GeroScience (2024) 46:737–750
https://doi.org/10.1007/s11357-023-01040-9

ORIGINAL ARTICLE

Comparative study of machine learning methods for modeling associations between risk factors and future dementia cases

Vaka Valsdóttir · María K. Jónsdóttir · Brynja Björk Magnúsdóttir · Milan Chang · Yi-Han Hu · Vilmundur Gudnason · Lenore J. Launer · Hlynur Stefánsson

Received: 16 September 2023 / Accepted: 7 December 2023 / Published online: 23 December 2023

Abstract A substantial portion of dementia risk can be attributed to modifiable risk factors that can be affected by lifestyle changes. Identifying the contributors to dementia risk could prove valuable. Recently, machine learning methods have been increasingly applied to healthcare data. Several studies have attempted to predict dementia progression by using such techniques. This study aimed to compare the performance of different machine-learning methods in modeling associations between known cognitive risk factors and future dementia cases. A subset of the AGES-Reykjavik Study dataset was analyzed using three machine-learning methods: logistic regression, random forest, and neural networks. Data were collected twice, approximately five years apart. The dataset included information from 1,491 older adults who underwent a cognitive screening process and were considered to have healthy cognition at baseline. Cognitive risk factors included in the models were based on demographics, MRI data, and other health-related data. At follow-up, participants were re-evaluated for dementia using the same cognitive screening process. Various performance metrics for all three machine learning algorithms were assessed. The study results indicate that a random forest algorithm performed better than neural networks and logistic regression in predicting the association between cognitive risk factors and dementia. Compared to more traditional statistical analyses, machine-learning methods have the potential to provide more accurate predictions about which individuals are more likely to develop dementia than others.

V. Valsdóttir (✉) · M. K. Jónsdóttir · B. B. Magnúsdóttir
Department of Psychology, Reykjavik University,
Reykjavik, Iceland
e-mail: vaka@ru.is

V. Valsdóttir · M. Chang
RHLÖ – Icelandic Gerontological Research Center,
Landspítali University Hospital, Reykjavik, Iceland

M. K. Jónsdóttir · B. B. Magnúsdóttir
Mental Health Services, Landspitali University Hospital,
Reykjavik, Iceland

Y.-H. Hu · L. J. Launer
Laboratory of Epidemiology and Population Sciences,
National Institute On Aging, National Institutes of Health
(NIH), Bethesda, MD, USA

V. Gudnason
Faculty of Medicine, University of Iceland, Reykjavik,
Iceland

V. Gudnason
The Icelandic Heart Association, Kopavogur, Iceland

H. Stefánsson
Department of Engineering, Reykjavik University,
Reykjavik, Iceland

Keywords Cognitive aging · Model performance · AGES-Reykjavik Study · Machine learning · Random forest · Cognitive risk factors

Introduction

Many believe that cognitive decline is an inevitable part of aging and that developing dementia is a normal part of cognitive aging, which we as individuals have no control over [1, 2]. Research has, however, shown that some individuals maintain their cognitive abilities even though they reach a very high age and that individuals can influence their cognitive aging process [1–4]. In fact, while it is known that both age and genetics are strong risk factors for cognitive decline [5, 6], a recent report on dementia prevention, intervention, and care states that modifiable risk factors contribute to up to 40% of dementia cases worldwide [3]. The risk factors that are highlighted in the report and considered to have the strongest evidence are low education level, hearing loss, traumatic brain injury, hypertension, excessive alcohol consumption, obesity, smoking, depression, social isolation, physical inactivity, diabetes, and air pollution. Other factors, such as diet [7, 8], multilingualism [9], sleep [10–12], and physical fitness [13–17], have also been shown to have an association with dementia and cognitive decline. All of these risk factors are modifiable to some extent. Consequentially, by targeting these known risk factors for dementia and cognitive decline in a systematic way, a substantial proportion of dementia cases could be prevented or delayed. Pinpointing the individuals who could benefit from modifying their lifestyle to improve their habits relating to those risk factors could be of great value. One approach that could improve the ability to identify these individuals would be to apply mathematical models to the available data.

Machine learning algorithms have increasingly been applied to healthcare data [18, 19]. Studies have shown that machine-learning algorithms have proven helpful in risk prediction, disease diagnosis, and assessment of disease severity [18, 20]. Many of these studies have involved data collected from older adults. One example is a study that applied random forest algorithms to examine how sleep predicted mortality in older adults compared to other known predictors of mortality [21]. The results showed that although sleep was not the strongest predictor of mortality, multidimensional sleep data can add to the predictive power of more routinely used predictors of mortality.

Another example is a study that compared the performance of different machine-learning methods in modeling social determinants of health using the Health and Retirement Study dataset [22]. Of the algorithms that were explored, neural networks performed best. The study's results suggested that interactive non-linear relationships between social factors and biological health indicators were identified by applying neural networks to the dataset.

Several machine-learning studies have been performed on data relating to dementia, with the choice of method often being determined by the characteristics of the dataset [23]. Spooner et al. [24] compared the performance of different machine learning methods in predicting the onset of dementia using high-dimensional clinical data from two different datasets. According to their findings, random forest algorithms can outperform linear regression models when non-linear data, with complex non-linear relationships, is present, which is the case with many datasets. Casanova et al. [25] used random forest classification to study predictors of cognitive trajectories. According to their results, nongenetic modifiable risk factors play an important role in the trajectory of cognitive decline. Qiu et al. [26] used a deep learning framework to classify individuals with Alzheimer's disease using MRI data and cognitive scores. The model's diagnostic performance was better than that of a team of neurologists. Garcia-Gutierrez et al. [27] assessed the ability of machine learning algorithms to classify individuals with and without dementia based on neuropsychological assessments. The algorithms performed well, and the results indicated that by applying such algorithms, it might be possible to rely on the results of fewer neuropsychological tests than specialists usually do, reducing the resources needed to perform neuropsychological examinations. The focus of these studies differs, but all of them indicate that useful information about cognition and dementia can be obtained by applying machine learning methods to datasets that contain relevant information.

Present study

The work presented in this paper modeled associations between risk factors and future dementia cases using machine learning methods. Predicting which individuals are more likely to develop dementia can provide valuable insight for healthcare professionals and the healthcare system. By creating more accurate models, costs can be reduced, and healthcare workers can provide care in a more targeted way. Individuals at risk could be identified sooner in the process, and dementia progression could possibly be delayed, improving the quality of life for those individuals. A previous analysis of the AGES-Reykjavik Study dataset using a traditional statistical approach has shown that a logistic regression model analyzed in SPSS can pinpoint risk factors associated with whether an individual will have dementia five years later [28]. This study aimed to take that assessment further and explore whether machine learning algorithms could better estimate future dementia cases based on known risk factors for dementia and cognitive decline.

Methods

Study Sample

The dataset was collected through a study called the Age Gene/Environment Susceptibility-Reykjavik Study (AGES-Reykjavik Study). The participants were recruited through previous participation in the Reykjavik Study, a study sponsored by the Icelandic Heart Association (IHA). A randomly selected subgroup of the participants in the Reykjavik Study, all of whom lived in the Reykjavik area of Iceland at the time of the study and were born between 1907 and 1935, were invited to continue and participate in the AGES-Reykjavik Study. Between 2002 and 2011, data were collected twice for each participant, with approximately five years between baseline and follow-up. The purpose of the AGES-Reykjavik Study was to evaluate risk factors for disability and disease within the older population. Data collected included results from cognitive tests and dementia assessments and information regarding risk factors that could have associations with cognitive performance. Overall, 5,764 participants took part in the baseline data collection of the study. Of those, 3,316 also participated in the follow-up data collection. A more detailed description of the study design can be found in previously published material [29, 30].

The dataset used for all analyses in this study is based on the dataset used in a previous paper where logistic regression was used to assess the relationship between risk factors for cognitive aging and dementia based on traditional statistical methods [28]. It should be noted that a substantial portion of the original AGES-Reykjavik dataset was excluded from the analysis. That is because many participants were excluded since they had missing data for any of the variables included in the logistic regression model analyzed using SPSS. Those who fulfilled the dementia criterion at the baseline measurement were not included in the final dataset since the purpose was to assess the likelihood of developing dementia later. Figure 1 shows a flowchart of participants included and excluded from the analysis. Of those participating in the follow-up data collection round, 16.2% (n = 536) fulfilled a dementia criterion when assessed at baseline. The proportion of participants included in the data analysis (n = 1,491) that fulfilled the criterion at follow-up was 8.2% (n = 123). Since the previous dataset analysis excluded participants with missing data for any of the variables in a model, a substantial number of participants were not included in the final dataset. For a more detailed description of the analytical sample, please refer to [28].

The Institutional Review Board of the U.S. National Institute on Aging, the National Institutes of Health, and the Icelandic National Bioethics Committee (VSN 00–063) have approved the study. All participants provided written consent for participation in the study.

Exposure variables

The choice of variables for the analysis was based on previous work that analyzed the same data [28, 30], which was supported by the cognitive aging literature [3, 9, 15, 25, 31–37]. The dataset used for the analyses performed here included eighteen exposure variables known as predictors of cognitive aging. Information about each variable is displayed in Table 1. Age and sex were included as control variables since both are recognized as predictors of cognitive aging [38, 39].

Dementia criterion

Participants were evaluated for mild cognitive impairment and dementia at baseline and follow-up. The screening process was the same in both instances, and the same panel of professionals was involved [29]. The three-step process started with a cognitive screening using the Mini-Mental State Examination

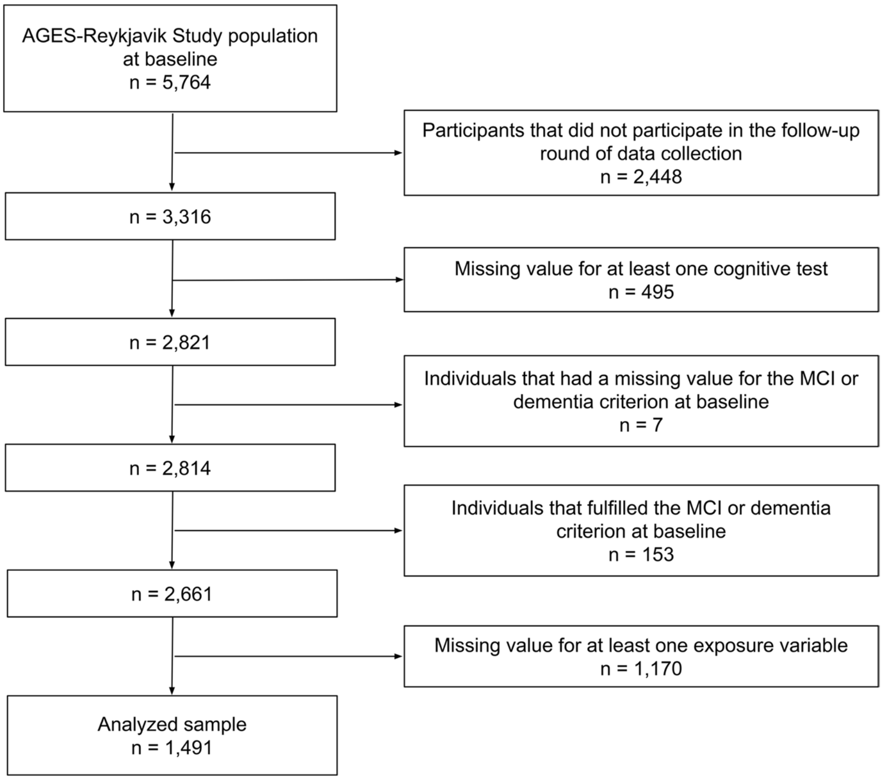

AGES-Reykjavik Study population at baseline
n = 5,764
Participants that did not participate in the follow-up round of data collection
n = 2,448
n = 3,316
Missing value for at least one cognitive test
n = 495
n = 2,821
Individuals that had a missing value for the MCI or dementia criterion at baseline
n = 7
n = 2,814
Individuals that fulfilled the MCI or dementia criterion at baseline
n = 153
n = 2,661
Missing value for at least one exposure variable
n = 1,170
Analyzed sample
n = 1,491

Fig. 1 Flowchart showing the participant selection for the study (the figure is a reproduction of work from a previous publication [28])

[42] and the Digit Symbol Substitution Test [43]. Next, a neuropsychological diagnostic battery was administered to participants with positive screening results. Based on two screening tests from the diagnostic neuropsychological test battery, participants were identified for further examination by a neurologist. In the third and final step, based on relevant cognitive, MRI, and health history data, a consensus diagnosis was made by a panel of relevant specialists (neuropsychologist, neurologist, geriatrician, and neuroradiologist) in line with international guidelines. For the purpose of this study, individuals considered by the panel to have either mild cognitive impairment or dementia were grouped together; therefore, all of them fulfilled the current study's dementia criterion.

Machine learning

Method of analysis

Three classification methods were compared to assess which could provide the most accurate prediction of future dementia cases based on known risk factors for cognitive aging. The scikit-learn library within the Python programming language was used for data analysis in all three cases [44]. The methods compared were logistic regression, random forest, and neural network.

Logistic regression Within traditional statistics, logistic regression is a method commonly applied to predict binary outcomes by analyzing the relationship between multiple independent variables [45]. The use of logistic regression has also been adopted by those who work on machine learning, and within that discipline, logistic regression is often used to classify incoming information based on existing data.[1]

Random forest Random forest is a machine-learning method that builds upon the idea of decision trees by combining in one model a forest of decision trees that work together [46]. Figure 2 shows an example of

[1] The work performed for this study falls under the machine learning approach to logistic regression, but there is, however, also a reference to a previous publication [28] that applies logistic regression using traditional statistics.

Table 1 Description of exposure variables

Sex	Male, female
Age	Subject age in years at first measurement
Mobility	Timed up and go test (TUG) measured in seconds [a]. A lower score represents better mobility
Leisure activities	Average of days per month engaged, reported for mental and social leisure activities (movies, lectures, church, crossword puzzles, board/card games, and computer games)
Foreign languages	Number of foreign languages spoken
Education	Education level completed: Primary school, secondary, college, university (Primary and secondary school were combined into one group)
Self-reported health	Self-estimation of general health: Excellent, very good, good, fair, poor (Two categories: 1 – poor and fair, 2 – good to excellent)
Physical strength	Maximum strength value in leg in Newtons
Smoking	Smoking status: Never smoked, previous smoker, current smoker
Coronary artery disease	Coronary artery disease diagnosis based on rose angina, MI ECG, and use of nitrates: Yes, possible case, no
Alcohol consumption	Grams of alcohol per week consumed
Depression	Geriatric Depression Scale score[b]
Diabetes	Diagnosed as diabetes by self-report, fasting glucose, or medication use: Yes, no
Hypertension	Hypertension, derived from physiological measurements (systolic blood pressure, diastolic blood pressure): Yes, pre-hypertension, no
Body Fat Percentage	Bioelectric Impedance (BIA): Percent body fat
ApoEε4carrier	Apolipoprotein E (ApoE) genotype positive carrier: Yes, no
Relative grey matter volume	Grey matter volume (ml) divided with intracranial volume (ICV[c])
Relative white matter volume	White matter volume (ml) divided with intracranial volume (ICV[c])

The table is an adapted version of Table 1 from Valsdóttir et al. (2022)

[a] [40]. [b] [41]. [c] The sum of grey matter volume, white matter volume, white matter lesion volume, and cerebral spinal fluid volume

how a random forest algorithm operates. Each decision tree within the model gives a prediction, and the results of the majority of the trees become the result the random forest algorithm gives. The purpose of the algorithm is to create many uncorrelated decision trees that can give more accurate results than any one decision tree could. The number of trees in the random forest and the complexity of the trees can be changed to test which combination gives the best results. The application of decision trees was also explored. However, since the algorithm's performance was underwhelming compared to the other three algorithms, the decision tree analysis will not be discussed further.

Using scikit-learn, data on feature importance can be extracted when a random forest algorithm is used, indicating which variables in the model were most important to the model's predictive power. A higher value indicates a higher importance.

Neural network Neural network is a machine learning algorithm modeled after how neurons in the brain communicate, allowing the algorithm to recognize patterns and relationships within datasets [47]. Figure 3 shows an example of how a neural network algorithm functions. The algorithm is based on multiple layers, an input layer and an output layer, and in between, there are hidden layers. The number of hidden layers can vary, and the number of nodes within each layer can also be changed. A node is a unit that operates similarly to a neuron; while training the network, a node is either activated or remains inactive. Different combinations of the number of hidden layers and nodes can be tested to identify which provides the most accurate predictions.

Data pre-processing

Validation strategy The dataset was split into training and test datasets to estimate the performance

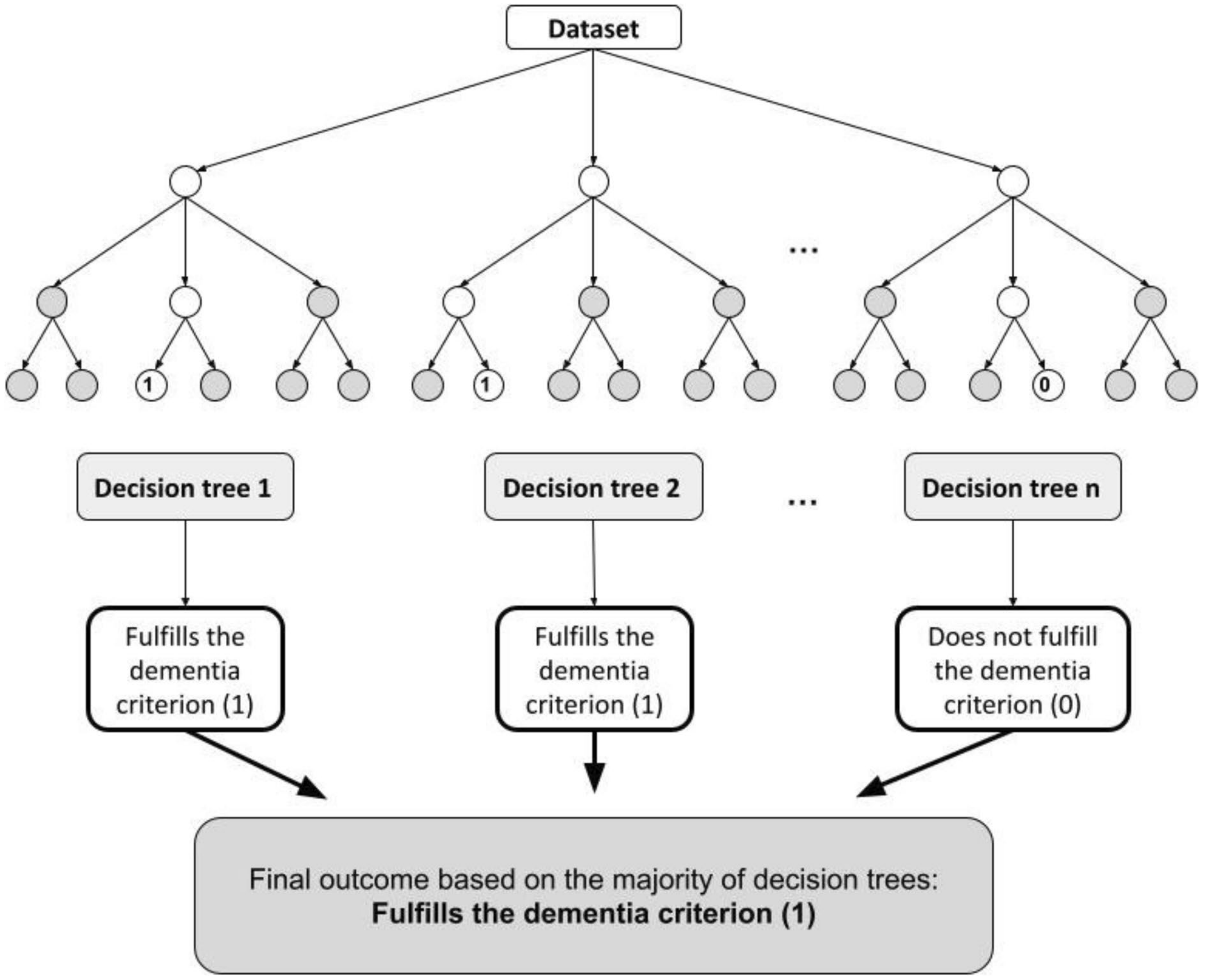
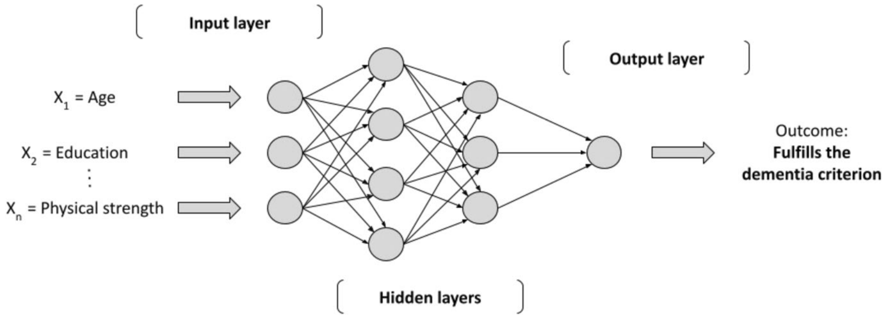

Dataset
1
1
0
Decision tree 1
Decision tree 2
Decision tree n
Fulfills the dementia criterion (1)
Fulfills the dementia criterion (1)
Does not fulfill the dementia criterion (0)
Final outcome based on the majority of decision trees:
Fulfills the dementia criterion (1)

Fig. 2 An example that explains a random forest algorithm

Fig. 3 An example that explains a neural network algorithm



of the different machine learning algorithms. The training dataset's size was 67% of the whole dataset; 493 participants were allocated to the test dataset and 998 to the training dataset. A statistical comparison was performed on the two datasets, and only one variable, diabetes, was significantly different ($p < 0.001$). Within the training dataset, 7.3% of the participants had diabetes, while 12.8% of participants within the test dataset had diabetes. Despite this anomaly, the comparison suggests that the two datasets are very similar.

Data scaling Some pre-processing was performed to prepare the data for the machine learning algorithms. Since the two groups within the outcome variable are unbalanced, features were standardized by removing the mean and then scaling to unit variance.

Performance criteria

Confusion matrix A confusion matrix gives valuable information about the model's performance [48]. The matrix displays which data points are classified correctly and which are classified incorrectly by showing true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Simple metrics calculated based on information from the confusion matrix can give a high-level idea of model performance on specific tasks. These include accuracy, precision and recall.

Accuracy The accuracy of a model is generally calculated by dividing the number of individuals correctly classified by the number of participants included in the dataset.

Precision Precision is calculated by dividing the number of true positives by the number of true positives plus the number of false positives. This value indicates the portion of correct positive predictions, showing the proportion of people classified into the dementia group that actually have dementia.

Recall Recall is calculated by dividing the number of true positives by the number of true positives plus the number of false negatives. This value indicates the portion of correct positive predictions out of all positive cases, giving insight into the proportion of individuals with dementia that the models do not detect.

Balanced accuracy When outcome variables are imbalanced, balanced accuracy is considered a better metric [49]. This metric considers that one of the classes has a much lower number of individuals that belong to it, giving that class more weight in the accuracy calculations.

Table 2 Misclassification rate for both training and test datasets for all three machine learning methods

	Misclassification rate		
	Training dataset	Test dataset	Difference
Logistic regression	8.1%	9.1%	1.0%
Random forest	6.9%	8.3%	1.4%
Neural network	7.7%	8.5%	0.8%

F1 score The F1 score focuses on incorrect classified results (false positives and false negatives), balancing precision and recall [50]. The higher the value, the better the performance of the model is.

ROC curve A receiver operating characteristic (ROC) curve is a graph that plots the true positive rate of a model against the false positive rate at different classification thresholds [51]. The area under the ROC curve (AUC-ROC), which gives a number between zero and one, indicates how well the classifier performs across all possible thresholds. The higher the number, the better performance is.

Matthews correlation coefficient The Matthews correlation coefficient (MCC) calculates the correlation between true and predicted values in a dataset [52]. The MCC takes a value between minus one and one, with one indicating perfect performance and minus one indicating that each item has been misclassified.

Overfitting

Each model's misclassification rate for the training and test datasets was compared to identify whether a model was overfitting, thereby creating a generalization error. If the misclassification rate for the test dataset was much higher than the training dataset, the model was overfitting, thereby not performing as desired. When tuning the models, the misclassification rate was continually assessed to ensure the model was not overfitting. Table 2 shows the misclassification rate of the final models for each method. Based on an analysis of trends in the misclassification rate for the training and test datasets, a difference of more than 2% was considered to suggest that a model was

Table 3 Descriptive statistics for exposure variables (measured at baseline) showing means, standard deviations, counts and percentages

	All participants (n = 1,491)		Healthy cognition at follow-up (n = 1,368)		Fulfil dementia criterion at follow-up (n = 123)	
	M	*SD*	*M*	*SD*	*M*	*SD*
Age, mean	74.11	4.41	73.85	4.30	77.10	4.45
Mobility	11.23	2.51	11.11	2.36	12.46	3.60
Leisure activity	5.79	3.70	5.92	3.72	4.43	3.10
Foreign languages	2.17	1.50	2.23	1.50	1.50	1.22
Physical strength	347.08	115.03	349.13	115.98	324.28	101.54
Alcohol consumption	17.82	34.07	17.92	33.68	16.71	38.23
Depression	1.88	1.89	1.83	1.88	2.43	1.91
Body fat percentage	29.16	8.14	29.36	8.11	26.95	8.22
Relative grey matter volume	46.02	3.13	46.14	3.12	44.71	2.93
Relative white matter volume	26.19	1.75	26.26	1.71	25.34	1.96
	n (%)		*n (%)*		*n (%)*	
Male	626 (42.0)		566 (41.4)		60 (48.8)	
Education (Primary and secondary school as reference)	1043 (70.0)		933 (68.2)		110 (89.4)	
College	263 (17.6)		254 (18.6)		9 (7.3)	
University	185 (12.4)		181 (13.2)		4 (3.3)	
Self-reported health—good to excellent	1168 (78.3)		1091 (79.8)		77 (62.6)	
Smoking (Never as reference)	659 (44.2)		610 (44.6)		49 (39.8)	
Previously	686 (46.0)		627 (45.8)		59 (48.0)	
Current	146 (9.8)		131 (9.6)		15 (12.2)	
Coronary artery disease (No as reference)	1152 (77.3)		1066 (77.9)		86 (69.9)	
Possible case	100 (6.7)		91 (6.7)		9 (7.3)	
Yes	239 (16.0)		211 (15.4)		28 (22.8)	
Diabetes—diagnosed with	136 (9.1)		122 (8.9)		14 (11.4)	
Hypertension (No as reference)	179 (12.0)		170 (12.4)		9 (7.3)	
Pre-Hypertension	604 (40.5)		555 (40.6)		49 (39.8)	
Yes	708 (47.5)		643 (47.0)		65 (52.8)	
ApoE e4 carrier	394 (26.4)		345 (25.2)		49 (39.8)	

The table is an adapted version of Table 2 from [15]

overfitting. If that was the case, a change was made to how the hyperparameters were tuned.

Hyperparameter tuning

All machine learning models have hyperparameters that must be specified when running the algorithm. For each machine learning algorithm, a randomized search was used to identify the combination of hyperparameters that performed the best on the dataset of interest (e.g., different combinations of hidden layers and numbers of nodes for neural networks, and the number of trees, maximum number of features, and complexity of each tree in random forest).

Results

Descriptive statistics

Table 3 shows the descriptive statistics for the different exposure variables in the model for the whole dataset and split into groups based on the dementia criterion. Of those with healthy cognition at baseline

($n = 1{,}491$), 123 (8.2%) fulfilled the dementia criterion five years later.

Model performance

Table 4 compares the performance of all three machine learning algorithms after hyperparameter tuning. According to the confusion matrix, the neural network model had the highest number of true negatives and the lowest number of false positives. The random forest model had the lowest number of false negatives and the highest number of true positives. The random forest algorithm performed best on balanced accuracy, precision, recall, F1 score, and MCC score metrics. The neural network algorithm performed best according to the AUC-ROC metric.

Random forest

Since the random forest algorithm performed well on most of the performance metrics, the results from the random forest model were analyzed further. Figure 4 shows the feature importance for each variable included in the model. According to the random forest algorithm, the five most important features are white matter brain volume, leisure activities, mobility, age, and grey matter brain volume, in that order.

Discussion

This study assessed the performance of three different machine-learning algorithms at estimating future dementia cases based on known risk factors for dementia and cognitive decline. Logistic regression, random forest, and neural network algorithms were compared. The results showed that random forest scored higher than logistic regression and neural networks on most performance metrics. The confusion matrix showed that random forest had the highest number of true positives and the lowest number of false negatives from the assessed algorithms. However, neural networks had the highest number of true negatives and the lowest number of false positives. Random forest scored higher than logistic regression and neural networks on balanced accuracy, precision, recall, F1 score, and Matthews correlation coefficient. However, logistic regression had more area under the ROC curve than random forest and neural networks.

Table 4 Comparison of performance metrics assessing the three different machine learning models being studied

	Confusion matrix										
	True negative[a]	False negative[b]	False positive[c]	True positive[d]	Accuracy	Balanced accuracy	Precision	Recall	AUC-ROC	F1 score	Matthews correlation coefficient
Logistic Regression	445	38	7	3	0.909	0.529	0.300	0.073	**0.764**	0.882	0.113
Random forest	448	**37**	4	**4**	**0.917**	**0.544**	**0.500**	**0.098**	0.686	**0.890**	**0.194**
Neural network	**449**	39	**3**	2	0.915	0.521	0.400	0.049	0.749	0.883	0.116

There were 493 participants in the test dataset, and these calculations are all based on those participants. [a]True negative: participants that are categorized into the healthy cognition group and truly belong to that group. [b]False negative: participants that are categorized into the healthy cognition group but do not belong to that group. [c]False positive: participants that are categorized into the dementia group but do not belong to that group. [d]True positive: participants that are categorized into the dementia group and truly belong to that group

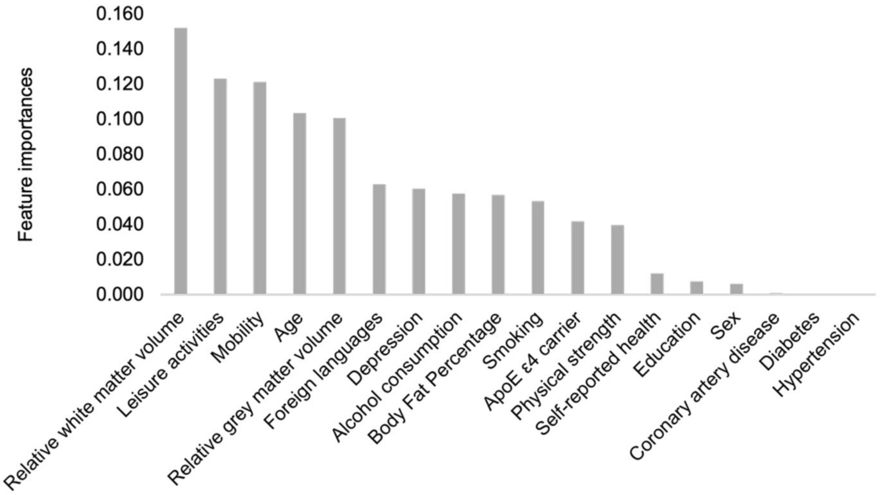


Fig. 4 Feature importance for random forest algorithm

The findings align with the results reported by Spooner et al. [24], which suggest that random forest algorithms are well-equipped to analyze data regarding predictors of dementia onset. In their findings, Casanova et al. [25] focused on genetic and modifiable risk factors for developing dementia, while Qiu et al. [26] emphasized the predictability of MRI data. Both studies showed promising results, indicating that information on modifiable risk factors and brain imaging data performed well at predicting the diagnosis of dementia. The results of this study agree with those findings.

Similar to this study, Seligman et al. [22] also compared the performance of different machine-learning algorithms in analyzing health-related data. Both studies showed that the machine learning models only performed moderately well, and Seligman et al. [22] even go so far as to state that some machine learning algorithms cannot produce better predictions than traditional simpler models. However, what differs between the two studies is that while the results of this study indicate that random forest performs better than logistic regression and neural networks, the results from Seligman et al. [22] showed that neural networks significantly outperformed linear regression methods and random forests.

Multiple performance metrics exist in the machine-learning literature, and several were calculated for this analysis. The random forest algorithm performed best according to balanced accuracy, the F1 score, and the Matthews correlation coefficient. All of them are widely used, but studies that have compared the quality of different performance metrics suggest that the MCC is the best classification performance metric for imbalanced datasets where both correct and incorrect classifications must be considered [53, 54]. The area under the ROC curve was the only metric where other performance metrics outperformed random forest. Both logistic regression and neural networks performed better, according to the AUC-ROC. However, the metric has been criticized for overestimating performance when the analyzed dataset is imbalanced, which is the case here [53, 55]. These findings emphasize that the random forest algorithm is the best choice for this dataset. What further supports that choice is the fact that out of the more popular machine learning methods, the random forest algorithm is one of the more easily understood and interpretable methods, in addition to providing information about the relative importance of the predictors included in a model [56].

Precision and recall do not perform well on all datasets but were included in the analysis to see how they compared to the other metrics used. Knowing the recall value is interesting since it indicates how well the model identifies the real positive cases, which is often the focus when analyzing medical data [57]. Recall is considered important when false negatives (classifying an individual to the healthy cognition group when he belongs to the dementia group) are undesirable. However, precision is considered important when false positives (classifying an individual to the dementia group when he belongs to the healthy cognition group) are undesirable. Stating which type of error would be more detrimental is complicated, and such a decision would depend on the models' eventual use. If the purpose were to identify which types of individuals would most likely benefit from lifestyle interventions to reduce the risk of developing dementia but not to inform an individual that he has a high risk of developing dementia in a few years, recall would be the metric of interest. Precision would be a more interesting metric if the purpose were to use the model as a risk calculator on an individual level. The random forest algorithm had higher precision and recall than neural networks and logistic regression. However, when neural networks and logistic regression are compared, the results show that the neural network algorithm performs better on the precision metric, and the logistic regression algorithm performs better on the recall metric.

According to Valsdóttir et al. [28], where logistic regression was performed on the same dataset using the statistical program SPSS, the results showed that leisure activities, self-reported health, education, age, being an ApoE $\varepsilon4$ carrier, and white and grey matter volume had a significant relationship with whether an individual would be assessed as having dementia five years later. When those results are compared to the feature importance analysis performed on the RF algorithm in this study, it is evident that the findings are not the same. For example, the number of foreign languages an individual speaks has much higher feature importance compared to education and self-reported health, even though the foreign language variable is not significant in the logistic regression model performed in SPSS. Additionally, these results do not align with the RF algorithm-based findings presented by Casanova et al. [25], which suggests that education is the top-ranked predictor for dementia progression. The reasons for this are unclear but might be explained by the characteristics of the cohort that participated in the AGES-Reykjavik Study [29]. The study's participants did not have the same educational opportunities that young people have today and may not have had the same access to education as their peers in other countries. That might be why other variables representing potential educational attainment are ranked higher in importance in this dataset.

Strengths and limitations

Even though the study's results might, at first glance, seem very decisive, it is vital to consider them in conjunction with the study's limitations and the data included in it. In the context of machine learning, the dataset being analyzed here is not very large and possibly not well suited for some of the machine learning methods available. Therefore, based on these findings, it is impossible to conclude that random forest algorithms would always perform better on similar but larger datasets.

The study was also limited by the information available in the dataset. The balanced accuracy metric gives insight into how accurate the models' predictions are. The balanced accuracy values in this study suggest that the classification could be significantly improved. The best way would be to include more suitable predictors in the models. Limitations relating to the dataset involve the number of variables available for analysis and that a portion of the variables is based on self-reported information, which needs to be interpreted cautiously. Of course, it is also important to consider that not all available and relevant data was fed into the models (such as MRI data for individual brain regions) for this study. Further studies could include more pertinent data in the models and produce more accurate predictions.

Finally, it must be considered that the results from traditional statistical analysis are generally more accessible to interpret than the outcomes from machine learning algorithms. Although some machine learning algorithms could produce much more accurate predictions for a particular dataset, it could also be the case that it could prove challenging to interpret what the results mean, which could be why researchers might be hesitant to try this more novel method. Therefore, it is essential to contemplate the purpose of analysis before deciding which approach would be most relevant.

The study also had strengths that are worth mentioning. In this study, the same dataset was analyzed as had been used in a previous analysis with the same goal that relied on traditional statistical methods. Since the results suggest that more accuracy could be reached by applying machine learning methods, such as RF algorithms, this could inspire researchers that have until now exclusively relied on traditional statistical approaches to branch out and incorporate in their work other less conventional ways of analyzing data.

Finally, the quality of the AGES-Reykjavik dataset significantly strengthens the results. Experts from many fields contributed to the study design, resulting in a high-quality dataset based on individuals with heterogeneous capabilities [29].

Future research

Since the decision was made to use the same dataset as had been used in a previous publication [28], the dataset used for this analysis was much smaller than the original AGES-Reykjavik dataset. That is because data points with missing data were excluded from the analysis when logistic regression was performed in SPSS. Future research could entail performing a principal component analysis or using other feature selection techniques on the original dataset from the AGES-Reykjavik Study to reduce the dimensionality of the dataset. This could draw out summary variables that could increase interpretability or other important variables not identified in previous dataset analyses.

Another avenue that could be taken with further research would be to focus on using such models within a clinical setting. Some people think that getting dementia is an inevitable part of aging. In contrast, others are convinced they are not a part of the group that will receive a dementia diagnosis when they get older. However, all these individuals could benefit from the prevention strategies proven to decrease the risk of developing dementia. The purpose of applying these models in a clinical setting would be to identify individuals with an increased risk of developing dementia to emphasize the importance of preventative measures among those individuals. In the future, a model such as this one could be applied on an individual level to indicate whether they need to take action to reduce the risk of developing dementia later in life. For older individuals, information about performance on neuropsychological tests could also be included to get even more accurate predictions about the likelihood of progression to dementia.

Conclusion

These preliminary results suggest that a random forest algorithm might better identify individuals in the AGES-Reykjavik Study dataset likely to develop dementia within five years compared to logistic regression and neural networks. However, it must be taken into consideration that these results only indicate better performance since statistical comparisons were not performed. The next step would be to perform such analyses on a more expansive portion of the AGES-Reykjavik Study dataset. These findings emphasize the possibility of gaining further insight into large and complex health-related datasets by exploring further the application of machine learning methods to such datasets. Further studies could support the development of a personalized dementia risk assessment tool that could be used to help people reduce their risk of developing dementia.

Funding This work was supported by The Foundation of St. Josef's Hospital in cooperation with The Icelandic Gerontological Research Center, National University Hospital of Iceland; Landspítali – University Hospital Research Fund; the Icelandic Gerontological Society; the Council on Aging in Iceland; Helga Jónsdóttir and Sigurliði Kristjánsson Memorial Fund and the Sustainability Institute and Forum (SIF) at Reykjavik University.

The AGES-Reykjavik Study received funding from the National Institutes of Health, Intramural Research Programs of the National Institute of Aging and the National Eye Institute (grant number ZIAEY00401); National Institutes of Health (contract number N01-AG-1–2100); the Icelandic Heart Association and the Icelandic Parliament.

Declarations

Competing interests The authors declare that there is no conflict of interest.

References

1. Smith GE, Bondi MW. Mild Cognitive Impairment and Dementia: Definitions, Diagnosis, and Treatment. Oxford, New York: Oxford University Press; 2013.
2. Cahill S, Pierce M, Werner P, Darley A, Bobersky A. A systematic review of the public's knowledge and understanding of Alzheimer's disease and dementia. Alzheimer Dis Assoc Disord. 2015;29(3):255–75. https://doi.org/10.1097/WAD.0000000000000102.
3. Livingston G, et al. Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. The

Lancet. 2020;396(10248):413–46. https://doi.org/10.1016/S0140-6736(20)30367-6.

4. Qiu C, Fratiglioni L. Aging without dementia is achievable: Current evidence from epidemiological research. J Alzheimers Dis. 2018;62(3):933–42. https://doi.org/10.3233/JAD-171037.
5. Gatz M, et al. Role of genes and environments for explaining Alzheimer disease. Arch Gen Psychiatry. 2006;63(2):168–74. https://doi.org/10.1001/archpsyc.63.2.168.
6. World Health Organization. Global status report on the public health response to dementia," World Health Organization, Geneva, 2021. Accessed: Jul. 26, 2022. [Online]. Available: https://www.who.int/publications-detail-redirect/9789240033245
7. Clare L, et al. Potentially modifiable lifestyle factors, cognitive reserve, and cognitive function in later life: A cross-sectional study. PLoS Med. 2017;14(3):e1002259. https://doi.org/10.1371/journal.pmed.1002259.
8. Román GC, Jackson RE, Gadhia R, Román AN, Reis J. Mediterranean diet: The role of long-chain ω-3 fatty acids in fish; polyphenols in fruits, vegetables, cereals, coffee, tea, cacao and wine; probiotics and vitamins in prevention of stroke, age-related cognitive decline, and Alzheimer disease. Revue Neurologique. 2019;175(10):724–41. https://doi.org/10.1016/j.neurol.2019.08.005.
9. Antoniou M. The advantages of bilingualism debate. Annu Rev Linguist. 2019;5(1):395–415. https://doi.org/10.1146/annurev-linguistics-011718-011820.
10. Bubu OM et al. Sleep, cognitive impairment, and Alzheimer's disease: A systematic review and meta-analysis. Sleep (2017); 40(1). https://doi.org/10.1093/sleep/zsw032
11. Shi L, et al. Sleep disturbances increase the risk of dementia: A systematic review and meta-analysis. Sleep Med Rev. 2018;40:4–16. https://doi.org/10.1016/j.smrv.2017.06.010.
12. Sindi S, et al. Sleep disturbances and dementia risk: A multicenter study. Alzheimer's & Dementia. 2018;14(10):1235–42. https://doi.org/10.1016/j.jalz.2018.05.012.
13. Beauchet O, et al. Poor gait performance and prediction of dementia: Results from a meta-analysis. J Am Med Dir Assoc. 2016;17(6):482–90. https://doi.org/10.1016/j.jamda.2015.12.092.
14. Blankevoort CG, et al. Physical predictors of cognitive performance in healthy older adults: A cross-sectional analysis. PLoS One. 2013;8(7):e70799. https://doi.org/10.1371/journal.pone.0070799.
15. Cooper R, et al. Objective measures of physical capability and subsequent health: a systematic review. Age Ageing. 2011;40(1):14–23. https://doi.org/10.1093/ageing/afq117.
16. Cui M, Zhang S, Liu Y, Gang X, Wang G. Grip strength and the risk of cognitive decline and dementia: A systematic review and meta-analysis of longitudinal cohort studies. Front Aging Neurosci (2021); 13 Accessed: Apr. 10, 2023. [Online]. Available: https://www.frontiersin.org/articles/https://doi.org/10.3389/fnagi.2021.625551
17. Duchowny KA, et al. Associations between handgrip strength and dementia risk, cognition, and neuroimaging outcomes in the UK Biobank Cohort Study. JAMA Netw Open. 2022;5(6):e2218314. https://doi.org/10.1001/jamanetworkopen.2022.18314.
18. Alanazi HO, Abdullah AH, Qureshi KN. A critical review for developing accurate and dynamic predictive models using machine learning methods in medicine and health care. J Med Syst. 2017;41(4):69. https://doi.org/10.1007/s10916-017-0715-6.
19. Wiens J, Shenoy ES. Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. Clin Infect Dis. 2018;66(1):149–53. https://doi.org/10.1093/cid/cix731.
20. Hashimoto DA, Rosman G, Rus D, Meireles OR. Artificial intelligence in surgery: promises and perils. Ann Surg. 2018;268(1):70–6. https://doi.org/10.1097/SLA.0000000000002693.
21. Wallace ML, et al. Multidimensional sleep and mortality in older adults: A machine-learning comparison with other risk factors. J Gerontol: Series A. 2019;74(12):1903–9. https://doi.org/10.1093/gerona/glz044.
22. Seligman B, Tuljapurkar S, Rehkopf D. Machine learning approaches to the social determinants of health in the health and retirement study. SSM - Population Health. 2018;4:95–9. https://doi.org/10.1016/j.ssmph.2017.11.008.
23. Javeed A, Dallora AL, Berglund JS, Ali A, Ali L, Anderberg P. Machine learning for dementia prediction: A systematic review and future research directions. J Med Syst. 2023;47(1):17. https://doi.org/10.1007/s10916-023-01906-7.
24. Spooner A, et al. A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. Sci Rep. 2020;10(1):1. https://doi.org/10.1038/s41598-020-77220-w.
25. Casanova R, Saldana S, Lutz MW, Plassman BL, Kuchibhatla M, Hayden KM. Investigating predictors of cognitive decline using machine learning. J Gerontol B Psychol Sci Soc Sci. 2020;75(4):733–42. https://doi.org/10.1093/geronb/gby054.
26. Qiu S, et al. Development and validation of an interpretable deep learning framework for Alzheimer's disease classification. Brain. 2020;143(6):1920–33. https://doi.org/10.1093/brain/awaa137.
27. Garcia-Gutierrez F, Delgado-Alvarez A, Delgado-Alonso C, et al. Diagnosis of Alzheimer's disease and behavioural variant frontotemporal dementia with machine learning-aided neuropsychological assessment using feature engineering and genetic algorithms. Int J Geriatric Psychiatry. 2021:1–13. https://doi.org/10.1002/gps.5667.
28. Valsdóttir V, et al. Exposure factors associated with dementia among older adults in Iceland: the AGES-Reykjavik study. GeroScience. 2023;45(3):1953–65. https://doi.org/10.1007/s11357-023-00804-7.
29. Harris TB, et al. Age, gene/environment susceptibility–Reykjavik study: Multidisciplinary applied phenomics. Am J Epidemiol. 2007;165(9):1076–87. https://doi.org/10.1093/aje/kwk115.
30. Valsdóttir V, et al. Cognition and brain health among older adults in Iceland: the AGES-Reykjavik study. Geroscience. 2022;44(6):2785–800. https://doi.org/10.1007/s11357-022-00642-z.

31. Abete P, et al. Cognitive impairment and cardiovascular diseases in the elderly. A heart–brain continuum hypothesis. Ageing Res Rev. 2014;18:41–52. https://doi.org/10.1016/j.arr.2014.07.003.
32. Cheng S-T. Cognitive reserve and the prevention of dementia: The role of physical and cognitive activities. Curr Psychiatry Rep. 2016;18(9):85. https://doi.org/10.1007/s11920-016-0721-2.
33. Feinkohl I, Price JF, Strachan MWJ, Frier BM. The impact of diabetes on cognitive decline: potential vascular, metabolic, and psychosocial risk factors. Alzheimer's Research & Therapy. 2015;7(1):46. https://doi.org/10.1186/s13195-015-0130-5.
34. Pugazhenthi S, Qin L, Reddy PH. Common neurodegenerative pathways in obesity, diabetes, and Alzheimer's disease. Biochimica et Biophysica Acta (BBA) - Mol Basis Dis. 2017;1863(5):1037–45. https://doi.org/10.1016/j.bbadis.2016.04.017.
35. Walker KA, Power MC, Gottesman RF. Defining the relationship between hypertension, cognitive decline, and dementia: A review. Curr Hypertens Rep. 2017;19(3):24. https://doi.org/10.1007/s11906-017-0724-3.
36. Weiss J, Puterman E, Prather AA, Ware EB, Rehkopf DH. A data-driven prospective study of dementia among older adults in the United States. PLoS ONE. 2020;15(10):e0239994. https://doi.org/10.1371/journal.pone.0239994.
37. Zaninotto P, Batty GD, Allerhand M, Deary IJ. Cognitive function trajectories and their determinants in older people: 8 years of follow-up in the English Longitudinal Study of Ageing. J Epidemiol Community Health. 2018;72(8):685–94. https://doi.org/10.1136/jech-2017-210116.
38. McCarrey AC, An Y, Kitner-Triolo MH, Ferrucci L, Resnick SM. Sex differences in cognitive trajectories in clinically normal older adults. Psychol Aging. 2016;31(2):166–75. https://doi.org/10.1037/pag0000070.
39. Singh-Manoux A, et al. Timing of onset of cognitive decline: results from Whitehall II prospective cohort study. BMJ. 2012;344:d7622. https://doi.org/10.1136/bmj.d7622.
40. Podsiadlo D, Richardson S. The timed 'Up & Go': a test of basic functional mobility for frail elderly persons. J Am Geriatr Soc. 1991;39(2):142–8. https://doi.org/10.1111/j.1532-5415.1991.tb01616.x.
41. Yesavage JA, et al. Development and validation of a geriatric depression screening scale: A preliminary report. J Psychiatr Res. 1982;17(1):37–49. https://doi.org/10.1016/0022-3956(82)90033-4.
42. Folstein MF, Folstein SE, McHugh PR. 'Mini-mental state': A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res. 1975;12(3):189–98. https://doi.org/10.1016/0022-3956(75)90026-6.
43. Wechsler DW. WAIS-III: Wechsler adult intelligence scale. Manual. New York: Psychological Corporation; 1955.
44. Pedregosa F, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825–30.
45. Cox DR. The Regression Analysis of Binary Sequences. J Royal Stat Soc Series B (Methodological). 1958;20(2):215–42.
46. Breiman L. Random Forests. Mach Learn. 2001;45(1):5–32. https://doi.org/10.1023/A:1010933404324.
47. McCulloch WS, Pitts WH. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys. 1943;5(4):115–33. https://doi.org/10.1007/BF02478259.
48. Stehman SV. Selecting and interpreting measures of thematic classification accuracy. Remote Sens Environ. 1997;62(1):77–89. https://doi.org/10.1016/S0034-4257(97)00083-7.
49. Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The Balanced Accuracy and Its Posterior Distribution. In 2010 20th International Conference on Pattern Recognition (2010); 3121–3124. https://doi.org/10.1109/ICPR.2010.764
50. Chinchor N. "MUC-4 Evaluation Metrics," in Fourth Message Uunderstanding Conference (MUC-4): Proceedings of a Conference Held in McLean, Virginia (1992); 1992. Accessed: Mar. 09, 2023. [Online]. Available: https://aclanthology.org/M92-1002
51. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn. 1997;30(7):1145–59. https://doi.org/10.1016/S0031-3203(96)00142-2.
52. Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics. 2000;16(5):412–24. https://doi.org/10.1093/bioinformatics/16.5.412.
53. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020;21:6. https://doi.org/10.1186/s12864-019-6413-7.
54. Luque A, Carrasco A, Martín A, de las Heras A. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognition. 2019;91:216–31. https://doi.org/10.1016/j.patcog.2019.02.023.
55. Hanczar B, Hua J, Sima C, Weinstein J, Bittner M, Dougherty ER. Small-sample precision of ROC-related estimates. Bioinformatics. 2010;26(6):822–30. https://doi.org/10.1093/bioinformatics/btq037.
56. Speiser JL, et al. Machine learning in aging: An example of developing prediction models for serious fall injury in older adults. J Gerontol: Series A. 2021;76(4):647–54. https://doi.org/10.1093/gerona/glaa138.
57. Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. Int J Machine Learning Technol. 2011;2(1):37–63. https://doi.org/10.48550/arXiv.2010.16061.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.