Predicting Dementia and Influencing Factors Investigation Based on Machine Learning Algorithms

Qiya Feng

School of Statistics and Data Science, Nankai University, Tianjin, 300071, China
email:2113071@mail.nankai.edu.cn

Abstract. Dementia is a serious problem worldwide, which brings great trouble to people's life. Therefore, to better control this disease, it is necessary to find out the main factors of dementia. Artificial Intelligence and machine learning can be very good at solving this kind of issue. This research is based on the usual physical data of patients to analyze the causes of dementia, using Dementia Patient Health, Prescriptions ML Dataset from Kaggle. First of all, data preprocessing and some other relative operations are carried out to make sure the dataset can be processed. Then 5 different models are taken into consideration, including Support Vector Machines (SVM), Naïve Bayes, decision trees, Random Forest (RF) and AdaBoost. After training these models and using them to predict dementia prevalence respectively, this research compares their performances by evaluation metrics like the accuracy, f1-score, confusion matrix and the Receiver Operating Characteristic (ROC) curve. Decision trees, RF and AdaBoost work best on this dataset, with accuracy up to 100%. Then RF is chosen for calculating and ranking feature importance. It is found that cognitive test scores, depression status and Apolipoprotein E (APOE ε4) carrying conditions are the three most important features that is related to dementia. However, some results are inconsistent with reality, such as judging important factors as unimportant. In this way, more work and improvement will need to be done in the future.

Keywords: Machine Learning, Dementia, Random Forest.

1 Introduction

As a mental illness, dementia is not a specific disease but is a syndrome that may result from different diseases [1]. Dementia can cause cognitive impairment in patients and can seriously destroy their ability of thinking, memorizing and other aspects that are associated with nerve cells and the brain [1]. Not only will patients themselves suffer the pain, but also their families will get harmed [1]. Dementia plagues a lot of people around the world. Taking China as an example, there will be a significant growth in the number of Chinese with dementia [2]. In 2020, 16.25 million people in China got dementia and the population is predicted to reach 48.93 million in 2050 [2]. Problems that come with it include increased costs, and the costs can be a huge burden on the individuals and society [3]. According to previous researches,

Y. Wang (ed.), *Proceedings of the 2024 International Conference on Artificial Intelligence and Communication (ICAIC 2024)*, Advances in Intelligent Systems Research 185,
https://doi.org/10.2991/978-94-6463-512-6_37

cardiovascular diseases, neuropsychiatric disorders, age, lifestyle and even genetics can affect the risk of dementia [1]. Timely intervention to target certain problems can reduce the risk of dementia to some extent [3]. Therefore, it is quite important to find the key influencing factors to help with dementia prevention efforts.

Dementia is usually diagnosed by a clinician based on a patient's symptoms, blood tests, brain scans and a number of other related tests. However, the entire process creates a lot of overhead for the patients and takes up a lot of their time, and some biomarkers are invasive to the human body. And even after patients make these efforts, test results may still be inaccurate since diagnosis is closely related to the doctor's subjective opinion [4]. Accordingly, some new perspectives should be applied to the diagnosis and prevention of dementia. Artificial Intelligence (AI) can be a potential tool to solve these problems due to its excellent feature extraction and prediction ability.

AI has received much attention in recent years. It mainly solves problems by simulating human intelligence. In daily life, AI has diverse applications and can be seen in games, self-driving cars, search engines and some other fields around us [5]. Machine learning (ML) is a technology to build an AI-driven application [5]. It has different learning methods, for instance, supervised learning and unsupervised learning [5]. Some common algorithms include k-means clustering, Support Vector Machines (SVMs), random forest and decision trees [6]. ML has been widely used in medical field such as the detection and diagnostic of autism, Mild Cognitive Impairment (MCI), depression and so on [7]. Recently, ML has also made some breakthroughs in the diagnosis of dementia and can deal with different types of data like neuroimaging, protein sequence and some other information [7]. Previous researches have mainly focused on analyzing data of patients' current physical conditions to determine whether they have dementia, but few studies have analyzed the causes of dementia which may be based on their usual physical data to a great extent.

This research is going to focus on Dementia Patient Health, Prescriptions ML Dataset from Kaggle to find the main factors that influence the risk of dementia. In this research, different ML algorithms will be used to predict patients' condition. The best predictor can be found from the result. On this basis, this research then will use the best predictor as a tool to give different features a ranking of importance.

2 Methods

2.1 Dataset Preparation

In this research, the dataset from Kaggle named Dementia Patient Health, Prescriptions ML Dataset is used [8]. It is a dataset about patients' health conditions including whether they have dementia and some other basic information and physical status, and the prescriptions and dosages for those with dementia. This dataset consists of 485 dementia patients and 515 non-dementia patients (1, 000 samples in total), which ensures the balance of the data shown in Fig. 1. This research focuses on

the dementia prediction, and there are two different categories in this classification task. That is 1 and 0, which means the patient have dementia or not have dementia respectively. 23 features are collected in this dataset, including their alcohol consumption level, the age, their past medical history, the education level etc. The distribution of some features classified by dementia is shown in the Fig. 2.

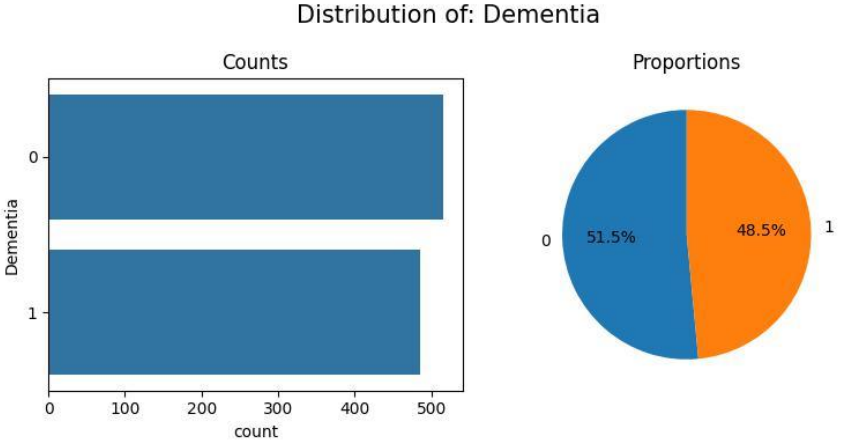

Fig. 1. Distribution of dementia (Photo/Picture credit : Original).

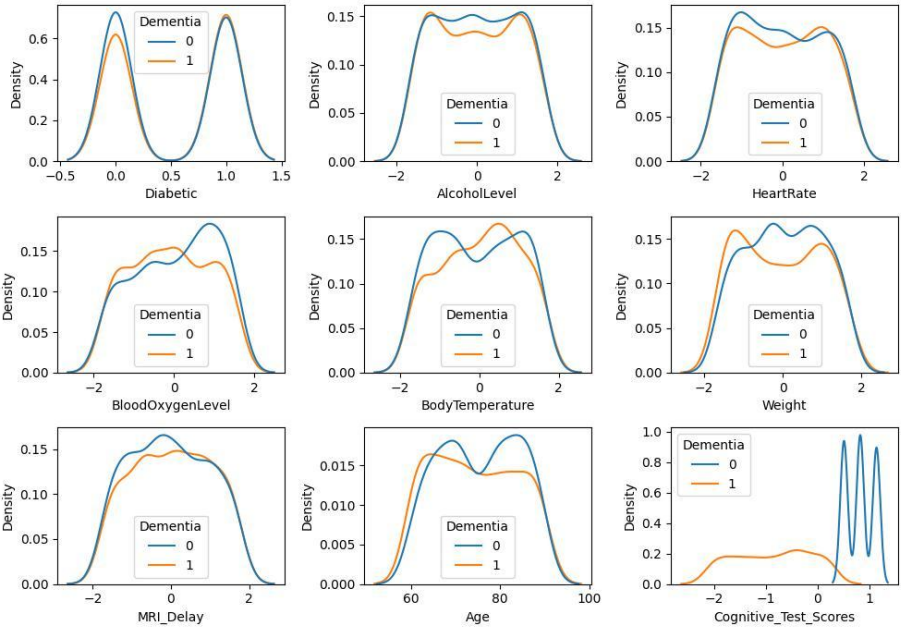

Fig. 2. Data distribution among dementia (Photo/Picture credit : Original).

Before building models, data preprocessing is implemented. The first step is cleaning the data. Through calculation, there are 515 missing values in Prescription and Dosage in mg, and 179 missing values in Chronic_Health_Conditions. Since the missing amount is quite large, it is not proper to delete these patient samples directly [9]. Prescription and Dosage in mg are treatments of dementia, and it is common that people without this illness have no relevant information. Because of that, these two features do not need to be considered as factors affecting dementia. For Chronic_Health_Conditions, it is also not proper to impute the data missed since this operation may cause the model performance to deteriorate [9]. In this way, this feature is out of consideration as well. The second step is encoding features that are not numeric, in order to facilitate model fitting. That is because there is no way to directly analyze strings. The third step is to standardize the data. This operation can help control the mean and variance of these values, which can eliminate dimensional effects between indicators. The last step is dividing training set and test set. This research takes 20% of data as test set and the remaining part for training.

2.2 Machine Learning Models

5 models are taken into this research, including Support Vector Machines (SVM), Naïve Bayes, decision trees, Random Forest (RF) and AdaBoost. These models are implemented based on built-in functions of sklearn. Evaluation metrics such as the accuracy, confusion matrix, f1-score and the receiver operating characteristic (ROC) curve are used to evaluate the quality of these models.

SVM. The basic idea of classification problem is to find a hyperplane to separate two types of training samples [10]. In the sample space, the hyperplane can be represented as followed:

$$\omega^T x + b = 0 \tag{1}$$

In this formula, ω is the normal vector and b is the replacement term. The sum of the distances from two support vectors, that are those points nearest to the hyperplane, of different categories to the hyperplane is called the margin. The basic type of SVM is to maximize the margin, equivalent to making $\|\omega\|$ minimum [10].

Naïve Bayes. Suppose that there are $\boldsymbol{N}$ types of samples, i.e. $\mathcal{Y} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_N\}$. $\boldsymbol{\lambda_{ij}}$ is the loss that misclassifies a sample of $\boldsymbol{c_j}$ to $\boldsymbol{c_i}$. Hence, based on the posterior probability, the conditional risk can be defined as:

$$\boldsymbol{R}(\boldsymbol{c_i}|\boldsymbol{x}) = \sum_{j=1}^{N} \boldsymbol{\lambda_{ij}} \boldsymbol{P}(\boldsymbol{c_j}|\boldsymbol{x}) \tag{2}.$$

The goal of Bayes classifier is to minimize the overall risk, which can be represented as:

$$\boldsymbol{R}(\boldsymbol{h}) = \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{R}(\boldsymbol{h}(\boldsymbol{x})|\boldsymbol{x})] \tag{3}$$

According to this, Bayes decision rule can be generated as choosing the categories that can make $R(h(x)|x)$ minimum, i.e.

$$\boldsymbol{h_*(x) = \arg\min_{c \in y} R(h(x)|x)} \quad (4).$$

In this formula, $h_*(x)$ is called the Bayes optimal classifier. Naïve Bayes is based on the principle above, and it also requires the dataset to satisfy the feature independence assumption [10].

Decision Trees. Decision trees have a structure of the tree. It contains the root node (that involves all of the samples), the internal node (that is the property testing), and the leaf node (that is the decision result). The goal of decision trees is to find a model that have strong generalization ability, and it follows a divide-and-conquer strategy [10]. An example of this algorithm process can be seen in Fig. 3 below.

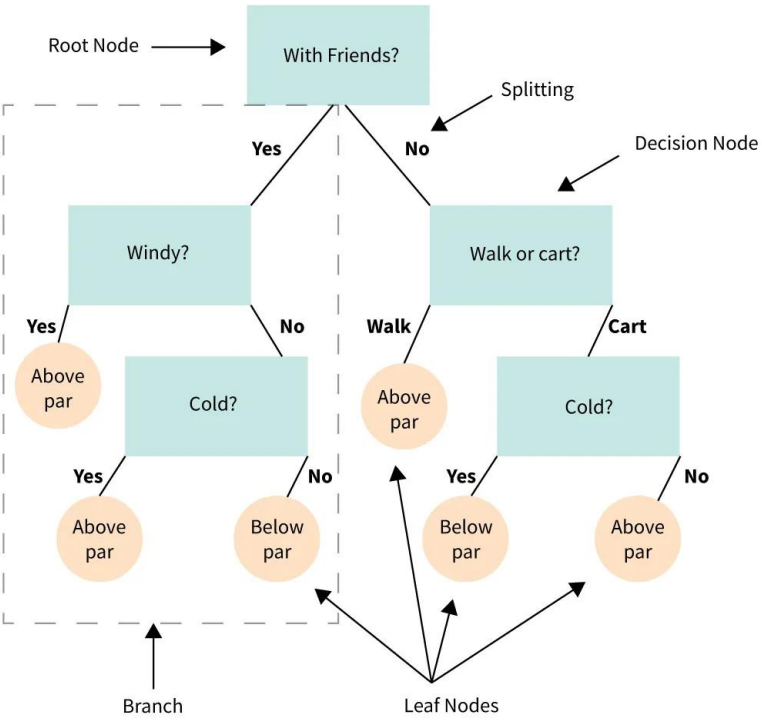

Fig. 3. Example of decision trees [11].

RF. Bagging is a parallel ensemble learning method. It uses bootstrap sampling to take several sample sets, and trains a base learner via each of them and then combines these base learners. RF is an extended variant of Bagging, using decision trees as the base learner. Furthermore, RF introduces randomized attribute selection in the training process of decision trees [10].

AdaBoost. Boosting is a family of improving-weak-learners-to-strong-ones algorithms. It trains a base learner and then adjust the contribution of training set to get a new base learner. Iterating like this, when enough base learners are collected, Boosting performs a weighted sum of them. AdaBoost is a kind of Boosting, which uses the linear combination of base learners to minimize the exponential loss function [10].

3 Results and Discussions

3.1 Model Performance

The classification results are represented in the following Table 1, Fig. 4 and Fig. 5.

Table 1. Prediction performance of different models.

	Accuracy	Precision	Recall	F1_score
Support Vector Machine	0.47	0.22	0.47	0.30
Naïve Bayes	0.96	0.96	0.96	0.96
Decision Trees	1.00	1.00	1.00	1.00
Random Forest	1.00	1.00	1.00	1.00
AdaBoost	1.00	1.00	1.00	1.00

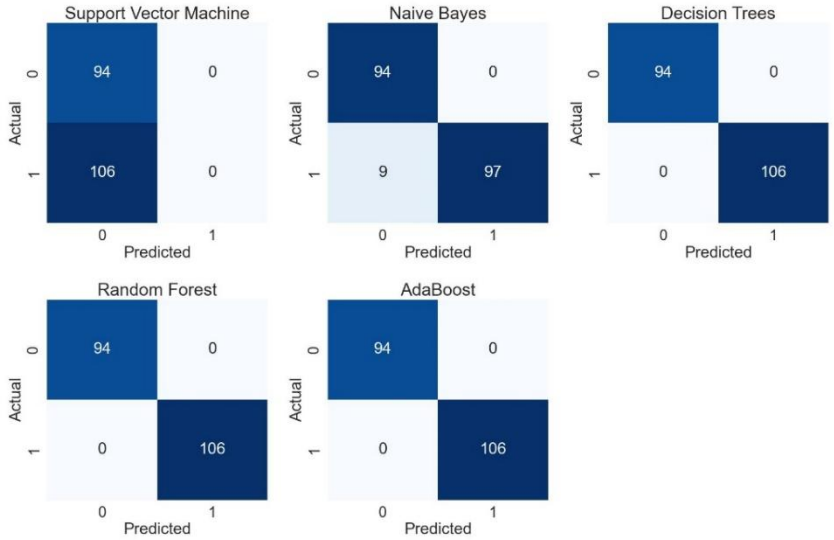

Fig. 4. Confusion matrices of different models (Photo/Picture credit : Original).

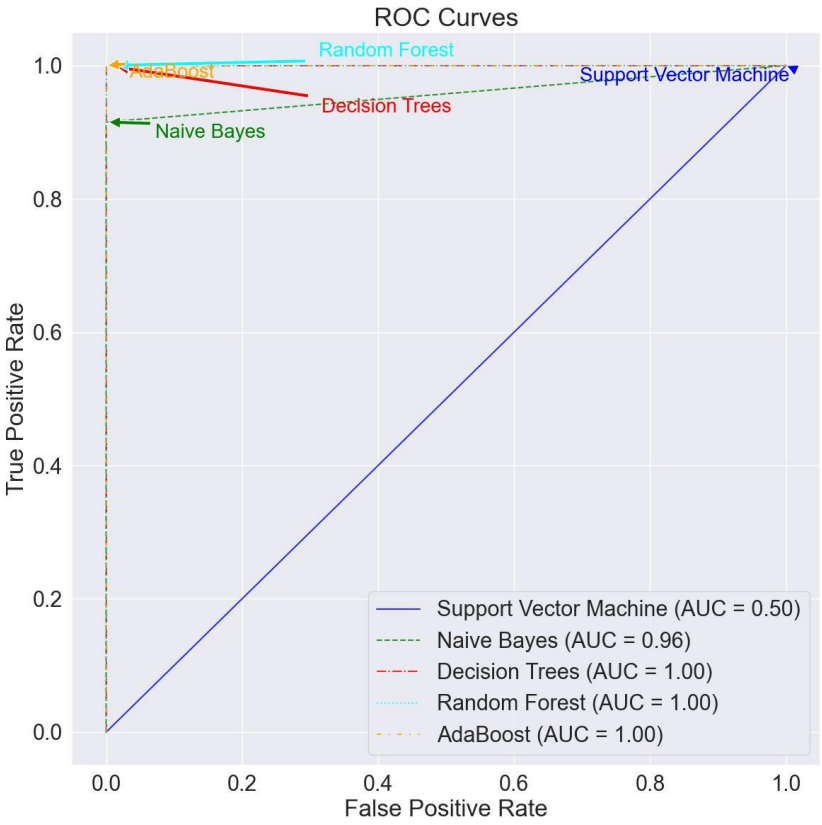

Fig. 5. ROC curves of different models (Photo/Picture credit : Original).

Comparing the detection results, it is obvious that decision trees, RF and AdaBoost perform the best on this dataset with 100% accuracy and f1-score of 1.0, and Naïve Bayes is also very good with 95.5% accuracy, which can be observed in Table 1. However, SVM is extremely inappropriate for this issue, with an accuracy of only 47% and f1-score of approximately 0.3. According to confusion matrices shown in Fig. 4, SVM judges all patients as not having dementia, while other models make few mistakes. Fig. 5 is about the ROC curves and corresponding Area Under the ROC Curve (AUC) values of these models. The closer the ROC curve is to the upper left corner and the AUC value is to 1, the better the model effect is. The same results can be found as the first two pictures.

One of the reasons why SVM has a poor performance may be overfitting, since this dataset has too many features but only with small sample size [12]. And numerous noises may also lead to the low accuracy [12]. Moreover, it is also possible that for high-dimensional data, small samples are sometimes not adequate to construct an effective hyperplane. In comparison, Naïve Bayes can delete irrelevant features during the training process and decision trees can handle noise data, so that they

perform much better than SVM [13]. And as ensemble learning methods, RF and AdaBoost combine predictions from multiple models, which can guarantee their success of forecasting in this matter [10].

3.2 Feature Importance

Based on the results of the model evaluation, this research chooses RF for feature importance ranking due to its excellent performance. The bar chart below in Fig. 6 demonstrates the order of importance from the greatest to the least. The most crucial feature is patients' scores of cognitive tests with importance of over 0.65, followed by their depression status and Apolipoprotein E (APOE ε4) carrying conditions but with significantly reduced importance of around 0.12 and 0.06 respectively. Other factors have much weaker links to whether patients develop dementia. In particular, features like the gender, medication history and diabetes prevalence are almost irrelevant to dementia, with importance of nearly 0.

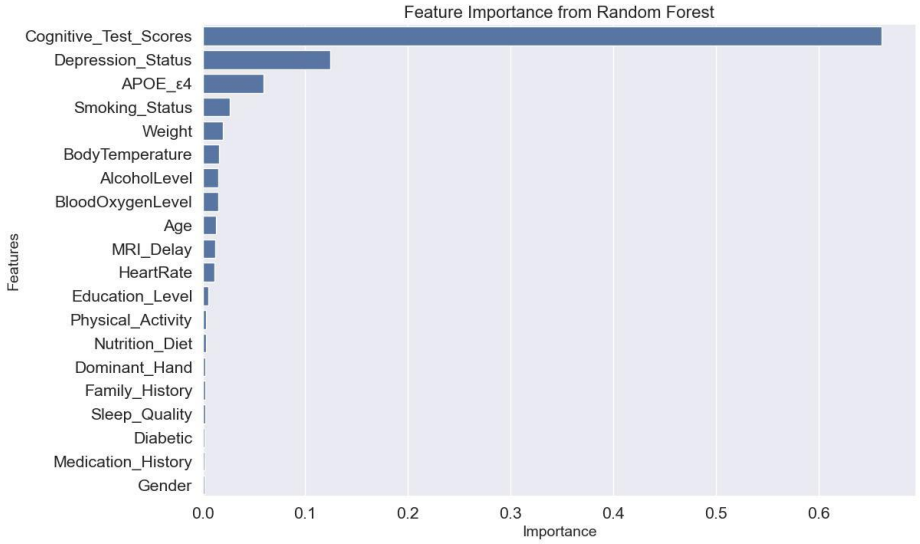

Fig. 6. Feature importance from RF (Photo/Picture credit : Original).

Many past studies provide valid evidence of these results. It can be found that many brief cognitive tests can be the accurate tools to diagnose dementia, and are harmless to the body so far [14]. In addition, APOE ε4 is the strongest genetic risk factor that can highly cause cognitive decline, which can explain why it ranks second in importance [15]. However, features like the gender actually have impact on the risk of developing dementia, which is contrary to the results of this research [16]. And according to WHO, main factors of dementia cases include diabetes, physical activities etc. [1]. Therefore, the sorting results obtained here still have deviations from reality, and the methods used still need to be improved.

4 Conclusion

To be concluded, this research introduced several machine learning methods to model the relationship between dementia and different traits, and then selects out the best predictor to calculate the importance of these trails. The main methods used here involve SVM, Naïve Bayes, decision trees, RF and AdaBoost, based on the dementia dataset from Kaggle. RF is chosen for the importance ranking when taking the accuracy and other evaluation metrics into account, and it is found that the cognitive test scores and depression status are two main factors. However, this research still has some limitations. The dataset used here is not enough to make deeper studies. More authoritative and extensive data should be considered in the future. There are also some deviations in the result of feature importance, although the model works well in predictions. In the future, it is necessary to further improve the parameters or even the structure of the model to obtain more reasonable results. And other ways to compute the importance will also be tried.

References

1. World Health Organization: Dementia.Available at: https://www.who.int/news-room/fact-sheets/detail/dementia (2023).
2. Li, F., Qin, W., Zhu, M., Jia, J.: Model-based projection of dementia prevalence in China and worldwide: 2020–2050. Journal of Alzheimer's Disease 82, 1823-1831 (2021).
3. Chowdhary, N., et al.: Reducing the risk of cognitive decline and dementia: WHO recommendations. Frontiers in Neurology 12, 765584 (2022).
4. Li, R., et al.: Applications of artificial intelligence to aid early detection of dementia: a scoping review on current capabilities and future directions. Journal of Biomedical Informatics 127, 104030 (2022).
5. Saranya, A., Subhashini, R.: A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends. Decision Analytics Journal, 100230 (2023).
6. Newby, D., et al.: Artificial intelligence for dementia prevention. Alzheimer's & Dementia 19(12), 5952-5969 (2023).
7. Mirzaei, G., Adeli, H.: Machine learning techniques for diagnosis of Alzheimer disease, mild cognitive disorder, and other types of dementia. Biomedical Signal Processing and Control 72, 103293 (2022).
8. Kaggle: Dementia Patient Health and Prescriptions Dataset. Available at: https://www.kaggle.com/datasets/kaggler2412/dementia-patient-health-and-prescriptions-dataset/data (2024).
9. Albahra, S., et al.: Artificial intelligence and machine learning overview in pathology & laboratory medicine: A general review of data preprocessing and basic supervised concepts. Seminars in Diagnostic Pathology 40(2), WB Saunders (2023).
10. Zhou, Z.H.: Machine Learning. Springer Nature (2021).
11. Masters in Data Science: Decision Tree. Available at: https://www.mastersindatascience.org/learning/machine-learning-algorithms/decision-tree/ (2024).

12. Bansal, M., Goyal, A., Choudhary, A.: A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning. Decision Analytics Journal 3, 100071 (2022).
13. Jadhav, S.D., Channe, H.P.: Comparative study of K-NN, naive Bayes and decision tree classification techniques. International Journal of Science and Research (IJSR) 5(1), 1842-1845 (2016).
14. Hemmy, L.S., et al.: Brief cognitive tests for distinguishing clinical Alzheimer-type dementia from mild cognitive impairment or normal cognition in older adults with suspected cognitive impairment: a systematic review. Annals of Internal Medicine 172(10), 678-687 (2020).
15. Gharbi-Meliani, A., et al.: The association of APOE ε4 with cognitive function over the adult life course and incidence of dementia: 20 years follow-up of the Whitehall II study. Alzheimer's Research & Therapy 13, 1-11 (2021).
16. Vila-Castelar, C., et al.: Sex and gender considerations in dementia: a call for global research. Nature Aging 3(5), 463-465 (2023).