A·P·U
ASIA PACIFIC UNIVERSITY
OF TECHNOLOGY & INNOVATION

INDIVIDUAL ASSIGNMENT

TECHNOLOGY PARK MALAYSIA

CT047-3-M-BDAT

BIG DATA ANALYTICS AND TECHNOLOGIES

APDMF2112DSBA(DE)(PR)

HAND OUT DATE: 22 JANUARY 2022

HAND IN DATE: 13 MARCH 2022

WEIGHTAGE: 20%

INSTRUCTIONS TO CANDIDATES:

1 **Submit your assessment via Moodle with specific folder provided.**

2 **Students are advised to underpin their answers with the use of references (cited using the Harvard Name System of Referencing).**

3 **Late submission will be awarded zero (0) unless Extenuating Circumstances (EC) are upheld.**

4 **Cases of plagiarism will be penalized.**

7 **You must obtain 50% overall to pass this module.**

A·P·U
ASIA PACIFIC UNIVERSITY
OF TECHNOLOGY & INNOVATION

Assignment Part A:

Techniques and Methods of Big Data Analytics

Module Code:	**CT047-3-M**
Module Name:	**Big Data Analytics & Technologies**
Student’s TP:	**TP065778**
Student’s Name:	**LEE KEAN LIM**
Lecturer’s Name:	**Dr. V. Sivakumar**
Intake Code:	**APDMF2112DSBA(DE)(PR)**

ABSTRACT

The extremely contagious severe acute respiratory syndrome coronavirus-2 has spread globally, infected millions, and overwhelmed the medical sector. High velocity and high volume of data is generated daily from the admission of large number of infected patients. This study aims to identify the need of big data analytic solutions in addition to propose several methods and techniques of big data analytics applicable to the domain. The proposed solutions would be based on mild to severe symptoms experienced by the patients which are evaluated by doctors using a star rating system. The study would be based on literature research on existing application of big data analytics in the medical domain. The need of big data solutions in the current domain is identified based on the data types and data characteristics exhibited namely volume, velocity, variety, veracity, and value. Five big data analytic solutions are proposed namely association rule mining, decision tree, logistic regression, artificial neural network, and naïve bayes classifier. The proposed methods are based on literature which are applicable to the domain to assist the medical professional in diagnosis of the patients more effectively and efficiently.

Table of Contents

1. INTRODUCTION

In December 2019, Wuhan, China emerged a virus that has since spread throughout every nation and caused a global pandemic (F. Wu *et al.*, 2020). The virus is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and officially known as the coronavirus disease-19 (COVID-19) (World Health Organization, 2020). COVID-19 is highly contagious and transmit through droplets from an infected person to another person. The virus targets primarily the throat and lungs of an infected person. The symptoms of the infected differ for everyone based on different factors such as underlying health conditions and age. Some would show no symptoms at all, some would show mild symptoms like fever, cough, tiredness, and loss of taste, while some would show severe symptoms that would lead to permanent irreversible damage such as respiratory failure, lung damage, heart muscle damage, nervous system problem, kidney failure, and even death (Johns Hopkins Medicine, 2022). As of 28th February 2022, about 434 million people are infected and about 6 million lives lost to the pandemic since the first outbreak in December 2019 (World Health Organization, 2022). Although vaccine for the virus is currently available, but the virus has evolved and caused multiple new variants to emerge which is more contagious than the previous strain.

The highly transmissible virus is infecting people at an alarming rate which if not properly managed, it can easily overwhelm and collapse the healthcare system. Healthcare professionals and facilities are unable to cope with the sudden surge of patients which could lead to some patients unable to receive treatment and healthcare professionals experiencing work burnout. Since COVID-19 patients experience a multitude of symptoms ranging from mild to severe. This would require healthcare professionals to perform multiple tests on a single patient which would require utilization of high amount of manpower and resource that consumes a lot of time whereby time is always a critical factor when it comes to saving lives. Due to the many tests performed, a large amount of data is recorded and to be processed to understand the health condition of the patients for deciding further treatment. With the high velocity and amount of data generated daily, a more effective and efficient method is required to process the data to aid the overburdened sector.

With the advancement of technologies, many tools have been developed with the capabilities to describe, diagnose, predict, and prescript solutions to many variations of problems. Such capabilities are part of the big data analytic solution. In the current domain, it is possible to utilize big data analytic solutions to increase the effectiveness and efficiency in various processes. By using big data analytics, the multitude of data from different test and reports can be analyzed and transformed into valuable insights in a faster manner to allow healthcare professionals to make quicker and better decision while having more time to focus on more critical tasks hence improving the efficiency of medical treatment and likely increase the survival rate of the patients. This report studies the need of big data analytics solution in the healthcare domain and proposing multiple methods of big data analytics that is applicable to the domain specifically on diagnosis of COVID-19 patients.

1.1 Aim and Objectives

The aim of this study is to propose big data analytic solutions that can be used to aid the healthcare domain to evaluate symptoms present in COVID-19 patients more effectively and efficiently.

The objectives of the study are as followed:

- To identify the need of big data analytics solution in the healthcare domain specifically in diagnosing COVID-19 patients with specified symptoms.
- To propose the suitable techniques and methods of big data analytics to evaluate symptoms of patients more effectively and efficiently.

2. BACKGROUND

Survival rate of a COVID-19 patient is highly dependent on the age and underlying health condition of the patient. Prognosis of COVID-19 elderly patients without comorbidities are of high risk of mortality while elderly patients with comorbidities are of even higher risk of mortality. Byeon *et al.* (2021) studied the age factor and different comorbidities on mortality of COVID-19 patients in South Korea and found that elderly men have a lower survival rate than elderly women when age is greater than 80. In addition, patients with comorbidities such as diabetes and urinary system diseases lead to a worse prognosis. Consistent with the findings, Yan *et al.* (2020) studied the association between COVID-19 patients with comorbidity diabetes and mortality rate and found that patients with comorbidity of diabetes lead to an increased risk of death and men posed a higher risk than women. Moreover, a study by Lee *et al.* (2020) in a different city in South Korea, found that mortality rate of elderly patients with comorbidities of diabetes, chronic lung diseases, and chronic neurologic diseases posed a significantly increased risk of patients mortality. However, the author mentioned that the comorbidity hypertension was not a significant risk factor in elderly patient mortality rate. On the other hand, a study on prognosis of COVID-19 elderly patients from China found that patients with comorbidities of hypertension, diabetes, or coronary heart disease posed a significant risk to mortality (J. Wu *et al.*, 2020; Zhou *et al.*, 2020). In another patient mortality study by Sousa *et al.* (2020) on the COVID-19 patients in Brazil, it was found that patients of age greater than 60 with comorbidities diabetes, neurological disease, lung disease, or cardiovascular disease contributed significantly to the patient mortality. Therefore, evidence suggests that prognosis of COVID-19 patients with comorbidities of either diabetes, cardiovascular disease, or lung cancer led to an increased risk of death. However, for patients with comorbidity of hypertension, a mixed opinion is provided.

There are several symptoms associated with a COVID-19 infected patient which are highly dependent on the severity of the illness. The patients can be asymptomatic, experiencing mild or severe symptoms. For the mild severity patients, the commonly found symptoms were fever, dry cough, shortness of breath, and fatigue while less commonly found symptoms were observed on some patients such as anorexia, diarrhea, nausea, and headache (Yan *et al.*, 2020). In another study, Liu *et al.* (2021) recorded symptoms present in COVID-19 patients and illustrated in a pie chart showing the percentage of symptoms present in the group of patients as shown in Figure 1. Majority percentage of symptoms present in the mild severity patients include fever (87.9%), dry cough (67.7%), fatigue (38.1%), and sputum production (33.4%). In addition, it was found that there exist a relationship between blood pressure and blood oxygen level where COVID-19 patients with severe symptoms showed a raised in blood pressure accompanied by a reduction in blood oxygen level and these symptoms may be associated with a degradation of the lung function which ultimately led to a heart failure as a result of hypoxia (Vicenzi *et al.*, 2020). J. Wu *et al.* (2020) tracked the blood glucose level of COVID-19 patients from the time of admission to the time of it reaching critical cases or death, found that critical cases patients tend to have a higher blood glucose level as compared to the blood glucose level during time of admission which had significantly poorer prognosis.

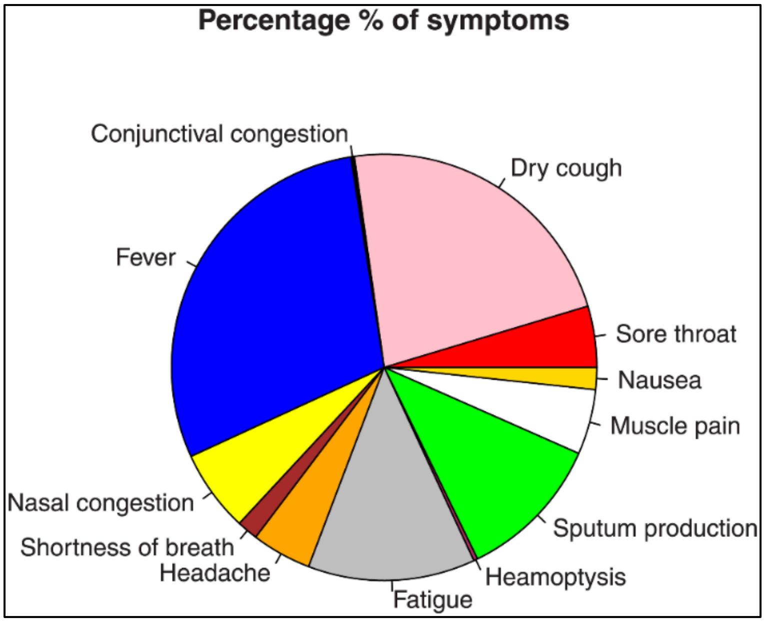

Figure 1: *Percentage of symptoms of COVID-19 patients (Liu et al., 2021).*

At the current time, data is being generated everywhere and at a high speed, the data can be utilized for a better purpose. Such data is coined big data which described as the data that is unable to be processed by using traditional methods or data storage. Big data can be used for development of a model which would generalize to a specific task. Big data analytics is the use of advanced analytical tools to handle structured, unstructured, or semi-structured data to derive valuable insights. Big data analytics solutions had been applied in the healthcare domain in various ways to aid the healthcare professionals such as providing diagnosis to patients, predicting the risk score, decision making, and pharmaceutical development (Alsunaidi *et al.*, 2021). Big data typically exhibit several characteristic namely volume, velocity, variety, veracity, and value (Kapil *et al.*, 2016). Tabulated in Table 1 shows the description of each big data characteristic as mentioned.

Table 1: *Big Data Characteristic Description*

Big Data Characteristic	**Description**
Volume	Concerns with the size of data which is collected and stored.
Velocity	Concerns with the transmission speed of data from the source to the destination.
Variety	Concerns with the type of data where data can exist in different format such as text, images, audio etc.
Veracity	Concerns with the quality of data which accuracy of collected data determines the outcome of the analysis.
Value	Concerns with the importance of data which brings value after processing.

There exist four categories of big data analytics namely descriptive analytics, prescriptive analytics, diagnostic analytics, and predictive analytics. Firstly, the descriptive analytics utilize historical data to uncover patterns and summarizes into a readable format. The aim is to identify and summarize what has happened in the past. Examples of descriptive analytics are summary statistics, association rules, etc. Secondly, the diagnostic analytics which provides a diagnosis to a problem. The aim is to identify why something happened or the cause of a phenomenon. Examples of diagnostic analytics are data mining, customer health score analysis, etc. Thirdly, the predictive analytics which provides a prediction into the future probable outcome within a specific scenario. Examples of predictive analytics are churn analysis, decision trees, etc. Finally, the prescriptive analytics which aimed to provide probable recommendations or actions to a problem based on the results of predictive analytics. Examples of prescriptive analytics are marketing strategies optimization, risk management, etc.

Individuals with suspected infection of COVID-19 can be diagnosed using reverse transcriptase polymerase chain reaction (RT-PCR) which the result from the test kit is highly reliable. However, due to the immense demand of the test kit, the supply is insufficient to cater for the demand. Example of a case where big data analytics solution are applied in combat with the COVID-19 is the use of questionnaire on patients to determine the likelihood of being contracted of COVID-19. Zoabi *et al.* (2021) utilized machine learning method that receive the questionnaire inputs from individuals to predict the likelihood of being infected by COVID-19. The questionnaire contains a series of true and false questions relating to symptoms experienced by the individual such as cough, fever, sore throat, shortness of breath, and headache. The inputs are fed to a gradient boosting machine learning model built with decision tree base learners and resulted a 0.9 area under the receiver operating characteristic curve (ROC) which is highly reliable. The method can be a complement to the RT-PCR test kit to diagnose individuals for COVID-19 where the test kit can be used on higher likelihood suspect based on the prediction to confirm the diagnosis.

3. DISCUSSION

The first section would include answering on the need of big data solution in the medical domain with respect to COVID-19 disease. In addition, the second section would propose several methods and techniques on performing big data analytics suitable for specific symptoms evaluation. Data in focus would be the symptoms experienced by a COVID-19 patient as followed:

a) Shortness of breath

b) Fever

c) Dry cough

d) Phlegm collection in lung / Sputum production

e) Acute pneumonia

f) High blood pressure

g) High blood glucose fluctuations

h) Coronary failure

i) Low blood oxygen level

Nine symptoms are listed whereby one or more symptoms are associated with the patient on different degree of illness severity. Evaluation of the symptoms are given with a star rating.

3.1 Need for Big Data solution

As COVID-19 has spread around the globe, huge amount and various type of data is generated and would likely increase as the number of infected people increases. Data comes in different types and from various sources which can be classified into several categories as suggested by Alsunaidi *et al.* (2021). Compilation of data and data types tabulated in Table *2* shows the typically recorded data in the medical sector and specifically data related to diagnosis of COVID-19 patients.

Table 2: Data sources and types (Alsunaidi et al., 2021)

Data Category	Data	Data Type
Demographics data	Gender, Age, Height, Weight, Body mass index, Language, Race, Ethnicity, Nationality, Religion, Marital status, Median income, Postal code, Location, Region, Insurance, Job, Number of family members	Structured data
Travel data	Recent outside travel history, Outside destinations	Unstructured data as travel history may contain text description
Activity data	Steps walked in a day, Hours of Sleep, Heart rate, Home-quarantine activities	Mostly structured data but unstructured data for home activities as it involves text description
Medical data	Vital signs, Symptoms, Comorbidities, Medical history, Routinely taken medications, Laboratory findings, CT scans, ICU length of stay, Readmission status	Mostly structured data but unstructured data for CT scans as it involves image data
COVID-19 data	Test date, Result, Symptom onset date, Incubation periods, Treatment measures, Infection feels	Mostly structured data but unstructured data for treatment measures as it involves text description
Samples	Throat swabs, Blood samples	Structured data
Vital signs	Temperature, Heart rate, Respiratory rate, Blood pressure systolic, Blood pressure diastolic, Oxygen saturation	Structured data
Symptoms	Fever, Shortness of breath, Chest pain, Cough, Sneezing, Chills, Runny nose, Anosmia, Headache, Sore throat, Sputum production, Fatigue, Muscle aches, Diarrhea, Vomiting, Loss of appetite, Trouble sleeping, Stomach pain, Rash, Neuralgia	Structured data

As shown in Table *2*, majority of the data can be categorized into structured data while a few of them exist in unstructured data format. Several big data characteristics are identified which relate to the domain namely Volume, Variety, Velocity, Veracity, and Value. Table *3* shows the justification on the data characteristic using data acquired in the medical sector with reference to Table *2*.

Table 3: *Data characteristic justification*

Data Characteristic	Justification
Volume	Vast amount of data recorded from a single patient which involved multiple data types and data categories as tabulated.
Velocity	As time is a critical factor in saving lives, data needs to be transmitted at a high speed for judgement to be made in time. Specifically in the vital signs data category, real time monitoring and feedback of the patient vital signs is important to track the health progress of the patient.
Variety	Structured and unstructured data types exist. While majority of the data is structured, some unstructured data type can be found which is CT scans, treatment measures description, and home activities description.
Veracity	The data recorded from performing testing on patient such as from the symptoms data category need to be accurate for a proper diagnosis of the patient where various illness may exhibit the same symptoms and inaccuracy of result may lead to misdiagnosis.
Value	Data such as from the symptoms, medical data, samples, vital signs, and COVID-19 data category are important for the medical professional to diagnose condition and provide suitable treatment for the patient.

In an example of utilization of big data to combat COVID-19, Jeong *et al.* (2020) developed a wearable device that continuously tracks the user data namely heart rate, respiratory rate, physical activity, and coughing to detect early signs of COVID-19 infection. The data is displayed in a dashboard accessible from a mobile application. Data are stored in the cloud platform and would alert user to seek medical attention when signs of infection is observed in the data. This demonstrates the use of big data where huge amount of data is generated from the continuous tracking and real time reporting to the user to obtain medical treatment in the earliest time to reduce escalation of illness severity.

Therefore, based on the data characteristic exhibited by the various types of data available in the healthcare domain. It is sufficient to conclude that data coming from this domain is considered as big data and big data analytic solutions are required to be implemented to assist and improve performance of the healthcare professionals in the medical sector.

3.2 Techniques and method proposal

The data in the current scenario provided is a series of symptoms evaluated by the medical professional rated with a star rating. The higher the star rating for a given symptom, the higher the severity of the symptom experienced by the patient. Hence, proposal of the big data analytic techniques and methods would have to take in consideration the type of data provided. The proposed big data analytic techniques and methods would be used to diagnose the health of the patient. The following section discuss the proposed big data analytic techniques and methods suitable for the specified domain.

3.2.1 Proposal 1 – Association Rule Mining

The first proposed big data analytic method is association rule mining (ARM). ARM uses a rule-based machine learning method for identifying unique patterns and relationship between features in a large dataset. There are few methods and algorithms available in ARM to generate association rules namely Apriori Algorithm, Frequent Pattern Growth Algorithm, Genetic Algorithm, etc (Kaur & Madan, 2015). Each of the algorithm ARM can be applied in many domains such as medical, market basket analysis, product clustering, etc. ARM can be very useful as it does not require labeled data which saves significant time in data collection and preprocessing. However, some algorithm may contain many parameters for tuning which requires knowledgeable personnel to build and test the algorithm before deploying for the task.

In the current domain, ARM can be used to find out the probability of illness based on the relationship between symptoms and the ratings provided by the medical professional. An illness may comprise of a different combination of symptoms and severity of the symptoms which requires the expert knowledge of the medical professional to diagnose what illness the patient is having. Thus, diagnosing the patient is not an easy task and time consuming. The output from ARM which are probabilities of a series of illness can aid the medical professional to diagnose the patient in a quicker manner although is it not advisable to fully rely on the results of the model. The model can be a complement to the medical workers to quickly diagnose and provide necessary treatment to the patient.

In an example where ARM is used in the medical domain, Chen *et al.* (2018) developed a disease diagnosis and treatment recommendation system to assist doctors when they are experiencing shortage of manpower or unavailability of senior doctors to provide quick diagnosis. The ARM in this scenario utilized the Apriori Algorithm to identify the relationship between features. The model was able to diagnose and provide treatment schemes intelligently and accurately to the doctors based on inspection reports of patients.

3.2.2 Proposal 2 – Decision Tree

The second proposed big data analytic method is decision tree. A decision tree machine learning method is an approach suitable for multiclass classification problem. The underlying concept of the decision tree is to convert data into decision trees or rules where a set of hierarchical decisions are output at each node. The input features of the model can accept by both numerical and categorical variables while requiring only a small dataset is sufficient for development of the model. In addition, the output from the tree is easily interpretable and understood as a visualization of the tree is provided showing how each result is obtained. Furthermore, an increase of data points would not hinder the speed of inference. However, decision trees are prone to overfitting. It is suggested to perform selection of input features, limiting to features with high significance to reduce the split of trees and outputting too many nodes. Moreover, dataset used for decision tree requires class balancing as the model is prone to bias when one class is significantly more than the others thus preprocessing of the data is required.

In the current domain, decision tree can be used to diagnose the illness and condition of the patient. Since decision tree is a supervised machine learning method, mapping of symptoms to illness is required initially to develop the model. Based on the output from the decision tree, medical professionals can easily interpret the probable illness at each node of the decision tree which is visualized hierarchically. However, parameter tuning would be required to limit the number of nodes output from the tree to avoid overfitting and producing too many results that would confuse the doctors. The training of the model would not require large dataset which can be beneficial in the early phase where data is scarce, but a larger dataset would produce more reliable result. Inference time from the model is fast thus allowing medical professionals to diagnose the patient and provide treatment and overall increases the efficiency and effectiveness of the process.

In an example where decision tree is used in the medical domain, Rochmawati *et al.* (2020) has utilized decision tree specifically with algorithm J48 and Hoeffding Tree to classify the condition of COVID-19 patients into mild, moderate, severe and not COVID disease using 13 symptoms experienced by patients. The study found that both algorithms provide similar result in terms of accuracy, processing time, and mean absolute error with the J48 just slightly better than the Hoeffding Tree. The algorithms were able to produce highly reliable rules in classifying the condition of the patients. It was also mentioned that Hoeffding Tree produced a simpler and a smaller number of nodes as compared to the J48.

3.2.3 Proposal 3 – Logistic Regression

The third proposed big data analytic method is logistic regression (LR). LR method is a classifier commonly used for binary target variable prediction, but extensions exist to perform a multiclass problem. The LR algorithm provides probability to the output with respect to the classes. The LR is a very simple machine learning algorithm to implement that does not require a large dataset to compute the model. It is also very efficient where it does not require high computational resource. However, the model is prone to overfitting if dataset exhibit high dimensionality which can be overcome with regularization. Preprocessing of the data is highly encouraged when using this algorithm such as scaling and normalization the data to reduce overfitting of the model.

In the current domain, LR can be used to predict the survival or mortality rate of patients based on the evaluated symptoms rating. Since LR is a supervised machine learning method, labeled data would be required for development of the model. Due to the sudden surge in patients infected by COVID-19, medical manpower and facilities are in extreme shortage. The prediction of the survival or mortality rate of patients would allow medical professionals to better manage resources to focus on the patients with higher survival rate. Currently, there does not exist a cure for patients with severe symptoms which they are fully dependent on the immune system itself to fight off the virus. Therefore, in dire times when hospitals are overcrowded, the doctors would need to shift the attention to prioritize medical treatment to the higher probable survival patients.

In an example where LR is used in the medical domain, Aljameel *et al.* (2021) studied the application of LR to predict the survival rate of COVID-19 patients using different combination of clinical features consisting of demographic of patient, symptoms, and chronic disease. The author identified the dataset to have imbalanced data thus suggest using synthetic minority oversampling technique to overcome the issue. The study found that performance of LR in predicting the survival rate was satisfactory with an accuracy achieving 0.849. It was also mentioned that the model should be validated using multiple datasets to improve performance.

3.2.4 Proposal 4 – Artificial Neural Network

The fourth proposed big data analytic method is artificial neural network (ANN). Inspired by the how the human brain works, the ANN consist of layers of nodes connecting to each other mimicking the neurons in brain. Each node is associated with a weight and threshold. ANN is a powerful tool which has been used in many fields with promising results such as computer vision, natural language processing, etc. ANN can be used for both regression and classification problems and is able to model non-linear data with large number of input features. However, producing an ANN model requires a large dataset and is expensive in terms of computation and time.

In the current domain, ANN can be used to predict the likely disease, prognosis, and the mortality of the patient. All the nine symptoms evaluated by star rating can be input features to the ANN to be used for prediction. The development and fine tuning of the model is a time consuming and computational resource expensive process. However, if such model is successfully deployed, it would provide significant help to the medical professionals, as ANN is very reliable in determining non-linear relationship between features and provide high accuracy results. This would in turn significantly boost the performance of medical staff and management of patients and resources.

In an example where ANN is used in the medical domain, Çotoy *et al.* (2021) developed a model to predict the likelihood of patient requiring admission to the intensive care unit. The model utilized 18 input features consisting of demographic, comorbidities, symptoms, and habits of patients for development of the model. The evaluated model achieved an accuracy of 79%. Since deployment, the model assisted the medical professionals to better manage medical resources and improved workflow performance.

3.2.5 Proposal 5 – Naïve Bayes Classifier

The fifth proposed big data analytic method is Naïve Bayes Classifier (NB). Based on the Bayes Theorem, NB utilize probabilities in predicting the output. A very important assumption when using NB is there must be independence among predictor variables. The NB algorithm is very easy to implement and provides quick computation to obtain the class of the dataset. It is able to work with a small dataset and provides reliable result although larger dataset would definitely increase the accuracy. However, NB is built on the assumption of independence among predictor variables while in practice, it is hardly ever to find a dataset that has no correlation between the independent variables.

In the current domain, NB can be used to predict the requirement of mechanical ventilator machine usage based on symptoms evaluated before severity of patient worsen. It is easy to implement and not computationally expensive. However, labeled data would initially be required to train the model. Such model deployment would allow medical professionals to anticipate the required resources and preparation thus improving the process flow.

In an example where NB is used in the medical domain, Silahudin *et al.* (2020) developed a NB classifier model to diagnose whether an individual is infected by COVID-19 disease based on questionnaire on symptoms experienced and travel history data. An individual can easily input the data in an online questionnaire and the output would show the likelihood of infected by COVID-19. A series of precautions and procedures are also informed to educate the user for further treatment requirement.

4. CONCLUSION

The number of COVID-19 cases has risen exponentially which has caused high velocity and large amount of data generated in the medical domain. As such, the data exhibit big data characteristic which is associated with volume, velocity, variety, veracity, and value. Therefore, five big data analytic solutions are proposed based on the symptoms evaluated to assist the medical professional to better manage and combat the COVID-19 disease. Many more big data analytic solutions can be applied as medical data exist in various format and specific models would need to be developed to tackle specific tasks.

REFERENCES

Aljameel, S. S., Khan, I. U., Aslam, N., Aljabri, M., & Alsulmi, E. S. (2021). Machine learning-based model to predict the disease severity and outcome in COVID-19 patients. *Scientific Programming, 2021.*

Alsunaidi, S. J., Almuhaideb, A. M., Ibrahim, N. M., Shaikh, F. S., Alqudaihi, K. S., Alhaidari, F. A., Khan, I. U., Aslam, N., & Alshahrani, M. S. (2021). Applications of big data analytics to control COVID-19 pandemic. *Sensors, 21*(7), 2282.

Byeon, K. H., Kim, D. W., Kim, J., Choi, B. Y., Choi, B., & Cho, K. D. (2021). Factors affecting the survival of early COVID-19 patients in South Korea: An observational study based on the Korean National Health Insurance big data. *International Journal of Infectious Diseases, 105*, 588-594. doi:https://doi.org/10.1016/j.ijid.2021.02.101

Chen, J., Li, K., Rong, H., Bilal, K., Yang, N., & Li, K. (2018). A disease diagnosis and treatment recommendation system based on big data mining and cloud computing. *Information Sciences, 435*, 124-149.

Çotoy, Y., Doğu, E., & Albayrak, Y. E. (2021). Predicting Intensive Care Unit (ICU) Requirement in COVID-19 With Artificial Neural Network. *EasyChair, 5753*.

Jeong, H., Rogers, J. A., & Xu, S. (2020). Continuous on-body sensing for the COVID-19 pandemic: Gaps and opportunities. *Science Advances, 6*(36), eabd4794. doi:doi:10.1126/sciadv.abd4794

Johns Hopkins Medicine. (2022, 24 February). *What is Coronavirus?* Retrieved from https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus

Kapil, G., Agrawal, A., & Khan, R. A. (2016, 21-22 Oct. 2016). *A study of big data characteristics.* Paper presented at the 2016 International Conference on Communication and Electronics Systems (ICCES).

Kaur, J., & Madan, N. (2015). Association rule mining: A survey. *International Journal of Hybrid Information Technology, 8*(7), 239-242.

Lee, J. Y., Kim, H. A., Huh, K., Hyun, M., Rhee, J.-Y., Jang, S., Kim, J.-Y., Peck, K. R., & Chang, H.-H. (2020). Risk Factors for Mortality and Respiratory Support in Elderly Patients Hospitalized with COVID-19 in Korea. *jkms, 35*(23), e223-220. doi:10.3346/jkms.2020.35.e223

Liu, X., Ahmad, Z., Gemeay, A. M., Abdulrahman, A. T., Hafez, E. H., & Khalil, N. (2021). Modeling the survival times of the COVID-19 patients with a new statistical model: A case study from China. *PLOS ONE, 16*(7), e0254999. doi:10.1371/journal.pone.0254999

Rochmawati, N., Hidayati, H. B., Yamasari, Y., Yustanti, W., Rakhmawati, L., Tjahyaningtijas, H. P., & Anistyasari, Y. (2020). *Covid symptom severity using decision tree.* Paper presented at the 2020 Third International Conference on Vocational Education and Electrical Engineering (ICVEE).

Silahudin, D., Henderi, & Holidin, A. (2020). Model Expert System for Diagnosis of Covid-19 Using Naïve Bayes Classifier. *IOP Conference Series: Materials Science and Engineering, 1007*(1), 012067. doi:10.1088/1757-899x/1007/1/012067

Sousa, G. J. B., Garces, T. S., Cestari, V. R. F., Florêncio, R. S., Moreira, T. M. M., & Pereira, M. L. D. (2020). Mortality and survival of COVID-19. *Epidemiology and Infection, 148*, E123. doi:10.1017/S0950268820001405

Vicenzi, M., Di Cosola, R., Ruscica, M., Ratti, A., Rota, I., Rota, F., Bollati, V., Aliberti, S., & Blasi, F. (2020). The liaison between respiratory failure and high blood pressure: evidence from COVID-19 patients. *European Respiratory Journal, 56*(1), 2001157. doi:10.1183/13993003.01157-2020

World Health Organization. (2020, 28 February). *Naming the coronavirus disease (COVID-19) and the virus that causes it*. Retrieved from https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it#:~:text=While%20related%2C%20the%20two%20viruses,the%20United%20Nations%20(FAO).

World Health Organization. (2022). *WHO Coronavirus (COVID-19) Dashboard*. Retrieved from: https://covid19.who.int/table

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., Hu, Y., Tao, Z.-W., Tian, J.-H., & Pei, Y.-Y. (2020). A new coronavirus associated with human respiratory disease in China. *Nature, 579*(7798), 265-269. doi:https://doi.org/10.1038/s41586-020-2008-3

Wu, J., Huang, J., Zhu, G., Wang, Q., Lv, Q., Huang, Y., Yu, Y., Si, X., Yi, H., & Wang, C. (2020). Elevation of blood glucose level predicts worse outcomes in hospitalized patients with COVID-19: a retrospective cohort study. *BMJ Open Diabetes Research Care, 8*(1), E001476.

Yan, Y., Yang, Y., Wang, F., Ren, H., Zhang, S., Shi, X., Yu, X., Dong, K., & care. (2020). Clinical characteristics and outcomes of patients with severe covid-19 with diabetes. *BMJ open diabetes research, 8*(1), E001343.

Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., Xiang, J., Wang, Y., Song, B., & Gu, X. (2020). Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The lancet, 395*(10229), 1054-1062.

Zoabi, Y., Deri-Rozov, S., & Shomron, N. J. n. d. m. (2021). Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *4*(1), 1-5.