Classification and Forecasting Model for Covid -19 Disease Severities based on Medical Diagnosis using Weighted Average Dynamic Time Warping Technique

Gopalakrishnan B¹, Manikantan M² and Purusothaman P³

¹Associate Professor, Department of Information Technology,
Bannari Amman Institute of Technology, Sathyamangalam, Erode, Tamilnadu, India

²Associate Professor, Department of Computer Applications
Kumaraguru College of Technology, Coimbatore, Tamilnadu, India

³Assistant Professor, Department of Information Technology
Bannari Amman Institute of Technology, Sathyamangalam Erode, Tamilnadu, India

ABSTRACT

Due to Covid-19 pandemic, mankind was affected with respect to mental and physical stress. The causes of the disease have to be analysed based on the correlation factors such as Medication, Age, Gender, physical fitness and habits. In this paper we have proposed a classification model based on weighted average dynamic time warping approach to detect the disease severity as High, Medium, and Low by considering the multi-variant dependent variables that affect the prediction of Covid -19 positive cases. We also proposed the forecasting model based on the time series exponential moving average to identify the growth of disease with respect to Age, Gender and Medical history of the covid-19 positive patient. The results are obtained by defining the correlation function to measure the disease severity in the range of High, Medium and Low. The time series analysis is done with respect to mean average disease severity and also number of positive cases. The forecasting is performed based on the age, gender and existing disorders in health. The results are analysed with other time series classification models such as weighted time wrapping to make the model fits maximum to the available input.

KEY WORDS: COVID-19 SYMPTOMS, TIME SERIES ANALYSIS, CLASSIFICATION, FORECASTING, WEIGHTED AVERAGE DYNAMIC TIME WRAPPING.

INTRODUCTION

The Coronavirus disease considers a huge disaster which not only affected the healthcare but it also affected the human life in larger proportion. In the

Covid-19 pandemic, people faced more crises in their education, business, transportation, supply chain etc. People on counting the days of recovery, the cases were exponentially increased. The World Health Organisation (WHO) took a biggest initiative in assessing the people all over the world. However, there are no vaccines, the WHO has released certain guidelines and some rules for the people to be safe. As this disease caused by human-human transmission as all the people must follow social distancing. Infected people undergo certain treatment according to their severity level.

ARTICLE INFORMATION

*Corresponding Author: gopalakrishnanb@bitsathy.ac.in
Received 20th Oct 2020 Accepted after revision 11th Dec 2020
Print ISSN: 0974-6455 Online ISSN: 2321-4007 CODEN: BBRCBA

Thomson Reuters ISI Web of Science Clarivate Analytics USA and Crossref Indexed Journal

NAAS Journal Score 2020 (4.31) SJIF: 2020 (7.728)
A Society of Science and Nature Publication,
Bhopal India 2020. All rights reserved.
Online Contents Available at: <http://www.bbrc.in/>
Doi: <http://dx.doi.org/10.21786/bbrc/13.11/27>

The Controlling of coronavirus disease has become a great challenge for the healthcare and also for the social workers. On further cases, they found that few are in asymptotic stages. The asymptotic condition is that the people don't have any symptoms regarding to coronavirus. As that people will easily leads to death. Technologies like Artificial Intelligence and Data Science plays an important role for analysing how severe can disease be and also in finding the disease in a shorter while. Many applications were developed in order to assess their risk level. They also recognized the Coronavirus dataset respective of different parameters.

As the cases increases exponentially, we have used the time series exponential moving average forecasting model because the analysis can be updated to any recent change in parameters and it will also provide you the quick and accurate analysis of the disease. We also used classification algorithm in order to classify the risk level into Low, Medium and High. From the dataset, we will compute a mean average on certain parameters to identify the severity of the disease and also to find the number of positive cases. In this paper, the analysis is done on age, gender and their existing disorders. By using these data, we will be able to find the risk level of a person. For accuracy, we have used the time series classification model called weighted time wrapping for best fit for the input data. The accuracy gets increase as increase in fit model percentage.

Literature Survey: Ahasan Ullah Khan et al., established a paper in which he describes how the human novel virus was initialized first in China, Wuhan in the month of December 2019. This has been the major serious situation most of the peoples are facing in the current world. There has been many serious diseases most people are getting affected. Now the Covi-19, has a vast effect on damaging the people's immune system by affecting the human population. He describes the history of the disease, prevention, and to cure the disease in the real world. Eghbal Hosseini et al., describes the Corona virus efficiency problem which is spread across the world. We proposed three solutions to solve the efficiency obstacles in the distribution process. This will minimise and reduce most of the infected people rate by lock downing the people inside the homes in order to reduce the affected rate. Many algorithms have been proposed to lessen the spread of coronavirus and they used CVA algorithms using the mathematical calculation to find the optimal solution of the graph. Quoc-Viet Pham et al., represents the Corona virus situation about Artificial Intelligence (AI) and analytics for Coronavirus.

This coronavirus has vastly connected over more than 214 nations and zones on the planet, and has fundamentally influenced each part of our everyday lives. This project aims at describing their values requesting the people

about the corona virus outbreak and taking caution about the huge calamities of the coronavirus situation. In this paper the analysts with different experiences furnish the manners in which the large information about the Corona virus situation results in halting the Corona virus breakdown.

Vinay Chamola et al., describes the unparalleled eruption of the 2019 novel Covid, named as COVID-19 by means of the World health Organization (WHO), has set various administrations about the world. COVID-1 epidemic united by the anxiety of abused healthcare schemes and it has enforced a widely held countries into a state of biased or full lock down. In response to such acts, we investigation about the utilization of advances, for example, the Technologies such as IoT, AI, big data analytics and Blockchain which helps us to cure the infections of Novel coronavirus. Muhammad Adnan Shereen et al., established paper in which it explains about the coronavirus situation which is severely spreadable and virus infection caused by severe SARS-CoV-2 in China spread across the country.

In this the human transfer has been wide and there's no legally permitted medicine to obtainable to fight against the Novel coronavirus. We have analysed the serious situation of Coronavirus disease and the previous diseases SARS-CoV and MERS-CoV which deals with the pandemic situation. Sanjay Kumar et al., describes the identification and creation of sets of transferable bug datasets. Data processing, a method of discovery of noiseless appearances of massive data is one amongst such techniques that become additional widespread for treating large disease information set. The primary objective is to enhance watching systems in exaggerated conditions and UTs in Republic of India which can be terribly valued to the management, clinicians, the forces concerned in sympathetic significance of the disclose of new coronavirus (COVID-19) to enhance the government plans, pronouncements, remedial amenities (apertures, testing kits, disguises etc.) Ram Kumar Singh et al., represents the Coronavirus situation in the country which is consistently expanding from January last.

It leads to serious lock down. It plays a crucial responsibility in finding the statistics of Coronavirus disease under the prescribed conditions which have the tendency to test the features of applied math prototype which supported the classic Hot-Winters methodology as suggests that of describes the Novel coronavirus disease forecasts for Asian nation at different health conditions. An entire recovery from COVID-19 can happen solely where Associate calculable 450 days from January 2020. Hao Xu et al., proposed a system in which the block chain enabled a privacy scheme for tracing the covid-19 infected patients which senses the user information and location ID. It helps to determine the battery

performance, storage and location by minimizing the public notification of block chain for lower computing power. This will declare the user private key and declare the notification message to the blockchain user.

This will enhance the privacy key notifications to the user and the patient for covid-19 Vasilis Z. Marmarelis et al., proposed a system in which developed a dynamic decomposition model for covid-19 which is conducted on a daily basis to show the newly confirmed cases from the early stage of covid-19 till date. They have analysed this using the time series model by forecasting methods. The graph distinguishes the infection waves using different representations. This will help the people to easily characterize the affected people by using different colour variations. The RM dynamic representation mainly characterize the mathematical parameters using Auto-Regressive Integrated Moving Average Heet savla et al., presented a paper in which the data's are gathered from Johns Hopkins coronavirus research center.

The data are trained using various machine learning models. This trained machine learning model will first diagnose the infection and inform it to the patient. This will be increased day by day and the number of infected cases rises accordingly. The overcoming SVM and Linear regression models helps to determine the prediction of cure cases all over the world. The KNN model will characterize the person who is infected or not. These algorithms define the monitoring and keep record of pandemic situations in our country.

PROPOSED MODEL

Overview: The proposed model consists of various steps to classify and forecast the disease severity based on the different parameters like Age, Gender, Medical history and disease severity of the persons identified through global data sources. The Proposed Process model is defined below in figure 1.

The model consists of three modules

1. Data Acquisition
2. Data Pre-processing
3. Model Examination

Data Acquisition: The personal information on covid-19 is collected from various formats such as datasets , web content, table database and live data collection from IoT devices. The data are retrieved through governmental sources, world bank open interface, mobility data from Google and apple reports. The package has been developed to download the information from the above sources using the following code .

```
# install the package
install.packages("COVID19");
# load the package
library("COVID19");
```

The Covid-19 data hub research committee provides unified data from world health organisations. The Package CRAN Package Covid-19 which has greater clarity and reliability in the data source that is collected with respect to the different language and geographical constraints. The WHO also provides current data related to Covid -19 for the researchers to download and upgrade the analysis based on their requirement. The data are collected from other sources like hospitals reports based on the Diagnosing SARS-COV-2 with antibodies and Diagnosing SARS-COV-2 with Nucleic-acid based tech to identify the other metrics which can be integrated to calculate the severity metric of the individual test report. The other parameters such as the patient's medical history, whether they have other medical issues like asthma , cancer, cardiology disorder , diabetes and high blood pressure etc.

Data Pre-processing: The data pre-processing is an important aspect in the classification process using machine learning techniques. There are several steps involved in the data pre-processing that enables the data to be cleaned and perfectly matches the input requirements for the machine learning models.

Data Cleaning

Data filtering

Joining and Splitting of data

Data pivoting

Data Cleaning: The processes of deleting the unwanted information in the data source that are collected from various sources of the covid -19 disease diagnosis. The data collected contains the name and other irrelevant data that doesn't have impact on the analysis of the data set collected. In our process each table consists of the name , age, source link , journal and study in which the

data has been collected. To delete these columns from all the diagnosis tables and be able to minimize the complexity of the predictive analysis in classification. The date of diagnosis is an important factor in the time series classification models. The process of the data cleaning is done through the following steps as shown in Figure 2.

Data Filtering: The process of filtering is accomplished for several benefits that accomplish the data to be more consistent and error prone. The first and foremost step is to identify the dependent and independent variables in the data sources. The columns are combined to form a pivoting variable that reduces the column size. For example if the covid - 19 diagnosis data collected every day with respect to specific timing then it can be pivoted with respect to date so that every day have only one entry in the dependent variable. The measuring the error in the specific column data by considering the data value two different sources which finds the deviation in the measurement obtained from sources and the mean square error is obtained using the formula

$$MSE(\hat{y}, y) = E[(\hat{y} - y)^2] = \frac{1}{N} \sum_{n=1}^N (\hat{y}_n - y_n)^2$$

The next process in filtering is identifying the clone data collected from different sources and removing the duplicate information from different sources for the same instance SAR reports collected every day.

$$\text{maximum...} \left[\frac{\text{hours}}{\text{day}} \times \frac{\text{minutes}}{\text{hour}} \times \frac{\text{seconds}}{\text{minute}} \right] = \frac{24 \times 60 \times 60}{30} = 2880 \left[\frac{\text{reports}}{\text{day}} \right]$$

The contradictory data should be identified in the data set that leads to inconsistency in the accuracy obtained by the proposed model. This can be achieved by the interpolated values between the values from different sources.

C. Model Examination: The proposed model consists of two modules

- Classification with respect to time series algorithm
- Forecasting on the severity of the disease covid-19

The classification algorithms are broadly divided into four categories

- Binary Classification
- Multi-Class Classification
- Multi-Label Classification
- Imbalanced Classification

The process of time series classification deals with the repeated pattern of the data analysis with respect to date and time field. The classification time series algorithm has been proposed as Weighted Average Dynamic Time Warping (WADTW) which can perform the average series of matrices that are generated through the repeated sequence in the linear regression prediction of the data source.

Let two division of sequence as $A = \{x_1, x_2, x_3, \dots, x_n\}$
 $B = \{y_1, y_2, y_3, \dots, y_n\}$ where x_1, x_2, x_3 are dependent variable and y_1, y_2, y_3 are independent variables. Average of the x_1, x_2, x_3 are taken with respect to the weight assigned in the every iteration.

The Euclidean Distance for $De(x, y) = \sum \text{Avg}(A - B)^2$
 The logistic weight function is defined as

$$w(a) = \frac{w_{\max}}{1 + e^{-g \cdot (a - m/2)}}$$

$W_{\max}$ is the upper bound on the weight and g is the penalty level for warping.

D. Algorithm for Weighted Average Dynamic Time Wrapping

Parameters: stiffness parameter v , penalty value λ
 Let D be an $n + 1 \times n + 1$ matrix initialised to zero.

$\text{Dist}(1, 1) \leftarrow 0$

$\text{Dist}(2, 1) \leftarrow x_1^2$

$\text{Dist}(1, 2) \leftarrow y_1^2$

for $a \leftarrow 2$ to $n + 1$ do

$\text{Dist}(a, 1) \leftarrow \text{Dist}(a - 1, 1) + |x_{a-2} - x_{a-1}|$

for $b \leftarrow 2$ to $n + 1$ do

$\text{Dist}(1, a) \leftarrow \text{Dist}(1, b - 1) + |y_{b-2} - y_{b-1}|$

for $a \leftarrow 2$ to $n + 1$ do

for $b \leftarrow 2$ to $n + 1$ do

if $a > 2$ and $b > 2$ then

$\text{dist1} \leftarrow \text{Dist}(a - 1, b - 1) + v \times |x_{a-b}| \times 2 + |x_{a-1} - y_{b-1}| + |x_{a-2} - y_{b-2}|$

else

$\text{dist1} \leftarrow \text{Dist}(a - 1, b - 1) + v \times |a - b| + |x_{a-1} - y_{b-1}|$

if $a > 2$ then

$\text{dist2} \leftarrow \text{Dist}(a - 1, b) + |x_{a-1} - x_{a-2}| + \lambda + v$


```

else
dist2  $\leftarrow$  Dist ( a - 1 , b ) + | x a - 1 | +  $\lambda$ 
if b > 2 then
dist3  $\leftarrow$  Dist ( a , b - 1 ) + | y a - 1 - y b - 2 | +  $\lambda$  + v
else
dist3  $\leftarrow$  Dist ( a , b - 1 ) + | y b - 1 | +  $\lambda$ 
Dist ( a , b )  $\leftarrow$  min ( dist1 , dist2 , dist3 )
return Dist ( n + 1 , n + 1 ).

```

The data has to be splinted as training and testing for the model that predicts the severity of the disease in the required output.

E. Forecasting model to predict the severity of the particular condition:

Symbolic aggregate approximation for window prediction

```

SAX_Window()
Set window length ln
Set word size ws
Set alphabet size  $\alpha$ 
Compute lookuptable
ln = length( $t_s$ )
for all t in [0, ln - 1] do
normed ts = z normalize(time series [ti : ti +l])
PAA = compute PAA (normed ts)
S = ""
for vi in PAA do
S += lookup(vi)
end for
Add S to the final representation
end for

```

Dividing testing and training data based on the window size

Training Time Series TrTS

```

Training Data TrTS ()
Set window size wi = 20
Set alphabet size  $\alpha$  = 5
for li = min li , wi <= Li, w+= sqrt(L) do
S = SAXwindow(raw data, li, wi,  $\alpha$ )
Train the SEQL model from the
symbolic representation Mi = TrTS(SAX)
M[li, wi,  $\alpha$ ] = Mi
end for

```

Testing Time series TeTS

```

Testing Data TeTS ( )
Set window size wi = 20
Set alphabet size  $\alpha$  = 5
score = 0
for l = (min li, wi <= L, wi += sqrt(L) do
sax = SAXwindow(raw data, li, wi,  $\alpha$ )
score += M[li, wi,  $\alpha$ ].predict(SAX)
end for
Predict = sign(score)

```

Feature Selection

```

FSelection ()
Set window size wi = 20
Set subsequence size  $\alpha$  = 5
Set of features Fi = {}
for l = min li, wi <= L, wi += sqrt(L) do
SAX = SAX_window (raw data, li, wi, ai)
Train the SEQL model Mi = TrTS(SAX)
for all subsequence in Mi do
Fi .add(new Feature(ssi, li, wi,  $\alpha$ ))
end for
end for

```

RESULT AND ANALYSIS

The process of implementing the above algorithms on the dataset collected from various sources such as open covid19 dataset. The dataset contains the more than 25 symptoms measured with date and country. The dataset has been pre-processed to eliminate the inconsistent and unwanted columns from the dataset. The analysis are made based on the following conditions

- The unsupervised clustering is performed based on Common_Symptoms (Co_S), Rarely_Symptoms (Re_S) and Critical_ Symptoms (Cr_S).
- The classification are done based on the Low, Medium and High Severity Level on the three clusters Co_S, Re_S and Cr_S,.
- The forecasting model that predicts the growth rate of the clustered symptoms Co_S, Re_S and Cr_S individually.

A. Cluster_symptoms time series analysis:

Figure 3: Common_Syptoms

The analysis report for the above Figure is shown in the following table: The Classification of Low, Medium and High are differentiated based on the severity levels of the Common symptoms. In the same way Rarely_symptomss and Critical_symptonss are classified based on Low, Medium and High with respect to time series from January – September 2020 covid-19 Datasets.

Variables:	Sum of symptom:Fever
	Sum of symptom:Headache
	Sum of symptom:Pain
	Sum of symptom:Cough
Level of Detail:	Month of Date
Scaling:	Normalised

Number of Clusters:	3
Number of Points:	13872
Between-group Sum of Squares:	492.45
Within-group Sum of Squares:	437.1
Total Sum of Squares:	929.55

		Centres			
Classification	Number of Items	Sum of symptom:Fever	Sum of symptom:Headache	Sum of symptom:Pain	Sum of symptom:Cough
LOW	4369	5.4943	4.5191	36.128	7.9761
Medium	820	12.836	4.0875	30.775	12.198
High	8683	3.9904	3.7829	32.102	3.3234

B. Rarely_symptoms time series analysis.

Analysis Report: The analysis is performed based on the four different parameters for the time series classification. The P-value becomes 0. it implies the classification is perfect match with these variables.

C. Critical_symptoms time series analysis.

The analysis of critical symptoms such as nausea, fatigue, anomia, agansia as shown in figure 5.

Analysis Report

D. Forecasting of Co_S, Re_S and Cr_S based on Time Series: The forecasting was performed on the cluster model to identify which symptoms will increase in the fore coming months based on the previous information's. The following figure 6 shows the forecast values for all the symptoms and their trend analysis.

Trend Analysis: A polynomial trend model of degree 4 is computed for common symptoms given Date Month.

Number of modeled observations: 9
 Number of filtered observations: 0
 Model degrees of freedom: 5
 Residual degrees of freedom (DF): 4
 SSE (sum squared error): 34.4841
 MSE (mean squared error): 8.62102
 R-Squared: 0.793371
 Standard error: 2.93616
 p-value (significance): 0.110442

Since the R-Square value is 79 % which is in the acceptable line that fits the actual values

Forecasting Analysis Options Used to Create Forecasts

Time series: Month of Date

Measures: common symptoms

Forecast forward: 4 months (October 2020 – January 2021)

Forecast based on: January 2020 – September 2020

Ignore last: No periods ignored

Figure 4: Rarely_Symptoms vs Time Series

			Model		Error	
Variable	F-statistic	p-value	Sum of Squares	DF	Sum of Squares	DF
Sum of symptom:Sore throat	4899.0	0.0	226.5	2	285.2	12336
Sum of symptom:Shortness of breath	4306.0	0.0	115.3	2	165.2	12336
Sum of symptom:Chills	3321.0	0.0	29.45	2	54.7	12336
Sum of symptom:Diarrhea	1386.0	0.0	6.751	2	30.03	12336

Figure 5: Critical_Symptoms vs Time Series

			Model		Error	
Variable	F-statistic	p-value	Sum of Squares	DF	Sum of Squares	DF
Sum of symptom:Nausea	2991.0	0.0	95.74	2	129.2	8071
Sum of symptom:Fatigue	2935.0	0.0	100.0	2	137.5	8071
Sum of symptom:Anosmia	105.5	0.0	0.5085	2	19.46	8071
Sum of symptom:Ageusia	82.82	0.0	0.6075	2	29.6	8071

Figure 6: Forecasting vs Time Series

Since the forecasting is based on additive and multiplicative to forecast the value of MAPE which is lower 10 % then the model is acceptable.

All forecasts were computed using exponential smoothing

Common Symptoms Results

Model			Quality Metrics					Smoothing Coefficients		
Level	Trend	Season	RMSE	MAE	MASE	MAPE	AIC	Alpha	Beta	Gamma
Additive	None	None	4	3	1.12	6.2%	31	0.500	0.000	0.000

Rarely Symptoms Results

Model			Quality Metrics					Smoothing Coefficients		
Level	Trend	Season	RMSE	MAE	MASE	MAPE	AIC	Alpha	Beta	Gamma
Multiplicative	Multiplicative	None	1	1	0.68	10.6%	10	0.000	0.047	0.000

Critical Symptoms Results

Model			Quality Metrics					Smoothing Coefficients		
Level	Trend	Season	RMSE	MAE	MASE	MAPE	AIC	Alpha	Beta	Gamma
Additive	None	None	0	0	1.44	6.0%	-10	0.000	0.000	0.000

CONCLUSION

The covid -19 is an unpredictable disease due to many syndromes that affect the human mankind. It also have a property of spreading across the globe which leads to continuous deaths. This paper mainly focused on the symptoms that are more relevant to the cause of the Covid-19 results positive. The model is classified based on the severity of the certain symptoms and their impact are classified as low, medium and high with respect time series data. The proposed model also forecast the severity of the symptoms such as Common-Symptom's, Rarely_Symptom's and Critical_Symptom's with respect to the future time line. In future it can be used to predict the consequences arise due to other diseases like Asthma, Cancer, and Diabetes. The results are analysed based on the values obtained by implementing the

algorithms such as Weighted Average Dynamic Time wrapping which prominently shown better output.

REFERENCES

- Ahasan Ullah Khan, Arnika Afrin Proma, Margia Akter, Md. Matiur Rahaman, Shobhan Das "A Review on Coronavirus Disease (COVID-19) Epidemic Threat for Global Health in 2020", IEEE, 2014.
- J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.
- Eghbal Hosseini, Kayhan Zrar Ghafoor, Ali Safaa Sadiq, Mohsen Guizani, and Ali Emrouznejad, "COVID-19 Optimizer Algorithm, Modeling and Controlling of Coronavirus Distribution Process", IEEE, 2020.
- Hao Xu, Lei Zhang, Oluwakayode Onireti, Yangfang, William J Buchanan, Muhammed Ali Imran, "BeepTrace: Blockchain-enabled Privacy preserving contact Tracing for Covid-19 Pandemic and Beyond", IEEE, 2020.
- Heet Savla, Vruddhi Mehta, Ramchandra Mangrulkar, "Prediction and Diagnosis of COVID-19 Using Machine Learning Algorithms", IJRTE, 2020.
- Muhammed Adnan Shereen, Suliman Khan, Abeer Kazmi, Nadia Bashir, Rabee Siddique, "Covid-19 infection: Origin, transmission and Characteristics of human coronaviruses", Journal of Advanced Research, 2020.
- Quoc-viet Pham, Dinh C. Nguyen, Thien Huynh-The, Woo-joo Hwang, Pubudu N. Pathirana "Artificial Intelligence and Big data for coronavirus pandemic: A Survey on the State-of-the-arts", IEEE, 2020.
- Ram Kumar Singh, Martin Drews, Manuel De la Sen, Manoj Kumar, S.S. Singh, A.K. Pandey, Prashanth Kumar Srivastava, Manmohan Dobriyal, Meenu Rani, Preeti Kumari, Pavan Kumar, "Short term statistical forecasts of Covid-19 Infections in India", IEEE, 2020.
- Sanjay Kumar "Monitoring Novel coronaviruses Infections in India by cluster analysis", Springer, 2020.
- Vasilis Z. Marmarelis, Fellow, "Predicting Modeling of covid-19 data in the US: Adaptive phase-space approach", IEEE, 2020.
- Vinay Chamola, Vikas Hassija, Vatsal Gupta, Mohsen Guizani, "A Comprehensive Review of the COVID-19 Pandemic and the Role of IOT, Drones, AI, Blockchain and 5G in managing its impact", IEEE, 2020.