A Bayesian Network Model for the Prognosis of the Novel Coronavirus (COVID-19)

Salisu Aliyu[1](✉), Aminu Salihu Zakari[2], Ibrahim Adeyanju[3], and Naseer Sanni Ajoge[4]

[1] Department of Computer Science, Ahmadu Bello University, Zaria, Nigeria
aliyusalisu@abu.edu.ng
[2] Department of Computer Science, Kaduna Polytechnic, Kaduna, Nigeria
[3] Department of Computer Engineering, Federal University Oye Ekiti, Oye Ekiti, Nigeria
[4] Open Distance and Flexible E-Learning Centre, Kaduna Polytechnic, Kaduna, Nigeria

Abstract. The World Health Organization (WHO) classified the new coronavirus disease 2019 (COVID-19) as a Public Health Emergency of International Concern on January 31, 2020. It is now clear that the dreadful virus has put a significant burden on the world's healthcare systems. Currently, the main techniques for diagnosing COVID-19 are viral nucleic acid testing and chest computed tomography (CT). Though proven to be effective, these methods are time consuming hence, the need for the use of non-clinical approaches for early detection, diagnosis and prognosis of the coronavirus. Using epidemiological dataset of COVID-19 patients, this study presents the use of Bayesian network model to predict contraction of the coronavirus disease. Following the application of several structural learning techniques, a causal Bayesian network based on the nodes: breathing problem, fever, dry cough, sore throat, running nose, asthma, chronic lung disease, headache, heart disease, diabetes, hypertension, fatigue, gastrointestinal, abroad travel, contact with COVID Patient, attended Large Gathering, visited public exposed places, family working in public exposed places and COVID-19 status, was estimated. The developed BN model correctly predicted the probability of contracting COVID-19 with an accuracy of 98%. The model also affirms that, individuals with contact history with a COVID-19 patient are the most susceptible to contracting the coronavirus disease.

.

Keywords: Coronavirus · COVID-19 · Pandemic · Bayesian networks · Model

.

1 Introduction

On January 31, 2020, the World Health Organization (WHO) proclaimed the new coronavirus disease 2019 (COVID-19) a Public Health Emergency of International Concern [1]. A disease that is said to have started in Wuhan, China, has now spread to other parts of the world. According to the situation reports by the World Health Organization as at

© Springer Nature Switzerland AG 2021
O. Gervasi et al. (Eds.): ICCSA 2021, LNCS 12957, pp. 127–140, 2021.
https://doi.org/10.1007/978-3-030-87013-3_10

10:21am CEST on September 14, 2020, there have been about 28,871,176 confirmed cases of COVID-19, including 921,801 deaths globally [2].

COVID-19, an infectious disease caused by a new coronavirus strain known as severe acute respiratory syndrome coronavirus 2 (SARS-CoV2), has as its primary source of infection patients who have already been infected with the disease, with or without clinical symptoms [3]. It is believed that the virus spread through respiratory droplets (e.g. cough, sneezes etc.). In the first week, symptoms like as fever, tiredness, cough, nasal congestion, runny nose, diarrhea, and headache may appear, with further symptoms such as dyspnea, cyanosis, malaise, bad appetite, and respiratory failure appearing as the disease develops [3].

There is currently no recognized medication or vaccine that has been clinically proven to treat COVID-19, and it is clear that the virus has wreaked havoc on healthcare systems throughout the world. Currently, the main techniques for diagnosing covid-19 are viral nucleic acid testing and chest computed tomography (CT) [4]. Though proven to be effective, these methods are time consuming hence, the need for the use of non-clinical techniques such as mathematical modelling, data mining, machine learning and expert system among other artificial intelligence techniques for efficient detection, diagnosis and prognosis of the coronavirus disease.

Bayesian Networks (BNs) are effective artificial intelligence tools for modeling problems in order to obtain a better understanding and perspective on uncertainties and complexities to assist decision makers. The use of BNs in modeling health care and disease-related problems has recently received remarkable interest from the scholarly community [5]. BNs have been used in a variety of applications in healthcare, including analyzing infection and death rates, modeling and forecasting patient outcomes, risk assessment, contact tracking, and epidemic curve prediction [6–8].

Despite the growing interest of the use of BNs in modeling infectious diseases and other health care related issues, the distribution of infectious diseases modeling in Bayesian Networks literature is only 4% [5].

At this point, the need for rapid diagnosis cannot be over-emphasized. Based on available patient data, a Bayesian Network model was created to estimate the likelihood of infection with the novel coronavirus disease. The model will help health workers to make effective use of the available scarce resources together with limited information, an advantage of modelling with BNs. The model was developed with the epidemiological information obtained from the World Health Organization. We also gave a thorough examination of the prediction of contracting the coronavirus diease and the connections between the disease's causative symptoms.

The key contributions of the work are summarized below:

- Identification of possible factors that are primary causes or poses a potential risk of contracting the coronavirus disease.
- Development of a Bayesian Network (BN), a probabilistic graphical approach for calculating the risk of specific variables/attributes in predicting the contraction of the dreadful coronavirus disease.
- Conduct a series of advanced studies, such as predictive inference reasoning and sensitivity analysis, to gain a better understanding of the effect of the generated model's outcomes.

The remaining sections of the paper are organized as follows: Sect. 2 explains the purpose of this work, Sect. 3 outlines the proposed model's general approach, including data collection and preparation, followed by an explanatory note on the Bayesian Network. The experimental results are explained in detail in Sect. 4. Finally, Sect. 5 concludes the paper.

2 Motivation

The following is the description of the motivation for doing this research:

- To avoid and reduce the effects of the coronavirus disease's spread, the primary causes of contracting the coronavirus disease must be thoroughly understood. With the huge amount of data collected with regards to coronavirus disease, the causes of one being infected with the disease can be mined out of these available dataset. Hence, the need for this study.
- There has been relatively few research into the prediction of the coronavirus disease. A number of prediction algorithms have been proposed towards enhancing healthcare decision making. Hence, the need to consider one for the coronavirus disease to aid decision makers. We therefore, propose utilizing the BN structure to predict coronavirus infection in this work.
- Bayesian networks have been used to assess risk and dependability, discover fault sources, minimize the likelihood of failures, examine how management activities influence sustainability, predict uncertainty and metrics in manufacturing, categorize data, and manage projects, among other applications. In Bayesian networks, conditionals are used to estimate risk based on a series of prior events that affects the outcomes. Without performing any further computations, one may deduce the causal linkages from the structure of a Bayesian model. These benefits associated with the use of BNs informed our decision of its usage in this study.

3 Related Works

Machine Learning algorithms are currently applied in a number of research [9]. In software development, Behera, Rath [10] examined the efficiency of machine learning models in predicting software dependability and reliability. Butt, Misra [11] identifies the major causes for the failure of the agile software development approach during the COVID-19 outbreak. In sentiment analysis, Behera, Jena [12] used the Convolutional LSTM learning model on big social media dataset. For stock prediction, Behera, Das [13] used machine learning algorithms on real-time streaming dataset. In healthcare, Guhathakurata, Kundu [14] used the Support Vector Machine (SVM) classifier to accurately classified patients' conditions into moderate, severe, or no infection based on coronavirus symptoms. Tezza, Lorenzoni [15] also used SVM to predict in-hospital mortality of COVID-19 patients in Italy. A number of machine learning algorithms some of which include: Decision Tree, K-Nearest Neighbour (KNN), Naïve Bayes, Random Forest and SVM were used in Rehman, Shafique [16] to effectively diagnose COVID-19 infection. Also, in the work of Muhammad, Algehyne [17], several machine learning

algorithms were used for predicting the coronavirus disease. Using publicly accessible clinical and laboratory data, Alballa and Al-Turaiki [18] were able to review the potential of machine learning algorithms in diagnosing COVID-19 and estimating both the mortality risk and its severity. Wang, Zhai [19] utilized a Bayesian network model to investigate the impact of COVID-19 on Chinese visitors' mobility decision–making processes. Government information, traffic information, family structure, and social interaction networks are among the key factors impacting Chinese passengers' mobility, according to their results. Ojugo and Otakore [20] also used a Bayesian network to construct a model that was trained on a target system and can help anticipate the optimal parameters for classification of the novel coronavirus. In this paper, we propose the use of BN model to predict the potential risk of being infected with the coronavirus. To further validate our work, we evaluated the proposed model side by side with other machine learning techniques, including Nave Bayes, decision trees, and support vector machines.

4 Methodology

4.1 Data Collection and Description

Epidemiological dataset of COVID-19 patients based on information collected from the World Health Organization (WHO) and made available on Kaggle Website [21] was used. The dataset has 5434 instances with 21 Boolean typed attributes which include breathing problem, fever, dry cough, sore throat, running nose, asthma, chronic lung disease, headache, heart disease, diabetes, hypertension, fatigue, gastrointestinal, abroad travel, contact with COVID Patient, attended Large Gathering, visited public exposed places, family working in public exposed places, wearing masks, sanitization from market and COVID-19 status. Table 1 contains a list of the variables in the dataset used in this study as well as their definitions.

Table 1. Variables and their definitions

Nodes	Definitions
Contact with COVID Patient	Associating with an infected patient that is not cured
Chronic Lung Disease	A disorder that affects the lungs and other respiratory system
Diabetes	The body's capacity to metabolize blood glucose is impaired in this condition
Abroad travel	History of traveling

(*continued*)

Table 1. (*continued*)

Nodes	Definitions
Heart Disease	Range of conditions that affects the heart
Fatigue	A feeling of tiredness or lack of energy
Visited Public Exposed Places	Visit to public places
Hyper Tension	A condition in which the force of blood against the artery walls is too high
Asthma	A condition in which a person's airways become inflamed, narrow, swollen and produces extra mucus making it difficult to breath
Sore throat	Pain and irritation of the throat
Breathing Problem	Trouble or difficulty breathing
Fever	Increase in body temperature above normal
Dry Cough	Cough that does not bring up any mucus
Running Nose	Persistent watery discharge from the nose
Headache	Pain or discomfort in the head
Gastrointestinal	Diseases or disorder of digestive tract
Attended Large Gathering	Present in a crowded environment
Family working in Public Exposed Places	Associating with people that are publicly exposed

4.2 Data Preparation

The dataset was prepped and sanitized such that only relevant attributes were chosen. Irrelevant attributes like wearing masks and sanitization from market were removed because they are constant values. Constant values are generally not required in learning BN structures. For example, if an attribute y takes the same value in each of the records, then it cannot be a predictor for any other at-tribute. Fig. 1 depicts the frequency of each dataset attribute.

4.3 Bayesian Networks

A Bayesian Network (BN) is a directed acyclic network in which each node is tagged with quantitative probability information [10]. It is an efficient encoding of a domain's probabilistic model. BN are essentially used to indicate the dependencies between variables/attributes. Each node in a BN corresponds to a random variable, which might be discrete or continuous. A pair of nodes is connected by a set of directed links or arrows. If an arrow connects node X to node Y, X is said to be Y's parent. Each node Y_i has a conditional probability distribution $P(Y_i|Parents(Y_i)$ that measures the influence of the parents on the node.

The network's topology defines the conditional independence relationships that exist in a given domain (see Fig. 2).

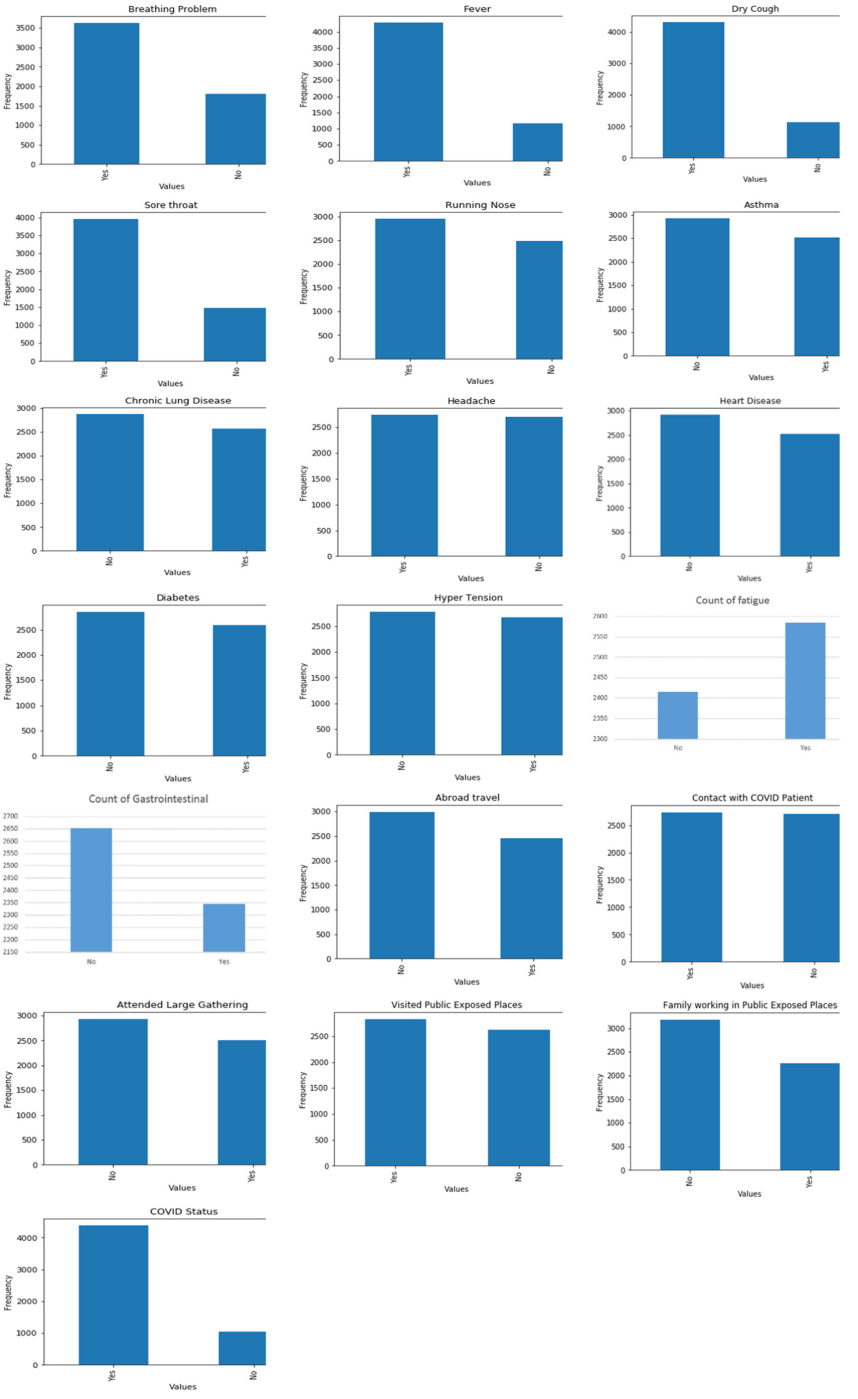

Fig. 1. Frequency of attributes used



Fig. 2. Topology of a Bayesian Network

From Fig. 2, node X has a direct influence on node Y. We often view the result of an unknown source as evidence, and we want to figure out what the cause is? In that case, Bayes' rule is expressed as:

$$P(cause|effect) = \frac{P(cause|effect)P(cause)}{P(effect)} \tag{1}$$

In Eq. 1, the conditional probability P(effect|cause) quantifies the causative link, whereas P(cause|effect) specifies the diagnostic direction. We often have conditional probabilities on causal links in tasks like medical diagnosis (that is, the doctor knows P (symptoms|disease) and wants to derive a diagnosis, P(disease|symptoms).

The domain expert is generally the one who determines what direct impact exists in a domain. We only need to describe a conditional probability distribution for each variable given its parents after the topology of the Bayesian network has been established.

There are basically two ways of building a BN structure/model:

i Manually- where an expert determines the direction of arrows/edges between nodes and their associated Conditional Probability Tables (CPT).
ii Structural Learning- where the direction and location of arrows/edges are determined by means of some learning algorithms.

In this study, the structural learning approach was used. There are several structural learning algorithms, among which include the Peter-Clark (PC) algorithm [22], Bayesian Search (BS) algorithm [23], Greedy Thick Thinning (GTT) algorithm [24] etc. The logscore in equation 2 by [24] was chosen as the model selection criteria since it assesses the performances of the candidate BN models developed by the structural learning techniques.

The logscore is calculated as follows:

$$logscore(X_1, \ldots, X_N) = \frac{\sum_{i=1}^{N} \log_2 P(X_i|model)}{nN} \tag{2}$$

where X_i represent the nodes, N represents the total number of levels, and n represents the number of variables in the model. The BN model with the highest logscore is selected.

Sensitivity analysis was also done to see how changes in the levels of other nodes influence a node in a network.

5 Experimental Findings

In this section, we discuss the development of the Bayesian model for predictive inference reasoning. The GeNie [15] software was used in the model development process. The

Bayesian Search (BS) Algorithm [12], the Greedy Thick Thinning (GTT) algorithm [13] and the Peter-Clark (PC) algorithm [11] were employed to build different BN structures. In other to select the best algorithm to use for the current problem, the logscore for each algorithm is computed. The algorithm with the maximum logscore value is in turn selected. The logscore of each of the BN structure learning methods are shown in Table 2.

Table 2. Algorithms for Structural Learning and their associated Logscore Values

Algorithms	Logscore
Greedy Thick Thinning	−46210.7
Bayesian Search	−46601.8
Peter-Clark	−53828.5

Table 2 shows that the GTT method has the greatest logscore, indicating superior estimate performance. As a result, the BN estimated by the GTT method was chosen. The BN structural model is generated is as shown in Fig. 3. Upon completion of parameter modelling and quantification in the process of parameter learning from the dataset used, Fig. 4 shows the estimated BN with probabilities associated with each node in.

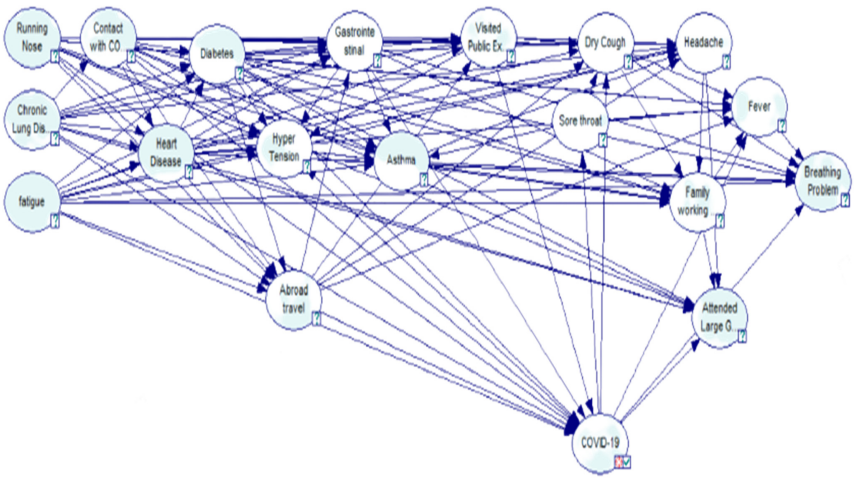

Fig. 3. GTT Algorithm-based Bayesian network structural model

The estimated BN in Fig. 4 can be used to predict the possibility of one being infected with the deadly coronavirus base on evidences observed. It can as well be used to analyze the factors causing one to be infected with the coronavirus and the relationships between those factors.

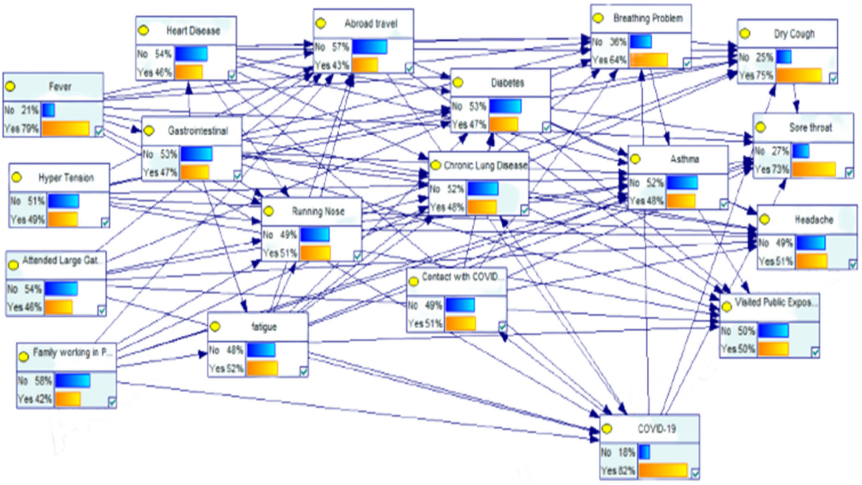

Fig. 4. GTT Algorithm-based Bayesian network model estimation

5.1 Validation of the Model

The developed model's prediction ability was verified using a measure of prediction accuracy and the Receiver Operating Characteristic (ROC) curve. Furthermore, a sensitivity analysis was carried out to determine which nodes are most likely to influence the likelihood of a patient contracting the deadly coronavirus disease.

Accuracy, as stated in Eq. 3, is the proportion of instances in the dataset correctly classified for the model. This is mathematically expressed as:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{3}$$

where *TP* denotes True Positive, TN denotes true negative, FP denotes false positive and FN denotes false negative.

The developed BN model correctly predicted the probability of contracting COVID-19 with an accuracy of 98%.

The ROC curve is a graph that compares the true positive rate (sensitivity) against the false positive rate (1- specificity). The area under the ROC curve is referred to as the Area Under the Curve (AUC). AUC values vary from 0 to 1. An AUC value of 1 indicates that the prediction is error-free, and as the AUC score drops, the model's prediction performance diminishes. Figure 5 depicts the ROC curve for the BN model prediction of contracting the coronavirus disease with an AUC of 0.997815.

Finally, a sensitivity analysis was performed to test the created BN model and assess the influence of additional variables/attributes (nodes) on the target node's result (COVID-19 status). In this study, the aim was to identify which node have the greatest impact on the probability of been infected by the deadly coronavirus.

Table 3 demonstrates how changes in the levels of other nodes impact the node "COVID-19". With a sensitivity value of 0.142, the most critical factor for predicting the contraction of the coronavirus disease is coming into contact with a COVID-19

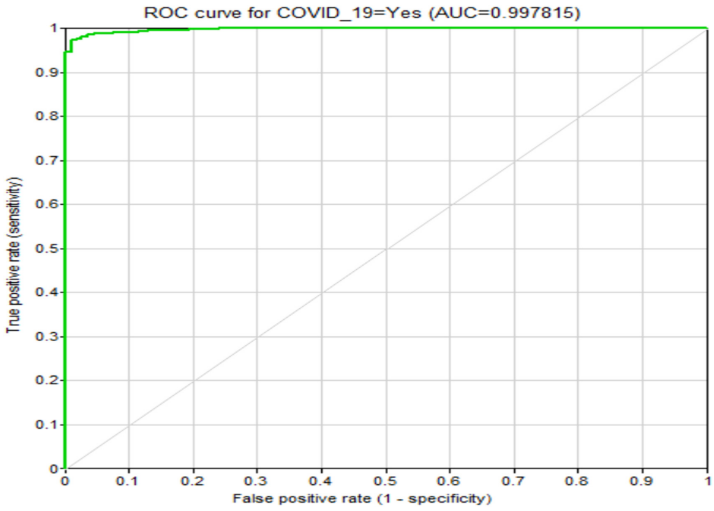

Fig. 5. ROC curve for the Bayesian network model.

patient, as shown in Table 3. This result is in accordance with the findings of [1, 3], the generally acknowledged views from centres of disease control across the globe and more importantly the World Health Organization.

This study hence concluded that patients with chronic lung diseases are the second category of people that are more susceptible to contracting the coronavirus disease. The third factor is diabetic patients followed by people with history of travelling abroad.

5.2 Other Machine Learning Algorithms

In addition to Bayesian Network, this study looked at the Nave Bayes Algorithm, J48 Algorithm, and Support Vector Machines, which are three common supervised machine learning techniques. We look at each machine learning process as thoroughly as possible in order to compute the best outcomes.

The Nave Bayes (NB) Algorithm is a popular machine learning approach based on the Bayes Rule. This approach is based on the independence of variables. The classes that would be calculated using the NB system must be unrelated to one another. One of the supervised learning algorithms is this one. Despite its simplicity, it yields excellent results in medical applications.

One of the most popular decision tree algorithms is J48. It also has the advantage of being able to make predictions using a smaller tree compares to other decision trees. As a result, the J48 algorithm is able to achieve more efficient results than its competitors.

Support Vector Machine (SVMs) are computational algorithms that use statistical learning theory to generate a coherent estimator from available data. It seeks to categorize the information into two groups. For this cause, the n-dimensional hyperplane is formed.

The WEKA software was used to develop programs for other machine learning algorithms used for comparison as shown in Table 4.

Table 3. Sensitivity analysis results using "COVID-19" as the target node.

Nodes	Sensitivity
Contact with COVID Patient	0.142
Chronic Lung Disease	0.046
Diabetes	0.022
Abroad travel	0.021
Heart Disease	0.018
Fatigue	0.018
Visited Public Exposed Places	0.012
Hyper Tension	0.000
Asthma	0.000
Sore throat	0.000
Breathing Problem	0.000
Fever	0.000
Dry Cough	0.000
Running Nose	0.000
Headache	0.000
Gastrointestinal	0.000
Attended Large Gathering	0.000
Family working in Public Exposed Places	0.000

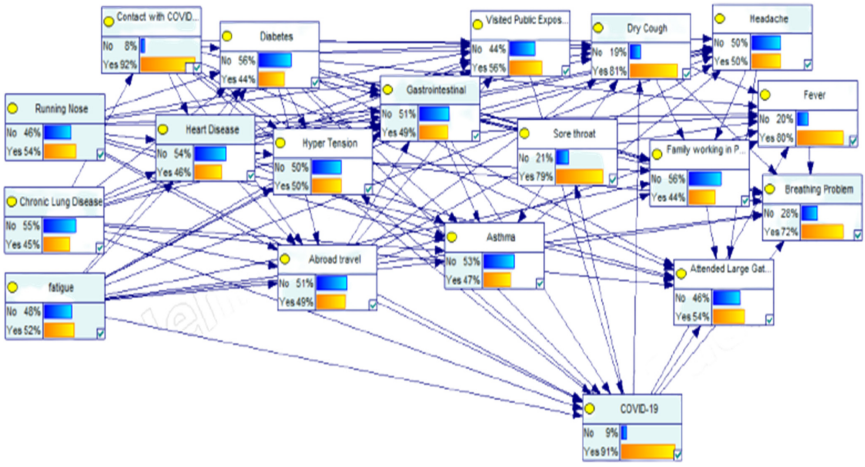

Fig. 6. The state of the Bayesian network in a situation where contact with a COVID-19 patient is estimated to be 92%.

Table 4. Comparison of machine learning algorithm

Algorithm	Accuracy	AUC
Naïve Bayes	96.5%	0.99
J48	98.2%	0.99
SVM	96.8%	0.91
Bayesian Network	**98.4%**	**0.99**

The Naïve Bayes, J48 Decision Tree, Support Vector Machine learning algorithms all correctly predicted the contraction of the coronavirus with an accuracy of 96.5%, 98.2% and 96.8% respectively.

The estimated BN depicting the level of probabilities of each node can be used to determine the patients that are more susceptible to contracting the coronavirus disease. For instance, given the evidence that there is 92% belief that a patient came in contact with another COVID-19 patient, the BN in Fig. 6 shows that there are 91% chances that the patient will likely contract the coronavirus disease. So, based on evidences provided by a patient, the BN model continues to adjust the probability of contracting the virus.

6 Conclusion

In the present study, a BN model was developed for estimating the probability of contracting the deadly coronavirus using epidemiological dataset. The BN quantifies the impact of several factors (attributes) on the probability of contracting the virus. With a 98% accuracy rate, the developed BN model successfully predicted the likelihood of contracting the coronavirus. Our model also affirms that contact with infected patient remains the sole cause for contracting the coronavirus disease. This is evident from the sensitivity analysis conducted which shows that contact with an infected patient has the highest sensitivity value of 0.142. The developed model can be used proactively to contain and manage the spread of the coronavirus in a cost effective manner. Furthermore, the BN model may be used to conduct scenario analysis in order to aid in the creation of a risk management strategy. In addition, the model may be modified to incorporate other features that are presently not captured, allowing for more effective tracking of vulnerable individuals and, as a result, more attention to be directed toward them. Finally, the proposed approach, as opposed to the existing random testing techniques, would aid in the early diagnosis of COVID-19 cases, decreasing the burden on healthcare institutions.

Funding:. No funding sources.

Compliance with Ethical Standards.

Conflict of Interest:. The authors declare that they have no conflict of interest

Ethical Approval:. This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Zhou, M., Zhang, X., Qu, J.: Coronavirus disease 2019 (COVID-19): a clinical update. Front. Med. **14**(2), 126–135 (2020). https://doi.org/10.1007/s11684-020-0767-8
2. World Health Organization, Coronavirus disease (COVID-19) situation reports (2020). https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/.
3. Chen, Z.-M., et al.: Diagnosis and treatment recommendations for pediatric respiratory infection caused by the 2019 novel coronavirus. World J. Ped. 1–7 (2020). https://doi.org/10.1007/s12519-020-00345-5
4. Wynants, L., et al.: Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. BMJ **369** (2020). https://doi.org/10.1136/bmj.m1328
5. McLachlan, S., et al.: Bayesian networks in healthcare: distribution by medical condition. Artif. Intell. Med. **107**, 101912 (2020). https://doi.org/10.1016/j.artmed.2020.101912
6. Fenton, N.E., et al.: COVID-19 infection and death rates: the need to incorporate causal explanations for the data and avoid bias in testing. J. Risk Res. **23**, 1–4 (2020)
7. Fenton, N., et al.: A privacy-preserving Bayesian network model for personalised COVID19 risk assessment and contact tracing. medRxiv (2020)
8. Ucar, F., Korkmaz, D.: COVIDiagnosis-Net: deep bayes-squeezenet based diagnostic of the coronavirus disease 2019 (COVID-19) from X-ray images. Med. Hypoth. 109761 (2020). https://doi.org/10.1016/j.mehy.2020.109761
9. Jena, M., Dehuri, S.: DecisionTree for classification and regression: a state-of-the art review. Informatica **44**(4) (2020). https://doi.org/10.31449/inf.v44i4.3023
10. Behera, R.K., Rath, S.K., Misra, S., Leon, M., Adewumi, A.: Machine learning approach for reliability assessment of open source software. In: Misra, S., et al. (eds.) ICCSA 2019. LNCS, vol. 11622, pp. 472–482. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-24305-0_35
11. Butt, S.A., Misra, S., Anjum, M.W., Hassan, S.A.: Agile project development issues during COVID-19. In: Przybyłek, A., Miler, J., Poth, A., Riel, A. (eds.) LASD 2021. LNBIP, vol. 408, pp. 59–70. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-67084-9_4
12. Behera, R.K., et al.: Co-LSTM: Convolutional LSTM model for sentiment analysis in social big data. Inf. Process. Manage. **58**(1), 102435 (2021)
13. Behera, R.K., et al.: Comparative study of real time machine learning models for stock prediction through streaming data. J. Univ. Comput. Sci. **26**(9), 1128–1147 (2020)
14. Guhathakurata, S., et al.: A novel approach to predict COVID-19 using support vector machine. In: Data Science for COVID-19, pp. 351–364. Elsevier (2021)
15. Tezza, F., et al.: Predicting in-hospital mortality of patients with COVID-19 using machine learning techniques. J. Person. Med. **11**(5), 343 (2021)
16. Rehman, M.U., et al.: Future forecasting of COVID-19: a supervised learning approach. Sensors **21**(10), 3322 (2021)
17. Muhammad, L., et al.: Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset. SN Comput. Sci. **2**(1), 1–13 (2021)
18. Alballa, N., Al-Turaiki, I.: Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: a review. Inf. Med. Unlocked **24**, 100564 (2021). https://doi.org/10.1016/j.imu.2021.100564
19. Wang, J., Zhai, X., Luo, Q.: How COVID-19 impacts chinese travelers' mobility decision-making processes: a bayesian network model. In: Wörndl, W., Koo, C., Stienmetz, J.L. (eds.) Information and Communication Technologies in Tourism 2021, pp. 557–563. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-65785-7_53
20. Ojugo, A., Otakore, O.D.: Forging an optimized bayesian network model with selected parameters for detection of the coronavirus in Delta State of Nigeria. J. Appl. Sci. Eng. Technol. Educ. **3**(1), 37–45 (2021)

21. Symptoms and COVID Presence (2020). https://www.kaggle.com/hemanthhari/symptoms-and-covid-presence.
22. Spirtes, P., Glymour, C.: An algorithm for fast recovery of sparse causal graphs. Soc. Sci. Comput. Rev. **9**(1), 62–72 (1991)
23. Cooper, G.F., Herskovits, E.: A Bayesian method for the induction of probabilistic networks from data. Mach. Learn. **9**(4), 309–347 (1992)
24. Cheng, J., Bell, D.A., Liu, W.: An algorithm for Bayesian belief network construction from data. in proceedings of AI & STAT'97. Citeseer (1997)