Researchers Journal of Science and Technology (REJOST)

SUPERVISED MACHINE LEARNING MODEL FOR EFFECTIVE CLASSIFICATION OF PATIENTS WITH COVID-19 SYMPTOMS BASED ON BAYESIAN BELIEF NETWORK

Anietie Ekong[1]*, Blessing Ekong[1], Anthony Edet[1]

Received: May 30th, 2022 / Accepted: June 3rd, 2022

ABSTRACT

Covid-19 is a contagious infection and should be managed properly. If it is identified early enough, the chances of survival are high. Symptoms presentations may be confusing to health practitioners since they are similar to other diseases and the number of health professionals are mostly insufficient, especially in developing countries, hence a correct and timely diagnoses based on symptoms' presentations can prove difficult. So far, efforts to take address these needs are still not adequate. In this paper, we leverage on the efficacy of machine learning algorithms to present a machine learning approach, Bayesian Belief Network, that aims at ensuring that the symptoms' set are correctly and timeously classified as either Covid-19^{+} or Covid-19^{-}. A dataset is gathered from observed symptoms of confirmed Covid-19 patients by healthcare practitioners. The dataset is trained so as to be able to classify, based on this prior knowledge, any supplied symptom set. This has been able to effectively handle the problem of wrong or delayed diagnoses with its attendant negative consequences. Our model yields 98% accuracy, showing a significant classification accuracy of patients with Covid-19 symptoms.

Keywords: Bio-inspired computing, Computer aided diagnosis, Bayesian Network, Machine learning, Covid-19.

Anietie Ekong *anietieekong@aksu.edu.ng*

[1]Department of Computer Science, Akwa Ibom State University, Mkpat Enin, Akwa Ibom State, Nigeria.

Introduction

According to Ekong (2021), the mortality rate from diseases as a result of incorrect or delayed diagnoses is high and worsened by high illiteracy level and lack of adequate medical personnel, particularly in low income countries.

A Bayesian network is a model that represents a joint probability distribution of a set of random variables with a possible mutual causal

relationship. It consists of different parameters which are used as random variables. The network is made up of different nodes representing the different random variables, edges that exist between pairs of nodes representing the causal relationship of these nodes and a conditional probability distribution in each of the nodes Rupesh *et al.* (2018). The main objective of using Bayesian Belief Network centers on making a framework that frames the posterior conditional probability distribution of outcome which are also referred to as causal variable(s) in practical approaches after new samples or evidences have been recorded. In Bayesian statistics, the posterior probability of an event is described as the updated probability of that event occurring after taking into consideration new information that have resulted.

Statistically, the posterior probability of an event 'A' is the probability of that event happening, given that another event 'B' has occurred. The mathematical model known as Bayes' theory is given in equation 1;

$$P(A|B) = (P(B|A)*P(A))/P(B) \qquad (1)$$

P(A|B) is the posterior probability of event A, P(B|A) is the likelihood of event B being the cause of the occurrence of event A, P(A) is the prior probability of event A while P(B) is the evidence that event A is true. Covid-19 has ravaged the world since 2019 causing the death of millions of people across the globe. Many methods have been employed by different countries to diagnose, treat or contain it. People that contact this disease present different symptoms some of which are also present in other diseases such as Malaria. As a result of this similarity in symptoms, people, including health practitioners have been finding it difficult to differentiate between Covid-19 and other diseases especially before the laboratory confirmatory test is done. This problem has been the major challenge this work has been modeled to solve. We design a Bayesian Belief Network (BBN) model to carry out the classification of medical tests performed on suspected Covid-19 patients to either COV^+ or COV^-, depending on the observed conditional variables, which in this case are the observed symptoms. The model in this work learns from the dataset which we provide and gives a concrete prediction that a test tuple is either CoV^+ or CoV^-.

Motivation

Two issues form our motivation in this work. Medical practitioners across board become unsure of what symptoms identify a patient as having been infected with Covid-19. This leads to delayed or outrightly wrong diagnoses implying that positive cases may be classified as negative and vice versa. This can lead to administering wrong treatment with its attendant negative consequences including but not limited to worsening health condition, unnecessary expenses in the wrong direction and possibly death.

Related Literature

Yazeed and Shomron (2020) proposed a Covid-19 diagnosis prediction by symptoms of tested individuals using a machine learning approach. The work classified test results based on age, previous test result of the patient and contacted persons. Heet *et al.* (2020) proposed a prediction and diagnosis of Covid-19 using Machine Learning Algorithms. The work focused on Linear Regression and Support Vector Machine (SVM) models for determining the curve of active cases. To diagnose, the developed application had certain questions that needed to be answered by the person under test. Based on these answers, the (K-Nearest neighbours) KNN model provides the maximum likelihood result of a person being infected or not.

Mustafa *et al.* (2021) carried out a research using Artificial Neural Networks. The work aimed at revealing scientific methods to aid carrying out effective testing for Covid-19. Arun *et al.* (2020) worked on Artificial Intelligence-Based Classification of chest X-ray images into Covid-19 and other infectious diseases. The worked studied X-ray images and classified them into CoV^{+} and other infectious diseases.

Saket *et al.* (2021) studied the severity of Covid-19 disease on cancer patients using purposefully designed machine learning approach. The work revealed three categories of Covid-19 infected cancer patients and how they react to the disease. They showed that some had severe symptoms early, others after three days while others did not have symptoms.

Vafa *et al.*(2021) proposed the use of machine learning approach that combined with some laboratory tests and vital signs to fairly infer that a patients was infected with ARSCoV-2.

Artika (2022) compared the Decision Tree and Logistic Regression Machine Learning Classification Algorithms with respect to the classification of patients with symptoms of diseases suspected to be Covid-19. The research aimed at classifying any test data to either CoV+ or CoV-. The author used, at varying times, two machine learning models namely; Logistic Regression and Decision Tree Algorithms. This was done in an attempt to determine which model gives more accurate result. The Decision tree model outperformed Logistic Regression model at the conclusion of the research. Farahat et al., 2020 proposed an automated diagnostic system using 3D Computer tomographic images to accurately diagnose Lung Function in Coronavirus Patients. He stressed that early discovery and grading of coronavirus disease 2019 (COVID-19), as well as the use of ventilator support machines of immense importance in combating COVID-19 and reduces mortality rate greatly. This system used an unsupervised approach called Appearance Model to segment the lung region from chest CT scans, and then applied 3D rotation invariant Markov–Gibbs Random Field MGRF)-based morphological constraints. He had 91.6% accuracy.

Aires et al., 2022 proposed an algorithm that classifies COVID-19 patients and differentiate severe COVID-19 patients from non-severe patients based on thromboelastometry parameters. He established that machine learning algorithms can help in this classification. However, he did not go into the extent and how accurate that was possible.

In all these either the degree of accuracy wasn't comfortable for health care practitioners to completely rely on them or the study dealt with the after-effect of contacting the disease and not the disease itself.

Methodology

We propose a Machine Learning Model for Classification of symptoms set as Covid-19 using Bayesian Networks.

The following steps were followed;

1. A unique ID was assigned to each of the 15 patients under consideration.
2. The irrelevant or noise data were removed. Some inconsistent data were corrected and re-entered.
3. The data was transformed into a format that could be understood by the computer to ensure that the data is of good quality before applying the Bayesian Model.
4. The Bayesian Belief Network (BBN) algorithm, a Machine Learning algorithm that is used for classification is applied. It works on the notion of joint probability distribution of a set of random variables with a possible mutual

causal relationship.

5. The system outputs the likelihood of an occurrence based on the lesson it learnt after training on the supplied dataset.

The necessary details of the above steps are explained in the ensuing paragraphs

Given that a patient has Covid-19, the observed evidences are Fever (F), Headache (H), Cough(C), Sour Throat (ST), Shortness of breath (SB), Lost of taste(LT), Sneezing (S). Having known the symptoms of Covid-19, the Directed Acyclic Graph (DAG) of the problem is shown in figure 1;

The symptoms have been reduced to only their initial letters or first 2 letters in the directed acyclic graph.

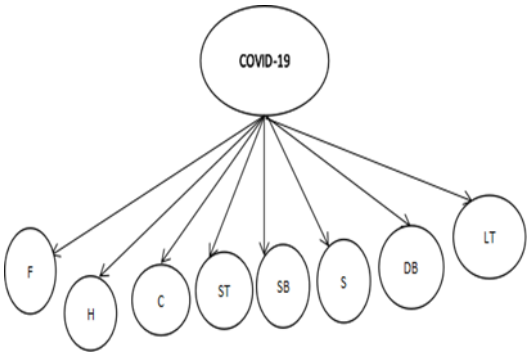

Fig. 1 Directed Acyclic Graph of Covid-19 symptoms

We model a Bayesian network under the prior hypothesis that a patient has Covid-19 given the above observed symptoms. There is another assumption that the patient is not Covid-19 patient.

These assumptions are presented as likelihoods such that $P(X_i|CoV+)$ represent the likelihood of a positive case and $(X_i|CoV-)$ represents the likelihood of a negative case, where i belongs to a set of natural numbers (n) from 1-8 representing symptoms. The probability of the patient testing positive for Covid-19 is P(CoV+) and the probability of testing negative for Covid-19 is as in equation 1.

$$P(CoV-) = 1-P(CoV+) \quad (2)$$

The Joint and conditional probability of the DAG can be given as follows;

P(F,H,C,ST,SB,S,DB,LT)

Using Bayesian Theory, we express our model as follows;

$$P(CoV+|X_i)= P(X_i|CoV+).P(CoV+)/P(X_i) \quad (3)$$

where $P(X_i|CoV+)$ is likelihood of the patient testing positive for Covid-19 and P(CoV+) is the prior probability of the patient testing positive for Covid-19 and the denominator $P(X_i)$ being evidence can be expressed as

$$P(X_i)=P(X_i|CoV+)P(CoV+)+P(X_i|CoV-)P(CoV-) \quad (4)$$

For our system, we have two classes namely, class - CoV+ and class - CoV-. Given the dataset's row running from

$X = (x_1,x_2,x_3,\ldots,x_n)$ and column running from A $= (a_1,a_2,a_3,\ldots,a_n)$.

Our prior probability may be estimated as $P(CoV+) = D_i/D$. Here D is the total number of training samples (row wise) and D_i is the total number of training sample of class - CoV+.

We can build Bayesian network by solving it out manually using Bayes' theorem or by the learning approach.

Machine Learning is used to construct Bayesian network for this problem. The dataset in table 1 is gathered from experts review of symptoms that suspected cases exhibits in Fresh Breath Specialist Hospital, Uyo and is found to be consistent with online review of the symptoms that Covid-19 patients exhibit. Each of the data input the exhibited symptoms represented by 0 or 1 which indicates the presence or absence of a symptom and was collected by different medical experts Though there are other symptoms, these selected sets were sufficient to objectively and correctly

classify the symptom set as either Covid-19+ or Covid -19-.

Using this training dataset and python as our tool, the data is pre-process and then trained to enable it classify correctly. The training set is fed into the system in in Comma separated variable (CSV) format.

Table 1 – Training Dataset

S/N	FEVER	HEADACHE	COUGH	SOUR THROAT	S.BREATH	SNEEZING	L.TASTE	CLASS
1	1	1	0	0	0	1	0	CoV-
2	1	0	1	1	1	1	1	CoV+
3	0	1	1	1	0	1	1	CoV+
4	0	0	0	0	0	0	1	CoV-
5	1	1	1	0	0	1	0	CoV-
6	0	0	0	1	1	1	1	CoV+
7	0	0	0	1	1	0	0	CoV+
8	1	1	1	1	0	0	0	CoV-
9	1	1	0	0	1	1	1	CoV+
10	0	1	0	0	0	1	0	CoV-
11	0	0	0	1	1	0	1	CoV+
12	0	0	0	0	1	0	1	CoV+
13	1	0	1	1	0	0		CoV-
14	1	1	1	1	0	0	0	CoV-
15	1	0	1	1	1	1	0	CoV+
X	0	0	0	1	1	0	1	?

From the dataset in table 1, the probabilities of symptoms occurrence are calculate thus; P(CoV+) =Di/D i.e. 8/15 = 0.533. P(CoV-) = Di/D i.e. 7/15 = 0.466. From the table, out of 15 entries, CoV+ = 8 while CoV- = 7 and we express the individual probability of the most important symptoms of Covid-19 which are (Sour Throat (S.throat), Shortness of Breath (S.Breath), Loss of Taste (L.Taste) and summarize in table 2

We can determine the status of a new tuple or person say X with symptoms (F=0, H=0,C=0,ST=1,SB=1,S=0,LT=1); where 0 represents the absence of the symptom that 0 is assigned and 1 represents the presence of the particular symptom that 1 is assigned by using these values and the associated probabilities as follows;

$P(X|CoV+)=6/8*7/8*6/8 = 0.4921875$

$P(X|CoV-) = 3/7*0/7*3/7 = 0.00$

Since X will either be CoV+ or CoV-, it follows that;

$P(X) = 0+0.44296875 = 0.44296875 \approx 0.44$

Table 2: Probabilities of the observed symptoms

Attribute	SYMPTOMS	Count		Probabilities	
		CoV+	CoV-	Cov+	Cov-
Principal Symptoms	Loss of Taste	6	3	6/8	3/7
	Shortness of Breath	7	0	7/8	0/7
	Sour Throat	6	3	6/8	3/7

Results and Discussion

From our classification process using Bayesian Networks Library in python, we have the result as shown in table 3 after training of the dataset in table 1.

Table 3: Computer generated prediction

CoV+	0.5750
CoV-	0.3500
?	0.0750

From table 3, 0.575 of the input data, corresponding to 57.5%, are classified as CoV+ while 0.35 of the data, corresponding to 35%, are CoV-. Incorrectly classified is 0.075 i.e 7.5%.

In classification, the test dataset always follows the pattern of the training set. In our training set, our model has been trained to classify all new COVID-19 test data tuple as CoV+ if the result is within 0.575- 1.00 i.e. 57.5% - 100% and as CoV- if the result is within 0.35 - 0.574 i.e. 35% - 57.4%. To test our model after training it, we an input marked as X in table 1 which has the following symptom set;
(F=0,H=0,C=0,ST=1,SB=1,S=0,LT=1) where F, C, S, SB, S and LT represents Fever, Cough, Sour Throat, Short of Breath, Sneezing and Lost of Tastes respectively. The value (1) represents the presence of that symptom while (0) represents the absent of the symptom.

Our Bayesian Network has classified the test data to be CoV- since the output for X is 0.44 i.e. 44%. This result is within 0.35-0.574 (35%-57.4%) range.

Conclusion

We have developed a machine learning model to classify Covid-19 symptoms set to be positive or negative. Our model has given us a good level of accuracy at 98%, hence, reducing the overall stress experienced by front line health workers and reducing the possibility of false or delayed diagnoses of Covid-19 cases to its lowest level.

References

Aires R., Soares A, Gomides A, Nicola A., Teixeira A, da Silva D (2022). Thromboelastometry demonstrates endogenous coagulation activation in nonsevere and severe COVID-19 patients and has applicability as a decision algorithm for intervention. *PLoS ONE 17(1)*: e0262600. https://doi.org/10.1371/journal. pone.0262600.

Artika Arista, (2022). Comparison Decision Tree and Logistic Regression Machine Learning Classification Algorithms to determine Covid-19. *Jurnal dan Penelitian Teknik Informatika,7,(1)*,59-65.

Arun, Sheeba,& Dinesh, (2020). Artificial Intelligence-Based Classification of Chest X-Ray Images into Covid-19 and Other Infectious Diseases. *International Journal of Biomedical Imaging*.

https://doi.org/10.1155/2020/8889023

Ekong A., Odikwa H., Ekong O.(2021). Minimizing Symptom-based Diagnostic Errors Using Weighted Input Variables and Fuzzy Logic Rules in Clinical Decision Support Systems. *International Journal of Advanced Trends in Computer Science and Engineering, 10(3)*, 1567 – 1575. https://doi.org/10.30534/ijatcse/2021/121032021

Farahat I., Sharafeldeen A., Elsharkawy, M., Soliman, A., Mahmoud, A., Ghazal M., Taher F., Bilal M., Abdel R., Aladrousy W.(2022), The Role of 3D CT Imaging in the Accurate Diagnosis of Lung Function in Coronavirus Patients. Diagnostics. https://doi.org/ 10.3390/diagnostics12030696.

Heet S., Vruddhi M., Ramchandra M. (2020). Prediction and Diagnosis of Covid-19 using Machine Learning Algorithms. *International Journal of Recent Technology and Engineering*, 9(3),678-683.

Mustafa K., Syed Fasih A., Fariha I., Asma T., Rabia A., (2021). Identification of efficient Covid-19 diagnostic test through artificial neural networks approach – substantiated by modeling and simulation. *Journal of Intelligent Systems*. https://doi.org/10.1515/jisys-2021-0041, 836–854.

Rupesh A., Amir F., Roderick D., Linda C., Gerben D., Monika H., Aly K., Habil Z.(2018). Applications of Bayesian network models in predicting types of hematological malignancies,8(1),1-2.

Saket N., Sejal M., Rocio P., Allen Z. and Ying T., (2021). Projecting Covid-19 disease severity in cancer patients using purposefully designed machine learning, Infectious Disease, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York.

Vafa B., Steven P. ,Chong L., Hemal P., Joel M., Farshid S., Payam E., Mark H. (2021). A Severe Acute Respiratory Syndrome Coronavirus 2 (SARSCoV-2): Prediction Model From Standard Laboratory Tests, Infectious Disease Society of America, DOI: 10.1093/cid/ciaa1175

Yazeed Zoabi, Noam Shomron, (2020), Covid-19 diagnosis prediction by symptoms of tested individuals: a machine learning approach,doi: https://doi.org/10.1101/2020.05.07.20093948.