Assessing the Risks of COVID-19 on the Health Conditions of Alzheimer's Patients Using Machine Learning Techniques

Prosenjit Karmaker and Muhammad Sajjadur Rahim(✉)

Department of Information and Communication Engineering, University of Rajshahi,
Rajshahi 6205, Bangladesh
sajid_ice@ru.ac.bd

Abstract. There is currently little evidence linking COVID-19 to Alzheimer's Disease (AD). The goal of this paper is to examine the correlation among COVID-19 symptoms to identify risks for AD patients and to determine the conditions that put AD patients in danger. We have developed a Machine Learning (ML) based model called *AD-Cov-CorrelationNet* that shows the relationship between various health issues and whether every attribute in the dataset is connected. We have discovered a direct link between several health issues in AD patients. The risk of getting an infection when they are directly contacted by the outside environment is very high. Although there is no direct contact with the outside environment, AD patients are still vulnerable to some health issues which cause serious problems and increase the risks of death. Supervised learning models such as Logistic Regression, K-Nearest Neighbor (KNN), Decision Tree, Random Forest, Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP) are utilized to understand the disease prognosis. The risk factors that the models predicted are clinically meaningful and relevant to reducing fatality. This comparative analysis achieves more than 98% accuracy, 97% precision, 97% recall, 97% F1 score, and accurate Receiver Operating Characteristic (ROC) curves.

Keywords: COVID-19 · Alzheimer's disease · Machine learning algorithms · Cross-validation · Correlation · Symptom · Predictive model · Logistic Regression · KNN · Decision Tree · Random Forest · SVM · MLP

1 Introduction

Late in 2019, the new coronavirus SARS-CoV-2 made its preliminary appearance in China, with a market in Wuhan, China, serving as its source. The COVID-19 virus spread across the world, causing the World Health Organization to proclaim a global outbreak in March 2020. There have been 3,349,786 COVID-19 cases and 238,628 deaths globally as of May 3, 2020. As our understanding of COVID-19 has grown, older age groups have emerged as one of the key risk variables linked to horrible outcomes following infection, with adults over 58 years old having a risk of dying from COVID-19

Md. S. Satu et al. (Eds.): MIET 2022, LNICST 490, pp. 253–266, 2023.
https://doi.org/10.1007/978-3-031-34619-4_21

that is double that of children [1]. COVID-19 symptoms can vary from mild to severe, and can even be fatal in some cases. Coughing, fever, loss of smell and taste are all common side effects, with migraine, nasal congestion, and respiratory problems being less so. In moderate to severe cases, other symptoms include severe stomach pains, sore throat, diarrhea, eye problems, swelling or purple toes, and breathlessness.

A neurological disorder known as Alzheimer's disease (AD) is characterized by memory loss, emotional disturbances, and abnormalities of the behavioral system. Alzheimer's disease affects more than 50 million individuals worldwide (Alzheimer's Report WHO), and most pharmaceutical medicines only have palliative effect. Aside from negatively impacting quality of life for patients and human health, Alzheimer's disease has a large financial impact. The most widespread degenerative nerve disorder globally is Alzheimer's disease (AD), indicating that up to 80% of Alzheimer's is caused. Among the 50 leading reasons for decreased life expectancy, this is one of the fastest-growing; if current trends continue, the number of Alzheimer's disease patients will exceed 150 million by 2050 [2, 3]. Patients with Alzheimer's disease often have short-term and long-term memory loss, as well as confusion, rage, violence, language issues, and mood changes as the disease progresses. Alzheimer's disease has a global economic cost of one billion dollars every year.

State-of-the-art supervised learning models such as Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, Random Forest, Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP) are used to assess the prognosis and course of the disease. This technique can classify enormous volumes of unstructured data, including correlations between symptoms and outcomes [4, 5]. Machine learning architectures and algorithms have developed recently because of their use in a variety of industries, including speech recognition, picture processing, and answering biological inquiries. In biological circumstances, their risk relationship might be dependent on other independent causative factors that have a strong correlation to the disease. However, a variety of unbalanced datasets frequently limits the model performance. In addition, all these models exhibit individual limitations. The results are superior to the accuracy of the diagnosis. The risk factors that the models predicted are clinically meaningful and relevant to reduce fatality. The Pearson's correlation model is expected to perform noticeably better than its competitors in simulating the complex interconnection of Alzheimer's disease (AD) patient risks to catch COVID-19.

As a result, three research questions (RQs) are investigated to evaluate the effectiveness of the suggested tactic for state-of-the-art approaches:

- RQ1: How can unbalanced data be effectively handled and prepared for machine learning (ML) models? Or, how can unbalanced data be made more balanced in order to process ML models?
- RQ2: How effectively can correlation method categorize an AD patient's COVID-19 infection risk based on symptoms?
- RQ3: What possible risk factors could lead to the serious problems of the Alzheimer's disease patients with COVID-19 (AD-COVID-19)?

The proposed solution is correlation study. To accomplish this, we have developed a ML-based *AD-Cov-CorrelationNet* model. Using the proposed model, this research shows the relationship between various health issues and whether every attribute in the dataset is connected to one another. Relational attributes vary with strength and direction when the main attribute changes. We discovered a direct link between several health issues that will be risky for Alzheimer's disease (AD) patients and increase their risk of death. The goal of our inquiry is to identify the risk factors that endanger people with Alzheimer's disease (AD). This paper's main contributions are as follows:

- In order to balance huge unbalanced datasets, the study investigated cutting-edge resampling approaches and used evaluation measures (Logistic Regression, KNN, Decision Tree, Random Forest, SVM and MLP). In comparison to the body of previous works, this dataset is huge and imbalanced. Researchers in the field are confident in the data balancing method used.
- Based on a real dataset, the study used a correlation method called Pearson correlation coefficient for classifying symptoms of AD-COVID-19 patients. Sweeping attributes are used to optimize the model throughout the experiment.
- Without deleting feature subsets, the study demonstrated a reliable correlation result for identifying risk variables from already-existing, diverse feature sets related to AD-COVID-19 patients.
- The study revealed that ML models may be applied in clinical practice by providing patients with risk variables that have clear therapeutic benefits in addition to improvements in performance and accuracy.

The following sections contains (2) Literature Review, (3) Data Description, (4) Research Methodology, (5) Learning Models, (6) Results and Analysis, (7) Comparative Study, and (8) Conclusions.

2 Literature Review

According to a study as in [3], Alzheimer's disease was the sixth-leading reason of death in the U.S in 2019, and the fifth-leading cause of mortality among Americans of age 65 and older. Deaths from stroke, heart problems, and HIV decreased between 2000 and 2019, whereas recorded Alzheimer's disease mortality has climbed by more than 145%.

Many researchers have already used a machine learning-based approach to predict COVID-19 using a cough dataset [6]. This research provides coronavirus positive or negative predictions for different age groups and regions but is not able to detect which illness or symptoms affect a patient badly. This study inspires us to do further research on coronavirus and Alzheimer's patients. Furthermore, we studied about the lifestyle and situation of Alzheimer's patients. According to a study on Alzheimer's patients [7], we found the proper reasons and difficulties for dementia patients. We found an artificial-based home solution too but were unable to identify how different illnesses can take part with coronavirus-positive Alzheimer's patients. The works in [8] provided us with the idea to study different illness of coronavirus on Alzheimer's patients. In addition, the research in [9] provided us the fatality rate idea of Alzheimer's patients due to COVID-19.

In this paper, using Google collaboration, the sensitivity, accuracy, specificity, and area under the ROC curve of the comparative analysis are evaluated. We compare every illness factor with each Alzheimer's patient who is COVID-19 positive. This study focuses on the correlation between different illnesses and coronavirus-positive Alzheimer's patients. We also identify the accuracy of our research to ensure proper outcomes.

3 Dataset Description

With the World Health Organization (WHO)'s open data repository (collected from kaggle.com), this research is focused to check over 5000+ AD-COVID-19 patients worldwide to identify how different health conditions and risk factors take effect on AD-COVID-19 patients. Using Pearson's correlation coefficient, we attempt to detect infection risks based on symptoms related COVID-19 infection. Therefore, we use a survey dataset which has the information of every patient about 13 health conditions and 8 risk factors as depicted in Table 1.

Table 1. Dataset description.

Attribute	Data Type	Equivalent Data Type	Non-Null Count
Breathing problem	Object	Binary	5434 non-null
Fever	Object	Binary	5434 non-null
Sore throat	Object	Binary	5434 non-null
Runny nose	Object	Binary	5434 non-null
Dry cough	Object	Binary	5434 non-null
Asthma	Object	Binary	5434 non-null
Chronic lung disease	Object	Binary	5434 non-null
Heart disease	Object	Binary	5434 non-null
Headache	Object	Binary	5434 non-null
Diabetes	Object	Binary	5434 non-null
Hyper tension	Object	Binary	5434 non-null
Fatigue	Object	Binary	5434 non-null
Gastrointestinal	Object	Binary	5434 non-null
Abroad travel	Object	Binary	5434 non-null
Contact with COVID-19 patient	Object	Binary	5434 non-null
Attended large gathering	Object	Binary	5434 non-null
Visited public exposed places	Object	Binary	5434 non-null
Wearing masks	Object	Binary	5434 non-null
Sanitization	Object	Binary	5434 non-null
COVID-19	Object	Binary	5434 non-null

4 Research Methodology

There are five subsystems in the proposed *AD-Cov-CorrelationNet* model as shown in Fig. 1. Data categorization and characterization are covered in the first subsystem. This subsystem explains how the symptoms are divided as attribute in dataset. The second subsystem deals with how imbalanced data is processed. The optimal approach to show the data using statistical indicators has been determined in the third subsystem utilizing a variety of machine learning algorithms. The fourth subsystem addresses the correlation method used to categorize the risks of getting infection. The fifth subsystem addresses the performance evolution part and provides the accuracy, precision, recall, and F1 Score. Another subsystem connected to the fourth subsystem provides comprehensive processing of the correlation method.

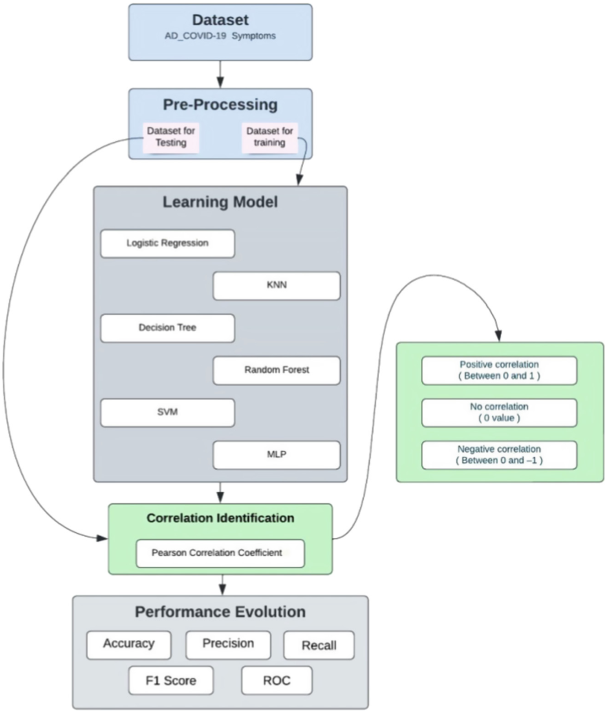

Fig. 1. The operational outline of the proposed *AD-Cov-CorrelationNet* model.

5 Learning Models

Table 2 gives a concise view of different ML algorithms used as learning models.

Table 2. Definition of ML algorithms and characterization of learning models.

Sl. No	ML Algorithm	Definition	Pros and Cons
1	Logistic Regression	In order to predict a binary outcome, logistic regression uses prior observations from a data collection	The training of logistic regression is very effective and easier to implement and analyze. If the number of data points is smaller than the number of features, logistic regression should not be used
2	K-Nearest Neighbors (KNN)	Classification and regression problems can be addressed using the supervised machine learning method known as the K-Nearest Neighbors (KNN)	It is instance-based learning. KNN is simple to use. To implement KNN, only two parameters are needed. It is unable to handle huge datasets, aware of noisy data, missing values, and outliers
3	Decision Tree	It is a method of decision support that utilizes a tree-like model to describe options and their possible results, including the possibility of chance events	Easily interpreted and understood, excellent for visual depiction. It has the ability to use both numerical and category features
4	Random Forest	A classification system made up of several decision trees is called the random forest	It is effective with non-linear data. Low probability of mistakes, and effectively uses a large dataset. Training is slow. For linear algorithms with numerous sparse features, it is not recommended
5	Support-Vector Machine (SVM)	SVMs, also referred to as support-vector machines, are supervisory learning models to analyze data for regression and classification	When there is a distinct margin of distinction, it works incredibly well. In high dimensional spaces, it works well. When we have a large data set, it does not perform as well because the training time is longer

(*continued*)

Table 2. (*continued*)

Sl. No	ML Algorithm	Definition	Pros and Cons
6	Multi-Layer Perceptron (MLP)	It is a completely connected feed-forward neural network	Ability to learn non-linear models. Real-time model learning capability (online learning). Scaling of features has an impact on MLP

6 Results and Analysis

6.1 Correlation Model Performance (Pearson Correlation Analysis)

The correlation model is used to quantify the linear relationship between two variables. It is possible for the correlation coefficient to fall between −1.0 and 1.0. The figures must not exceed 1.0 or fall below −1.0. A correlation of −1.0 denotes a perfect negative correlation, whereas a correlation of 1.0 denotes a perfect positive correlation. The performance of the correlation model is given in Table 3.

Table 3. Correlation model performance (Pearson correlation analysis).

Pearson correlation coefficient (r) value	Strength	Direction	Main Attribute	Relational Attribute (When main attribute changes relational attribute Changes too with strength and direction)	Findings
Greater than 0.5	Strong	Positive	Breathing problem, Fever, Dry cough, Sore throat, Asthma, Lung disease, Heart disease, Diabetes, Hypertension	Contact with COVID-19 patients, attended large gathering, abroad travel	Patients with direct contact with outside world are mostly suffer from COVID-19

(*continued*)

Table 3. (*continued*)

Pearson correlation coefficient (r) value	Strength	Direction	Main Attribute	Relational Attribute (When main attribute changes relational attribute Changes too with strength and direction)	Findings
Between 0.3 and 0.5	Moderate	Positive	Breathing problem, Fever, Dry cough, Sore throat, Runny nose	Asthma, Lung disease, Heart disease, Diabetes, Hypertension	Patients with serious chronic illness with COVID-19 symptoms also suffer from infection in spite of no direct contact with outside world
Between 0 and 0.3	Weak	Positive	Breathing problem, Fever, Dry cough	Diabetes, Hypertension	Patients with only Diabetes, Hypertension with mild COVID-19 symptoms also suffer from infection. But, the cases are not that significant
0	None	None	Headache, Fatigue, Gastrointestinal	Contact with COVID-19 patients, attended large gathering, abroad travel	Patients with Headache, Fatigue, Gastrointestinal problems are less likely to suffer from COVID-19 in spite of direct contact with outside

(*continued*)

Table 3. (*continued*)

Pearson correlation coefficient (r) value	Strength	Direction	Main Attribute	Relational Attribute (When main attribute changes relational attribute Changes too with strength and direction)	Findings
Between 0 and –0.3	Weak	Negative	Headache, Hypertension, Fatigue, Gastrointestinal	Diabetes, Fatigue, Breathing Problem	Patients are suffering low COVID-19 positive rate (In rare cases)
Between –0.3 and –0.5	Moderate	Negative	None	None	No correlation
Less than –0.5	Strong	Negative	None	None	No correlation

The results of correlation heat map are presented in Fig. 2.

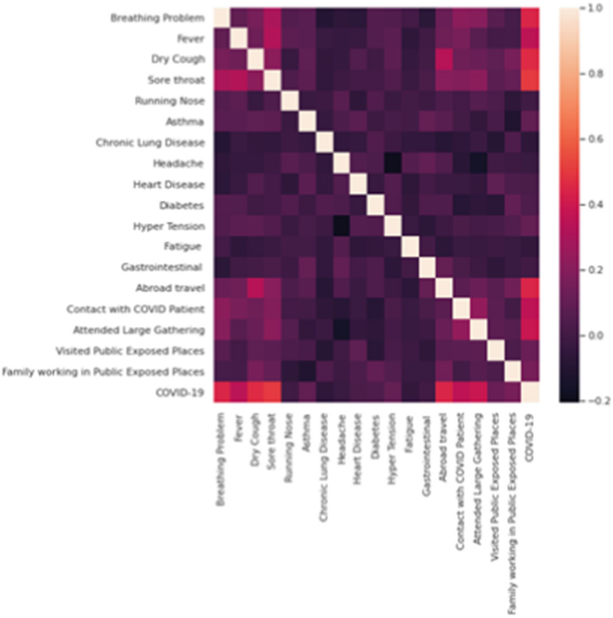

Fig. 2. Correlation heat map results (simple format).

6.2 Confusion Matrix

Confusion matrix includes data on actual and expected classifications. Four types of combination are given as follows. The number of true predictions that an event is positive is known as True Positive (TP), the number of false negatives (FN), or positive classes that are wrongly categorized as negative, is the number of improperly anticipated negative cases. The term “false positive” (FP) describes the quantity of incorrectly positive predictions made regarding a specific example, indicating that a negative class was inadvertently labeled as positive. The number of correctly predicted instances where an example is negative is known as True Negative (TN). The confusion matrix of Logistic Regression is shown in Fig. 3. Table 4 gives the measured values of all the confusion matrices.

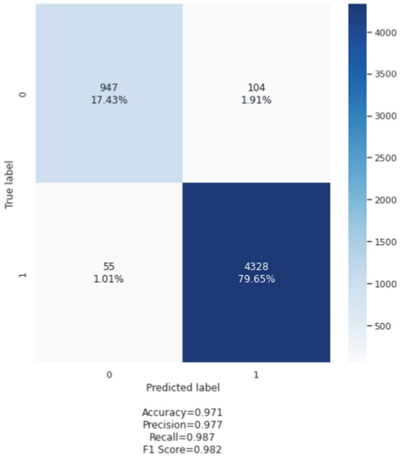

Fig. 3. Confusion matrix of Logistic Regression.

6.3 Accuracy, Precision, Recall, and F1-Score

It is important to measure the values of accuracy, recall, F1-Score, precision.

1. Accuracy: Accuracy is defined as the proportion of correct guesses in the number of projections overall.
2. Recall: True Positive Rate (TPR) or Recall is other term for sensitivity. It is a measurement for how many positive cases the classifier recognized consequently. It ought to be higher.
3. Precision: It is also known as the proportion of all positively classified instances to all positively projected cases.
4. F1 score: It is calculated using a weighted average of recollection (sensitivity) and reliability.

Figure 4 depicts the accuracy comparison of different ML classifier models.

Table 5 presents the performance comparison of different classifier models in terms of accuracy, precision, recall, and F1 score.

Table 4. Confusion matrix of six ML models.

Logistic Regression			KNN		
N=5434	Predicted (0)	Predicted (1)	N=5434	Predicted (0)	Predicted (1)
Actual (0)	TN	FP	Actual (0)	TN	FP
	947	104		896	155
Predicted (1)	FN	TP	Predicted (1)	FN	TP
	55	4328		8	4375
Decision Tree			Random Forest		
N=5434	Predicted (0)	Predicted (1)	N=5434	Predicted (0)	Predicted (1)
Actual (0)	TN	FP	Actual (0)	TN	FP
	1029	22		822	229
Predicted (1)	FN	TP	Predicted (1)	FN	TP
	72	4311		1	4382
SVM			MLP		
N=5434	Predicted (0)	Predicted (1)	N=5434	Predicted (0)	Predicted (1)
Actual (0)	TN	FP	Actual (0)	TN	FP
	963	88		958	93
Predicted (1)	FN	TP		FN	TP
	33	4350	Predicted (1)	61	4322

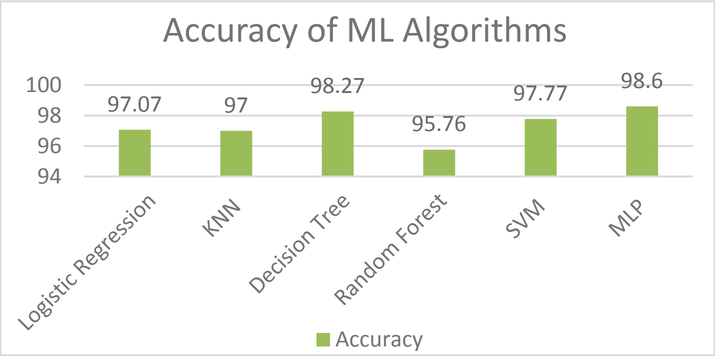

Fig. 4. Accuracy of ML Algorithms.

6.4 Receiver Operating Characteristic (ROC)

Figure 5 shows the Receiver Operating Characteristic (ROC) comparison of different ML classifiers models.

Table 5. Performance comparison of different ML Classifier models.

ML Algorithm	Accuracy	Precision	Recall	F1 Score
Logistic Regression	97.07%	97.65%	98.74%	98.19%
KNN	97.00%	96.57%	99.81%	98.17%
Decision Tree	98.27%	99.49%	98.35%	98.92%
Random Forest	95.76%	95.03%	99.97%	97.44%
SVM	97.77%	98.01%	99.24%	98.62%
MLP	98.60%	97.92%	98.61%	98.25%

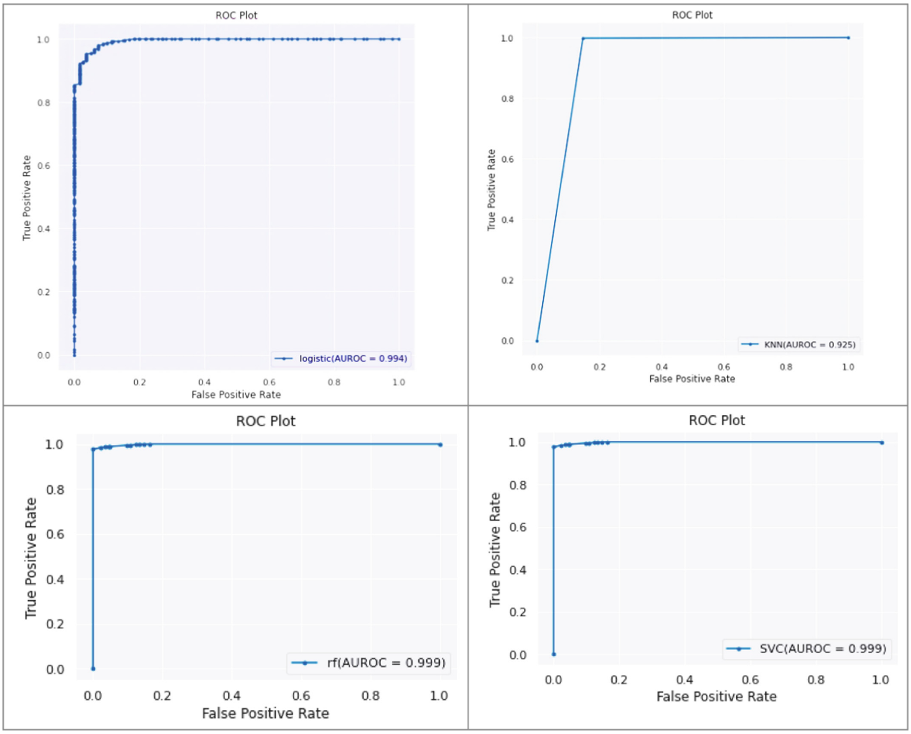

Fig. 5. ROC comparison of different ML Classifiers models.

7 Comparative Study

Table 6 compares the results of the correlation model with relevant studies and provides examples. These results show that the correlation model performed competitively in comparison to different models and studies. Nevertheless, we could not find more than one binary COVID-19 patient datasets containing AD patients and medical information for comparison. This correlation study provides around 98% accuracy from the *Ad-Cov-CorrelationNet* model.

Table 6. Comparative study.

Description	Dataset Type	Implemented Method/Algorithm	Accuracy	Reference
Predicting COVID-19	X-ray image	LSTM-RNN	96.0%	[10]
Predicting COVID-19	X-ray image	LSTM-RNN	93.0%	[11]
Predicting COVID-19	X-ray image	Res-CovNet	86.0%	[12]
Predicting AD-COVID-19 Mortality	Binary	AD-CovNet	97.0%	[9]
AD-COVID-19 symptoms correlation	**Binary dataset**	**AD-Cov-CorrelationNet**	**98.60%**	**Our work**

8 Conclusions

This study discovers substantial connections between distinct COVID-19 symptom cases and the worldwide burden of dementia. Health policymakers must have thorough plans in place to identify those at risk (including older people) and limit the risk of infection, even while paying attention to clinical and psychiatric well- being, at this key stage of the epidemic, when countries are ready to lift their national lockdown and begin opening their borders. Such patients may be prioritized based on their risk level if a vaccination becomes more broadly available. As a result, it is critical to assess the impact of COVID-19 on Alzheimer's patients' health. Whenever it comes to vaccine, Alzheimer's sufferers will be given extra attention and importance. The mortality rate of Alzheimer's patients may be lowered as a result of the research.

A comparative analysis is conducted using Google collaboration research, which has evaluated the performance of each ML technique included in the *Ad-Cov-CorrelationNet* model in terms of accuracy, precision, recall, and F1 score. The accuracy of Logistic Regression, KNN, Decision Tree, Random Forest, and SVM classification models is greater than 95%, and MLP yields the best accuracy of 98.60%. The outcomes of this study also show accurate ROC curves. A patient who is directly contacted in the outside world suffers more illnesses associated with coronavirus. Symptoms like breathing problems, fever, dry cough, and sore throat are very sensitive to COVID-19 cases. So, it is safe to stay at home for sensitive patients. Hypertension, headache and gastrointestinal are not the serious illness for Alzheimer's patients. So, symptoms with these minor problems remain in a less risky position.

Finally, the findings after the correlation study are given below:

Finding 1: AD Patients with direct contact with outside world mostly suffer from COVID-19.
Finding 2: AD Patients with serious chronic illness with COVID-19 symptoms also suffer from infection in spite of no direct contact with outside world.
Finding 3: AD Patients with only diabetes, hypertension with mild COVID-19 symptoms also suffer from infection. But, the cases are not that significant.
Finding 4: AD Patients with headache, fatigue, and gastrointestinal problems are less likely to suffer from COVID-19 in spite of direct contact with outside.
Finding 5: AD Patients are suffering low COVID-19 positive rate (in rare cases).

References

1. WHO: Coronavirus disease (COVID-19). https://www.who.int/emergencies/diseases/novel-coronavirus-2019. Accessed 17 Oct 2022
2. Wang, Q., Davis, P.B., Gurney, M.E., Xu, R.: COVID-19 and dementia: analyses of risk, disparity, and outcomes from electronic health records in the US. Alzheimer's Dement. **17**(8), 1297–1306 (2021)
3. Wiley, J.: Alzheimer's disease facts and figures. Alzheimer's Dement. **17**, 327–406 (2021)
4. Bzdok, D., Altman, N., Krzywinski, M.: Statistics versus machine learning. Nat. Methods **15**(4), 233–234 (2018)
5. Min, S., Lee, B., Yoon, S.: Deep learning in bioinformatics. Brief. Bioinform. **18**(5), 851–869 (2016)
6. Laguarta, J., Hueto, F., Subirana, B.: COVID-19 artificial intelligence diagnosis using only cough recordings. IEEE Open J. Eng. Med. Biol. **1**, 275–281 (2020)
7. Jesmin, S., Kaiser, M.S., Mahmud, M.: Artificial and internet of healthcare things based Alzheimer care during COVID 19. In: Mahmud, M., Vassanelli, S., Kaiser, M.S., Zhong, N. (eds.) Brain Informatics. BI 2020. LNCS, vol. 12241, pp. 263–274. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59277-6_24
8. Villavicencio, C.N., Macrohon, J.J., Inbaraj, X.A., Jeng, J.H., Hsieh, J.G.: Development of a machine learning based web application for early diagnosis of COVID-19 based on symptoms. Diagnostics **12**(4), 821 (2022)
9. Akter, S., et al.: AD-CovNet: an exploratory analysis using a hybrid deep learning model to handle data imbalance, predict fatality, and risk factors in Alzheimer's patients with COVID-19. Comput. Biol. Med. 105657 (2022)
10. Alassafi, M.O., Jarrah, M., Alotaibi, R.: Time series predicting of COVID-19 based on deep learning. Neurocomputing **468**, 335–344 (2022)
11. Alorini, G., Rawat, D.B., Alorini, D.: LSTM-RNN based sentiment analysis to monitor COVID-19 opinions using social media data. In: ICC 2021-IEEE International Conference on Communications, pp. 1–6. IEEE (2021)
12. Madhavan, M.V., Khamparia, A., Gupta, D., Pande, S., Tiwari, P., Hossain, M.S.: Res-CovNet: an internet of medical health things driven COVID-19 framework using transfer learning. Neural Comput. Appl. 1–14 (2021)