Iran J Public Health, Vol. 52, No. 10, Oct 2023, pp.2179-2185

Original Article

Prediction of Alzheimer's in People with Coronavirus Using Machine Learning

***Shahriar Mohammadi** [1], **Soraya Zarei** [1], **Hossain Jabbari** [2,3]

1. *Information Technology Group, Department of Industrial Engineering, K.N. Toosi University of Technology, Tehran, Iran*
2. *Neurology Department, Penzing Teaching Hospital, Vienna, Austria*
3. *Digestive Diseases Research Institute, Tehran University of Medical Sciences, Tehran, Iran*

***Corresponding Author:** Email: mohammadi@kntu.ac.ir

(Received 10 Dec 2022; accepted 19 Feb 2023)

Abstract

Background: One of the negative effects of the COVID-19 illness, which has affected people all across the world, is Alzheimer's disease. Oblivion after COVID-19 has created a variety of issues for many people. Predicting this issue in COVID-19 patients can considerably lessen the severity of the problem.

Methods: Alzheimer's disease was predicted in Iranian persons with COVID-19 in using three algorithms: Nave Bayes, Random Forest, and KNN. Data collected by private questioner from hospitals of Tehran Province, Iran, during Oct 2020 to Sep 2021. For ML models, performance is quantified using measures such as Precision, Recall, Accuracy, and F1-score.

Results: The Nave Bayes, Random Forest algorithm has a prediction accuracy of higher than 80%. The predicted accuracy of the random forest algorithm was higher than the other two algorithms.

Conclusion: The Random Forest algorithm outperformed the other two algorithms in predicting Alzheimer's disease in persons using COVID-19. The findings of this study could help persons with COVID-19 avoid Alzheimer's problems.

Keywords: Alzheimer; COVID-19; Machine learning

Introduction

Since 2019, the coronavirus pandemic as an acute respiratory disease has produced unique circumstances all across the world. This disease epidemic has produced problems in people's lives around the world (1). Initially, the COVID-19 virus was discovered in Wuhan, China, and quickly spread to many countries. Although certain shutdowns and control efforts reduced the disease's spread to some level, the mutation of the virus and the appearance of new species prevented the control of that completely (2-4). The elderly and people with particular illnesses (diabetes, kidney failure, cardiovascular disease, etc.) were particularly susceptible to COVID-19 since this viral disease has a severe impact on the body's immune system. As a result, those infected with coronavirus had a higher mortality rate than those who were not (3,4).

Patients have had a variety of consequences as a result of the Coronavirus. So far, the most serious side effects have included olfactory disorders, gastrointestinal disorders, mental health problems, liver complications, etc. (5-7).

One of the most common CNS comorbidities detected in COVID-19 patients in a considerable way is Alzheimer's disease (AD) (8). The most common neurodegenerative disease and the leading cause of dementia was considered by Alzheimer's disease. In Alzheimer's disease patients, the deposition of amyloid beta or neurofibrillary tangles damages key parts of the brain that is responsible for memory and learning (9). According to the WHO Trusted Source, approximately 55 million individuals worldwide suffer from dementia annually, and 60–70% of these cases account for Alzheimer's disease (8). The results of serum neurodegenerative biomarkers of patients of COVID-19 revealed symptoms related to encephalopathy. According to this study, hospitalized COVID-19 individuals are more prone to develop neurologic problems, notably encephalopathy, and around half of them are more likely to develop long-term cognitive deficits for unknown reasons (9). Naturally, elderly people are more susceptible to Alzheimer's disease. Although, the pandemic COVID-19 increased the odds ratio for Alzheimer's disease (10). About 20% to 40% of COVID-19 patients are over sixty years old (11).

The risk of acquiring Alzheimer's disease is higher for people who carry the APOEe4 allele in their sixth decade. Likewise, the APOEe4 allele increases the chance of COVID-19 pathogenesis. The receptor for entry of COVID-19's causal agent, severe acute respiratory syndrome corona virus-2 (SARS-CoV-2), is angiotensin converting enzyme 2 (ACE 2). In Alzheimer's patients, ACE2 expression has been found to be 10 times higher, posing a hazard to their survival (12,13).

Machine Learning, a type of Artificial Intelligence (AI) is becoming more prevalent in the healthcare industry. Machine Learning is study of algorithms that improve with the use of data and experience. Machine learning is divided into two phases: training and testing. Training and testing are the two phases that make up the primary framework of machine learning (14). There is no systematic strategy that can be used to discover the best machine learning method for a specific problem a priori (15). As a result, testing different top algorithms on a new application to classify insemination events into health or illness outcomes based on the aforementioned explanatory variables is a frequent strategy in machine learning investigations (14).

Using biological factors to predict the emergence of disease is one strategy to control it. As a result, the repercussions will be less severe. The goal of this study was using three machine learning methods to predict the occurrence of Alzheimer's disease in patients with COVID-19 illness.

Materials and Methods

Patients with COVID-19 (n=3384) and Alzheimer's disease (n=998) were found in five hospitals in Tehran, Iran. The information was gathered from the Oct 2020 to Sep 2021, inclusive. Data on health, education level (illiterate, high school, and above the diploma), inherited records, age, and sex were also accessible. The obtained information was related to the age groups of 35 to 70 years.

We've used a variety of machine learning techniques to forecast Alzheimer's disease. As classifiers, these algorithms include Nave Bayes, Random Forest, and K-Nearest Neighbor. Python-based machine learning and data mining software. This is the prediction model's final phase. We use several assessment measures, such as classification accuracy, confusion matrix, and f1-score, to evaluate the prediction results.

Preparation and Splitting the Data

Our project seeks to create a web application that uses machine learning models to detect ailments COVID-19 patient that experienced Alzheimer. The web application developed using machine learning approaches to detect Alzheimer's patients is illustrated in Fig. 1.

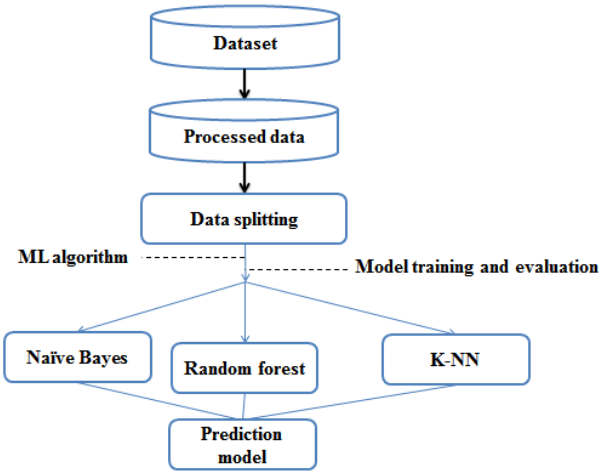

Fig. 1: Process work flow

Machine Learning Algorithms

There is no systematic strategy that can be used to discover the best machine learning method for a specific problem a priori. As a result, testing different leading algorithms on a given application is a popular method in machine learning studies. The leading algorithms for learning Nave Bayes, random forest, and K-Nearest Neighbors algorithms were put to the test in this study. Following is a quick description of each technique.

Naïve Bayes

Nave Bayes (NB) is a Bayes rule-based statistical classifier that is one of the most efficient and effective inductive learning algorithms in which all features are independent, given the outcome value. Its simplicity, computational feasibility, and resilience make it suited for practical usage (16). Nave Bayes can tolerate feature dependencies quite well, and they frequently outperform more complex approaches like rule learners and decision tree learners (17). Furthermore, NB are very intuitive and simple to grasp, which is a major problem in the realm of machine learning. Linear dependencies between features, on the other hand, can diminish the power of NB, therefore cautious selection among highly dependent characteristics can be advantageous. Another issue with NB is that the assumption of normality for numeric features is not always valid, and employing kernel density estimation can help in this situation (18). Assume B is a vector of features (b1, b2,..., bn), and c is a two-valued class variable (O=unaffected, P=affected). Given the feature vector, one can determine the probability (p) of the class variable (C):

$$p\ (C = c\ |b1,\ b2,... bn) = \frac{p\ (b1,\ b2,... bn|C = c) \times p\ (C = c)}{p\ (b1,\ b2,... bn)}$$

Given the class variable, we assume that all features are independent (conditional independence),

$$p\ (b1,\ b2,... bn|C = c) = p\ (b1|C = c)p\ (b2|C = c)...p\ (bn|C = c);$$
$$p\ (b1,\ b2,... b_n) = p\ (b1)p\ (b2)...p\ (bn)$$

The prior probability for a given class, p (C = c), can be easily determined from the training data.

Random Forest

Another ensemble method is random forest (RF), which combines training numerous classifiers using m bootstrap samples from the training set with random selection of a subset of features for each of those classifiers (19). Thus, an RF algorithm is essentially similar to bootstrap, except that in each iteration of tree construction, RF selects a random subset of features and divides the instances based on the most informative feature. RF is a high-dimensional data-processing technology that is computationally efficient. One of the most significant advantages of RF, and the primary reason it was chosen for this study, is that it is very efficient at estimating missing values and can maintain high accuracy even when a large proportion of the data is missing, which is a common scenario when analyzing producer-reported health data.

K-Nearest Neighbors

This algorithm, often known as KNN or k-NN, which is a non-parametric, supervised learning classifier that makes classifications or predictions about the grouping of individual data points based on closeness. KNN algorithms have been employed in numerous applications such as sta-

tistical estimation and pattern recognition for more than 50 years. KNN is a non-parametric classification approach that can be divided into two categories. 1) NN approaches with less structure 2) NN approaches based on structure. The total data is classified into training and test sample data in structure less NN algorithms. The distance between the training point and the sample point is calculated, and the point with the shortest distance is referred to as the nearest neighbor. Data structures such as the orthogonal structure tree (OST), ball tree, k-d tree, axis tree, closest future line, and central line are used in structure-based NN approaches (20).

Performance criteria like as accuracy, precision, recall, and F1 score were assessed to establish the ideal settings for each model. Each model's accuracy was compared. The confusion matrix is used to define performance evaluations, which can be binary or multiclass.

Accuracy is a metric for determining the percentage of correctly classified results among all instances.

$$Accuracy = (TN/TP + TN + FP + FN) \times 100$$

Precision: By dividing the number of accurately anticipated positive rates by the total number of expected positive rates is calculated. If the Precision value is 1, the classifier is considered good.

$$\text{Pr}ecision = TP/TP + FP$$

True positive rate (recall) is a true positive rate. the classifier is considered good when the recall is obtained 1.

$$\text{Re}call = TP/TP + FN$$

F1 Score: This is a metric that takes into account both the Recall and Precision factors. Only when both of the measures, such as Recall and Precision, are 1, does the F1 score become 1.

$$F1Score(\%) = 2 \times \frac{Recall \times Precision}{Recall + Precision}$$

Results

Figure 2 depicts the proportion of patients with Alzheimer's disease in various age groups. The percentage of adults over the age of 65 who have Alzheimer's disease is larger than the other two groups.

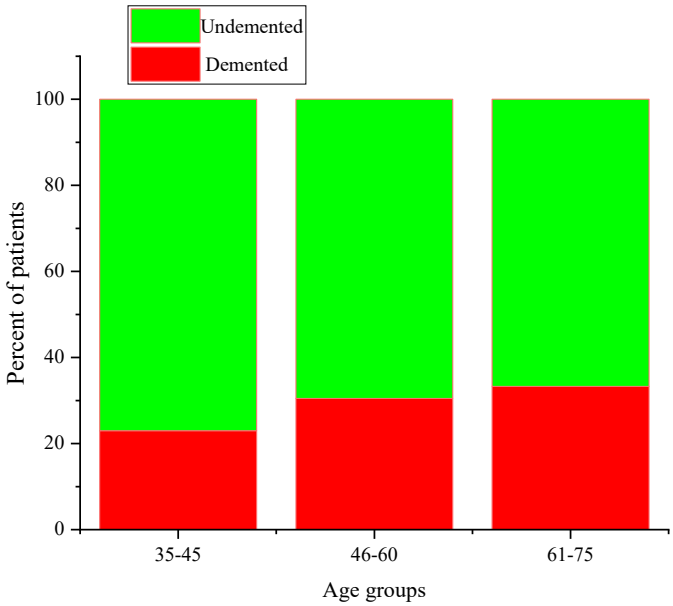

Fig. 2: The percentage of patients with COVID-19 who get Alzheimer's disease at various ages

Classification Accuracy

The conversions of the True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) measurements are the most popular metrics. The confusion matrix for Naive Bayes, Random Forest, and K-NN classifier ML models is shown in Fig. 3. The FP rate is the percentage of negative examples (unaffected) that are wrongly predicted as positive examples (affected). The true positive rate is the percentage of true positive cases that are accurately predicted as true positive examples (affected). The number of accurately predicted persons with Alzheimer's disease was higher in the random forest model than in the other two models. The KNN model was linked to the lowest percentage of proper Alzheimer's diagnosis.

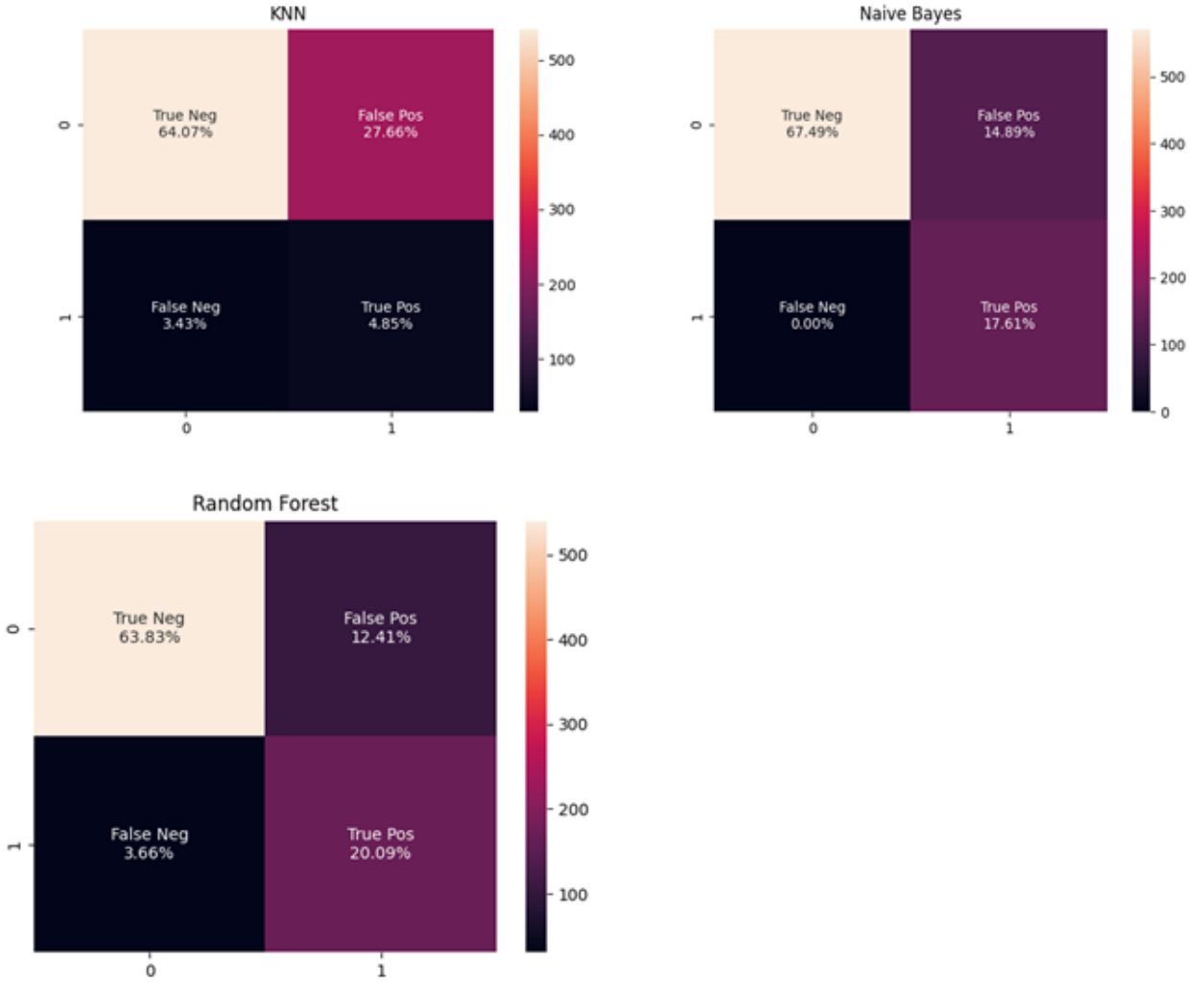

Fig. 3: The confusion matrix of ML algorithms used

The comparing accuracy, precision, recall, and F1-score of Alzheimer's prediction shown in Table 1. The prediction accuracy of the two algorithms Nave Bayes and Random Forest is expected to be around 80% higher than the KNN algorithm. The estimation accuracy of the KNN algorithm was less than 20%, while the estimation accuracy of the Random Forest approach was greater than 60%. (Table 1).

The Nave Bayes method had a greater True positive rate share than the other two algorithms (Table 1). In general, the accuracy of the Nave Bayes and Random Forest algorithms in predicting Alzheimer's disease in persons using COVID-19 was higher than the KNN method (Table 1).

Table 1: The comparison of accuracy, precision, recall, and F1-score of Alzheimer's prediction in studied algorithm

	Algorithms		
Variables	Random forest	KNN	Naïve Bayes
Accuracy	83.9	68.9	85.1
precision	61.9	15.7	54.5
Recall	85.1	58.9	98.9
F1-score	71.9	24.1	70.0

Discussion

In our study the elder people by corona patient had more Alzheimer disease than younger that was agree with previous studies. Recent research has found a link between aging and the disease of amnesia, the severity of which can be influenced by factors other than age, such as the immune system (12,13). Persons with COVID-19 who had amnesia were more likely to be older.

The use of random forest algorithm in the first place and Naïve Bayes after that have higher prediction accuracy and precision for predicting the occurrence of Alzheimer's disease in people with corona. The use of phenotypic data is less expensive than MRI images to predict the occurrence of disease. Of course, it depends on the amount of data collected and also the algorithm used for it. Moreover, our study showed that the use of information related to the social and economic status of people with Corona patient can play a significant role in predicting complications related to this disease such as Alzheimer's, and the use of machine learning algorithms can significantly help in controlling severe complications

In general, using various circumstances to forecast disease in persons suffering from various complications such as Alzheimer's disease can play a significant role in disease prevention and control. Different algorithms' prediction accuracy varies based on the nature of the data and the amount of information utilized (14-18). The usage of Nave Bayes and Random Forest algorithms for Alzheimer's illness was demonstrated to be appropriate in our investigation.

Conclusion

As a complication of COVID-19 disease, Alzheimer's disease has caused numerous complications in some individuals, and forecasting its emergence can help to mitigate these issues to some extent. In general, using the Random Forest method, which is known for its excellent accuracy, as well as the Nave Bayes algorithm, could help persons with COVID-19 control their Alzheimer's disease.

Journalism Ethics considerations

Ethical issues (Including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy, etc.) have been completely observed by the authors.

Acknowledgements

The authors express their gratitude to the managers and staff of Tehran Province hospitals for their assistance in gathering data for this study.

Conflicts of interest

The authors of this article declare that they have no conflict of interests.

References

1. Chang SL, Harding N, Zachreson C, et al (2020). Modelling transmission and control of the COVID-19 pandemic in Australia. *Nat Commun,* 11 (1):5710.
2. Sookaromdee P, Wiwanitkit V (2020). Imported cases of 2019-novel coronavirus (2019-nCoV) infections in Thailand: Mathematical modelling of the outbreak. *Asian Pac J Trop Med,* 13 (3):139-140
3. Motamed-Jahromi M, Kaveh MH (2021). The Social Consequences of the Novel Coronavirus Disease (COVID-19) Outbreak in Iran: Is Social Capital at Risk? A Qualitative Study. *Interdiscip Perspect Infect Dis,* 2021: 5553859.
4. Hui DS, Azhar EI, Madani TA, et al (2020). The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health—The latest 2019 novel coronavirus outbreak in Wuhan, China. *Int J Infect Dis,* 91:264-266.
5. Mohamed DZ, Ghoneim ME-S, Abu-Risha SE-S, et al (2021). Gastrointestinal and hepatic diseases during the COVID-19 pandemic: Manifestations, mechanism and management. *World J Gastroenterol,* 27(28): 4504-4535.
6. Sahu T, Mehta A, Ratre YK, et al (2021). Current understanding of the impact of COVID-19 on gastrointestinal disease: Challenges and openings. *World J Gastroenterol,* 27 (6):449-469.
7. Brüssow H, Timmis K (2021). COVID-19: long COVID and its societal consequences. *Environ Microbiol,* 23(8):4077-4091.
8. Fotuhi M, Mian A, Meysami S, Raji CA (2020). Neurobiology of COVID-19. *J Alzheimers Dis*, 76:3-19.
9. Frontera JA, Boutajangout A, Masurkar AV, et al (2022). Comparison of serum neurodegenerative biomarkers among hospitalized COVID-19 patients versus non-COVID subjects with normal cognition, mild cognitive impairment, or Alzheimer's dementia. *Alzheimers Dement,* 18 (5):899-910.
10. Association As (2019). Alzheimer's disease facts and figures. *Alzheimers Dement,* 15 (3):321-387.
11. Etard J-F, Vanhems P, Atlani-Duault L, et al (2020). Potential lethal outbreak of coronavirus disease (COVID-19) among the elderly in retirement homes and long-term facilities, France, March 2020. *Euro Surveill,* 25 (15):2000448.
12. Kuo C-L, Pilling LC, Atkins JL, et al (2020). APOE e4 genotype predicts severe COVID-19 in the UK Biobank community cohort. *J Gerontol A Biol Sci Med Sci,* 75(11):2231-2232.
13. Hultman K, Strickland S, Norris EH (2013). The APOE ε4/ε4 genotype potentiates vascular fibrin (ogen) deposition in amyloid-laden vessels in the brains of Alzheimer's disease patients. *J Cereb Blood Flow Metab,* 33 (8):1251-8.
14. Chauhan RH, Naik DN, Halpati RA, et al (2020). Disease Prediction using Machine Learning. *Int Res J Eng Technol,* 7 (5):2000-2002.
15. Shahinfar S, Guenther JN, Page CD, et al (2015). Optimization of reproductive management programs using lift chart analysis and cost-sensitive evaluation of classification errors. *J Dairy Sci,* 98 (6):3717-3728.
16. Domingos P, Pazzani M (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning,* 29 (2):103-130.
17. Cestnik B (1990). Estimating probabilities: A crucial task in machine learning. In: Proceedings of the European Conference on Artificial Intelligence, Stockholm, Sweden, 1 January 1990, 1990. pp 147-149.
18. Witten I, Frank E, Hall M (2005). The explorer. Data Mining Practical Machine Learning *Tools and Techniques,* 445-479.
19. Breiman L (2001). Random forests. *Machine Learning*, 45 (1):5-32.
20. Bhatia N, Vandana (2010). Survey of nearest neighbor techniques. *Int J Comput Sci Inf Technol,* 8 (2):302-305.