training processes are performed automatically and can adapt to different domains. The main modeling work in this paper was mostly carried out by computer scientists with limited domain knowledge, and the workflow used can be easily transferred to a new problem of finding non-linear features for a target variable. Therefore, the methodology used here can be further developed to suit a new general data-driven paradigm, which uses big data to find appropriate features and build an accurate predictive model.

Despite the contribution of the non-linear feature selection algorithms and the advantage of non-parametric machine learning models in terms of prediction performance, the limitations of these models should also be acknowledged. It is well-established that there is a tradeoff between prediction accuracy and model explainability. Models such as the ANN or SVM may perform better than linear regression models in terms of prediction accuracy, but they are also less interpretable. There has been an increasing interest in the development of machine learning models with higher explainability in recent studies (Gilpin et al., 2018; Christopher, 2019). However, since the main objective of our experiment is to identify the methods and procedures which optimize the feature selection process and facilitate prediction, we have not provided a detailed discussion about the explainability of the predictors. Such an investigation would merit a separate study in the future.

6. Conclusion

In this paper, several machine learning models are used to predict the health status of the elderly in recent years. By comparing the experimental results, the artificial neural network is considered to be the best performing model because it displayed the highest accuracy in the classification experiments. This also proves that the artificial neural network used to predict the health status of elderly people is reliable.

Machine learning methods differ from the traditional methods used in social science. The former’s advantages include two aspects: on the one hand, machine learning methods can capture non-linear features. The previous traditional methods of analysis in the social sciences, such as multiple linear regression, cannot discover non-linear features and, in some cases, perhaps the impact of non-linear features is the most critical factor; on the other hand, the feature selection project is built around the data and the results are directly calculated from the data. Therefore, it saves a lot of time compared to manually checking the characteristics based on the literature review. It can also avoid the interference of prior knowledge and bring a new perspective and way of thinking.

In summary, the entire workflow of the machine learning methods can reduce the impact of experience and lead to a greater reliance on the data to obtain accurate predictions. It can be used to build data-driven, general paradigm research that can be migrated to any domain. However, this does not mean that the traditional knowledge of social sciences has become redundant. Combining machine learning technology with traditional methods is the future direction.

Funding statement

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Declaration of Competing Interest

No conflict of interest has been declared by the authors.

Author contributions

All authors contributed to the drafting or revisions of the manuscript and have read and approved the final version of this submitted manuscript. Fy. Qin and Zq. Lv contribute equally to this work.

References

Aldaz, J. M., Barza, S., Fujii, M., & Moslehian, M. S. (2015). Advances in Operator Cauchy-Schwarz inequalities and their reverses, *Annals of Functional Analysis, 6*(3): 275-295. https://doi.org/10.15352/afa/06-3-20.

Artaud, F., Dugravot, A., Sabia, S., et al. (2013). Unhealthy behaviors and disability in older adults: Three-City Dijon cohort study. *BMJ, 347*: f4240. https://doi.org/10.1136/bmj.f4240.

Avlund, K., Vass, M., Lund, R., Yamada, Y., & Hendriksen, C. (2008). Influence of psychological characteristics and social relations on receiving preventative home visits in older men and women. European Journal of Ageing, 5(3), 191-201. https://doi.org/10.1007/s10433-008-0086-4.

Benfante, R., Reed, D., & Brody, J. (1985). Biological and social predictors of health in an aging cohort. *Journal of Chronic Diseases, 38*(5): 0-395.

Bonsang, E. (2009). Does informal care from children to their old people parents substitute for formal care in Europe? *Journal of Health Economics*, 28, 143–154. https://doi.org/10.1016/j.jhealeco.2008.09.002.

Bookman, A., Harrington, M., Pass, L., & Reisner, E. (2007). *Family Caregiver Handbook*. Cambridge, MA: Massachusetts Institute of Technology.

Boult, C., Kane, R. L., Louis, T. A., & McCaffrey, D. (1994). Chronic conditions that lead to functional limitation in the elderly. *Journal of Gerontology, 49*: M28-M36.

Breiman, L. (2001). *Random Forests. Machine Learning 45*(1), 5-32.

Casanova, R., Saldana, S., Lutz, M. W., Plassman, B. L., Kuchibhatla, M., & Hayden, K. M. (2017). Investigating predictors of cognitive decline using machine learning. *Alzheimer's & Dementia, 13*(7), 876.

Chakravarty, E. F., Hubert, H. B., Krishnan, E., et al. (2012). Lifestyle Risk Factors Predict Disability and Death in Healthy Aging Adults. *The American Journal of Medicine, 125*(2): 0-197.

Chen, T. Q. , & Carlos, G. (2016, March). XGBoost: A Scalable Tree Boosting System. the 22nd ACM SIGKDD International Conference. San Francisco, USA. https://doi.org/10.1145/2939672.2939785.

Christoph, M. (2019). "Interpretable machine learning. A Guide for Making Black Box Models Explainable". https://christophm.github.io/interpretable-ml-book/.

Cortes, C., & Vapnik, V. (1995). *Support-Vector Networks. Machine Learning, 20*, 273-297.

Facal, D., Valladares‐Rodriguez, S., Lojo‐Seoane, C., Pereiro, A. X., Anido‐Rifon, L., & Juncos‐Rabadan, O. (2019). Machine learning approaches to studying the role of cognitive reserve in conversion from mild cognitive impairment to dementia. *The International Journal of Geriatric Psychiatry, 34*(7): 941-949. https://doi.org/10.1002/gps.5090.

Dunlop, D., Hughes, S. L., & Manheim, L. M. (1997). Dis-ability in activities of daily living: Patterns of change and hierarchy of disability. *American Journal of Public Health*, 87: 378–383.

Fried, L. P., Herdman, S. J., Kuhn, K. E., et al. (1991). Preclinical Disability: Hypotheses About the Bottom of the Iceberg. *Journal of Aging and Health, 3*(2): 285-300.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. Paper presented at the *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. Turin, Italy, pp. 80-89.

Greenglass, E., Fiksenbaum, L., & Eaton, J. (2006). The relationship between coping, social support, functional disability and depression in the old people. *Anxiety, Stress & Coping, 19*(1): 15-31.

Hebert, R, Brayne, C., & Spiegelhalter, D. (1997). Incidence of functional decline and improvement in a community-dwelling, very old people population. *American Journal of Epidemiology*, 145: 935–944.

Hosmer, D. W., & Lemeshow, S. (2000). Applied logistic regression. Wiley, New York. https://doi.org/10.1002/0471722146.

Idler, E. L., & Benyamini, Y. (1997). Self-rated health and mortality: a review of twenty-seven community studies. *Journal of Health and Social Behaviour*, 38:21-37.

Jimenez, M., Chiuve, S. E., Glynn, R. J., Stampfer, M. J., Camargo Jr, C. A., Willett, W. C., et al. (2012). Alcohol consumption and risk of stroke in women. *Stroke, 43*: 939-45.

Krijger, K. , Schoofs, J. , Marchal, Y. , Van De Vijver, E. , Borgermans, L. , & Dervroey, D. (2014). Association of objective health factors with self-reported health. *Journal of preventive medicine and hygiene, 55*: 101-107. https://doi.org/10.1163/9789004278585_011.

Li, C. C., Meng, X. H., Wang, J. R., Ma, H. J., Chen, C., & Liu, Y. Q. (2017). Association between sociodemographic, psychosocial, lifestyle factors, and self-reported health among migrant laborers in China. *Journal of the Chinese Medical Association* 80 204-211. https://doi.org/10.1016/j.jcma.2016.10.011.

Oscar, N., Fox, P. A., Croucher, R., Wernick, R., Keune, J., & Hooker, K. (2017). Machine Learning, Sentiment Analysis, and Tweets: An Examination of Alzheimer's Disease Stigma on Twitter. *The Journals of Gerontology Series B Psychological Sciences and Social Sciences, 72*(5), 742–751. https://doi.org/10.1093/geronb/gbx014.

Nummela, O. P., Sulander, T. T., Heinonen, H. S., & Uutela, A. K. (2007). Self-rated health and indicators of SES among the ageing in three types of communities. *Scand Journal of Public Health*; 35:39-47. https://doi.org/10.1080/14034940600813206.

Oman, D., Reed, D., & Ferrara, D. (1999). Do old people women have more physical disability than men do? *American Journal of Epidemiology*, 150: 834–842.

O'Donnell, O., Van Doorslaer, E., Wagstaff, A., & Lindelow, M. . (2008). Analyzing health equity using household survey data: a guide to techniques and their implementation. *world bank, 86*(10), 816-816.

Parker, M., Thorslund, M., & Lundberg, O. (1994). Physical function and social class among Swedish oldest old. *Journals of Gerontology, 49*(4). https://doi.org/10.1093/geronj/49.4.S196.

Shariful, I., Ismail, T., Nazrul, I. M., Ahbab, M. F. R., Hafiz, T. A. K. , & Sharifa, B. (2017). Urban-rural differences in disability-free life expectancy in Bangladesh using the 2010 HIES data. *PLoS One, 12*(7): e0179987. https://doi.org/10.1371/journal.pone.0179987.

Smola, A. J. , & Bernhard Schölkopf. (2004). A tutorial on support vector regression. *Statistics and Computing, 14*(3), 199-222.

Suykens, J. A. K. , & Vandewalle, J.. (1999). Least squares support vector machine classifiers. *Neural Processing Letters, 9*(3), 293-300.

Szmitko, P. E., & Verma, S. (2005). Antiatherogenic potential of red wine. clinician update. *American Journal of Physiology-Heart and Circulatory Physiology, 288*: H2023-30.

Tak, E., Kuiper, R., Chorus, A., et al. (2013). Prevention of onset and progression of basic ADL disability by physical activity in community dwelling older adults: A meta-analysis. *Ageing Research Reviews, 12*(1): 329-338.

Ties, B., Ahmad, R. H., Emese, V., & Somnath, C. (2016). A global assessment of the gender gap in self-reported health with survey data from 59 countries. *BMC Public Health*. https://doi.org/10.1186/s12889-016-3352-y.

Tsai, J., Ford, E. S., Li, C., Pearson, W. S., & Zhao, G. (2010). Binge drinking and suboptimal self-rated health among adult drinkers. *Alcohol Clin Exp Res, 34*: 1465-71.

Williams, B. (2014). Consideration of Function & Functional Decline. *Current Diagnosis and Treatment: Geriatrics, Second Edition*. New York, NY: McGraw-Hill. pp. 3–4. ISBN 978-0-07-179208-0.

Williams, C. (2011). *CURRENT Diagnosis & Treatment in Family Medicine, 3e. Healthy Aging & Assessing Older Adults*. New York, NY: McGraw-Hill, (*Chapter 39*).

Zell, A. (2003). *Simulation neuronaler Netze [Simulation of Neural Networks]*. (1st ed.). German, (Chapter 5.2).