Websites:
http://www.lifesciencesite.com
http://www.sciencepub.net

Emails:
editor@sciencepub.net
sciencepub@gmail.com

Life Science Journal

MARSLAND PRESS
Multidisciplinary Academic Journal Publisher

COVID-19: A Framework for SARS-CoV-2 Mutations Prediction using Machine Learning Techniques

Dr Saeed QY Al Khalidi[1], Dr Prakash Kuppuswamy[2]

[1]Department of Information science, King Khalid University, Abha, KSA
[2] Computer Networks Engineering Department, College of CS & IT, Jazan University, Jazan, KSA
prakashcnet@gmail.com, salkhalidi@yahoo.com

ABSTRACT: COVID-19 has affected entire nations and millions of people, prompting that it has been called by the WHO is known to be a worldwide pandemic. COVID-19 is difficult to compare to pandemics that have occurred in the last decade or so, including bird flu, swine flu, and SARS. Pandemic now represents an enormous challenge for clinicians, health-care workers, epidemiologists, and decision-makers outside the government sectors. As of yet, no good data are available on how the mutation of the COVID-19 variant changes the risks associated with underlying comorbidities. Risks might vary by population or setting. Mutations refer to single changes in the genetic code (genome) of a virus. Although mutations are common, they rarely affect the virus' characteristics. To prevent further human deaths, the COVID-19 variant mutation must be studied, analyzed, and predicted. It may be possible to tackle this challenge by using artificial intelligence or deep learning algorithms. There should be a global discussion regarding the prediction of mutation situations and how they should be dealt with. Artificial intelligence can be a valuable tool for healthcare operations in this pandemic situation. Moreover, healthcare professionals and a data analysis algorithm will probably arrive at the same conclusion based on the data set, however, the use of machine learning will allow for quicker and earlier diagnosis of any type of disease. This paper uses machine learning to help clinicians and medical researchers better understand COVID-19 variant mutations at various stages. In our proposed taxonomy, machine learning-based schemes are categorized based on the data pre-processing methods, the evaluation methods, and the applications.
[Dr Saeed QY Al Khalidi, Dr Prakash Kuppuswamy. **COVID-19: A Framework for SARS-CoV-2 Mutations Prediction using Machine Learning Techniques**. Life Sci J 2023;20(4):39-56]. ISSN 1097-8135 (print); ISSN 2372-613X (online). http://www.lifesciencesite.com.05.doi:10.7537/marslsj200423.05.

Key Words: Covid-19, Machine Learning, Mutation, Variant, Coronavirus, SARS-CoV-2, Prediction algorithm

1. INTRODUCTION

The novel coronavirus (SARS-CoV-2) was detected for the first time in the Hubei Province of China in December 2019 [1-2]. This virus causes a severe respiratory disease (COVID-19) that has spread rapidly around the world [3]. A COVID-19 outbreak was declared a pandemic by the World Health Organization on March 11, 2020 [4]. Because of the novelty of this disease, its clinical course and treatment are largely unknown [5]. By the end of December 2021, The EU Clinical Trials Register 41359 clinical trials with a EudraCT protocol, of which 6790 are clinical trials conducted with subjects less than 18 years old [6-7]. Over 289 million people worldwide have been infected since then, and over 5.44 million people have died from Coronavirus Disease as on December 2021 [7]. Despite unprecedented efforts by scientists and physicians to sequence, diagnose, treat, and prevent COVID-19, the disease has yet to reveal its lasting effects on individuals after the acute phase [12]. Managing the challenges during the COVID-19 pandemic [8]. As well as city lockdowns, measures such as COVID-19 are imposed when it is a serious threat [8]. As the COVID-19 pandemic accelerates, governments are warning people at high risk to be especially vigilant in observing social distancing measures because they are more likely to need critical care, including ventilation, and to die if they become ill [11].

However, in real-world situations, considering risk over multiple periods is more realistic and more interesting because we can observe the dynamic nature of the problem. COVID-19 typically spreads as "waves" of infection. It emphasizes the importance of having a multi-period analysis and being myopic is not enough. A multi-period analysis of risk poses technical challenges [8]. A number of new challenges arise as a result of COVID-19 [9-10].

As a result of inadequate adjustment for important confounding factors such as age, gender, and smoking status, insufficient follow-up [37], as well as likely under-reporting of pre-existing conditions, it is unclear the relative importance of different underlying health

conditions. Study of similar outbreaks of SARS, MERS, and influenza H1N1. As with previous outbreaks, such as SARS in 2002–2003, clinicians are confronted with new diseases for which there are few effective treatment options [13]. COVID-19 has no targeted agent at this early stage of the outbreak, so repurposing of existing antiviral drugs and corticosteroids is discussed [13–20]. It is expected that large observational clinical studies will be performed until promising targeted randomized controlled trials become available in order to evaluate potential treatment effects, as was done, for instance, for SARS, MERS and influenza H1N1on hospital mortality [28–32]. Because observational studies cannot provide causal conclusions, they cannot replace randomized controlled trials. On the other hand, they can serve as a catalyst for further research into potential treatment options.

Researchers, clinicians, health-care workers, epidemiologists, and decision makers face a formidable challenge as a result of this pandemic. As a part of this global effort, BMC Medical Research Methodology has set up a collection of articles called "Methodologies for COVID-19 research and data analysis". Rapid assessment of a dynamic research area such as COVID-19, where the body of evidence has grown at an impressive rate, requires an approach that is more direct and has a wider scope than the current gold standard methods, such as scoping and systematic reviews [31]. Many research articles have discussed the potential use of machine learning and artificial intelligence in the fight against the COVID-19 crisis [32]. A number of systematic reviews have been published on the impact of comorbidities, symptoms, and treatments of the pandemic [33-35]. The available scientific literature on COVID-19 does not allow for a comprehensive assessment. As a result, we examined the published scientific literature on COVID-19, evaluating it appropriately [36].

The level of infectious virus needed to make COVID-19 transmissible is not known; however, a carefully conducted epidemiological study that identifies the proportion of pre-symptomatic to post-symptomatic spread is the best way to determine the potential for transmission. There have been multiple instances of transmission of the COVID-19 virus before symptoms appear. Individual case studies cannot determine the proportion of pre- and post-symptom onset transmission because little or no information is available on COVID-19 asymptomatic/pre-symptomatic cases that had close contacts but did not result in transmission. Covid-19 models are highly dependent on the assumptions built into them and can sometimes be misinterpreted as, noted previously, because modeling studies can provide valuable information in the interpretation of epidemiological data [21-24].

The SARS-CoV-2 virus mutation (also called a viral mutation) occurs when the genetic sequence changes when compared to a reference sequence like Wuhan-Hu1 or USA-WA1/2020. New variants or sub-lineages of the SARS-CoV-2 virus could contain mutations that distinguish them from the reference sequences or predominant types already circulating in the population. In contrast with previous or currently circulating viruses, different strains of SARS-CoV-2 can have different characteristics, including the ability to spread more easily, exhibiting resistance to existing treatment options, and exhibiting no impact compared to previous strains [25-27].

The presence of mutations of the SARS-CoV-2 virus in a patient sample can affect test performance. Several factors can affect the performance of a test with mutations, including the sequence of the variant, the design of the test, and its prevalence in the population. Machine learning (ML) is a subclass of artificial intelligence, where algorithms analyze large data sets to detect patterns, learn from them, and execute tasks autonomously without being told how to address the problem in a particular manner. In the healthcare industry, machine learning algorithms have a lot of potential because there is a lot of data generated for each patient. Therefore, it's no surprise that there are multiple successful applications of machine learning in healthcare today that have been mentioned in figure1. A wide variety of tasks can be solved using machine learning techniques in healthcare as well as many other research applications. In healthcare, machine learning facilitates the following services.

- Classification- By using machine learning algorithms, it can determine and label the type of disease or medical case.
- Recommendations- The use of machine learning algorithms can provide medical information without the need for an active search.
- Clustering- Similar medical cases can be grouped using machine learning to analyze patterns and conduct future research.
- Prediction- Machine learning can predict future events based on current data and common trends.
- Anomaly detection- With machine learning, you can find patterns that differ from the norms in healthcare and determine whether any actions need to be taken.

- ❖ Automation- Data entry, appointment scheduling, inventory management, and other routine repetitive tasks that require too much time and effort from doctors and patients can be automated by machine learning.
- ❖ Ranking- Using machine learning, relevant information can be displayed first, making searching for it easier.

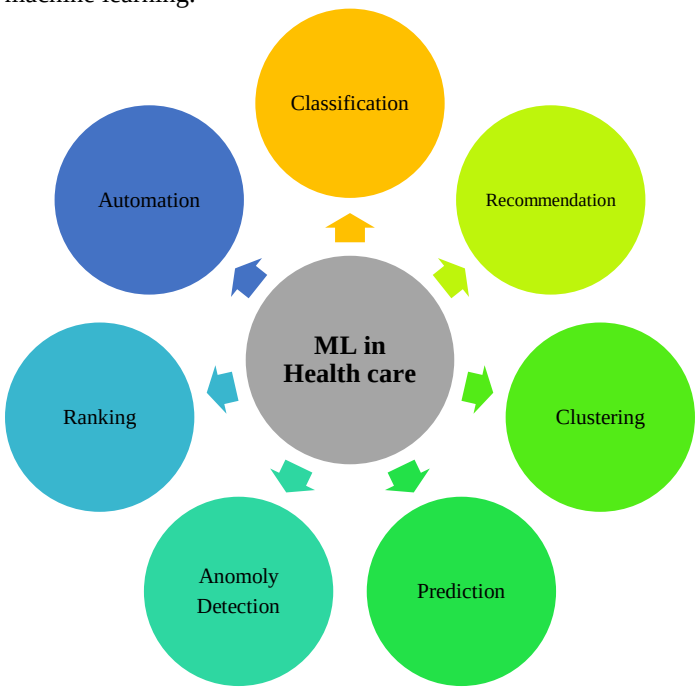

Figure 1. Machine Learning in Health care industry

2. RELATED STUDY

The paper discusses risk analysis research in logistics with special relevance to COVID-19. It reviews some of the previous TRE publications on logistics risk analysis. In this research, the article inspires more innovative risk analysis research to overcome logistics challenges during and after the COVID-19 pandemic. This will stimulate further research related to risk analysis for logistics systems. TRE, as a well-established leading journal in logistics, will continue to publish risk analysis studies in the future related to logistics and transportation. Specifically, studies looking at innovative measures to address logistics challenges during and after COVID-19 [8].

The article explains the mechanics of dataset construction, highlights key design decisions, provides an overview of how CORD-19 has been used, and previews the upcoming tools and shared tasks built around the dataset. In order to find effective treatments and management policies for COVID-19, researchers will continue to bring together the computing community, biomedical experts, and policy makers. Among the many tools that can be directed to combat the COVID-19 epidemic, automated text analysis stands out as a tool that cannot provide a full solution. Additionally, this work illustrates the value of public access to full-text literature, since it allows users to interact with and discover texts through computational access [38].

COVID-19 can currently only be treated with comprehensive support. Covid-19 mild cases, patients can often return to full health after appropriate clinical treatment. In order to determine what happened during cancer treatment, this study will analyze patients' blood parameters before and after their complete remission. The immune system is made up of immune cells is every part of the body, they help maintain a balance between health and illness. The paper examines four potential mechanisms leading to lymphocyte deficiency. Initially, the virus may directly attack those cells, killing them. Infection of lymphocytes may be facilitated by the presence of the coronavirus receptor ACE2. Viruses can also attack the lymphatic system directly. There is evidence of acute lymphocyte decline, which may be due to dysfunction in the cells themselves. It might also be due to a direct invasion by coronavirus, which attacks organs like the thymus and spleen [40].

Clinical research companies have been under tremendous pressure since the COVID-19 pandemic broke out. A redistribution of resources and a temporary suspension of face-to-face visits will inevitably restrict research in other therapeutic areas.

The clinical research center will screen participants for COVID-19 symptoms, fever, and exposure history before admitting them to the study. Participants can choose to visit remotely, when possible. Masks and physical distance are required for face-to-face interactions, and visitors are not allowed. SARS-CoV-2 infections may cause lower or higher blood pressure in hypertensive patients, worsen diabetes control in diabetics, or accelerate kidney disease progression in chronic kidney disease patients. A delay in treatment due to fear of contracting the virus can result in acute illnesses, hospitalization, and even death. Researchers, coordinators, and clinicians rekindle a sense of urgency and mission to use science to solve problems that are important to patients and the public [41].

In this systematic review and meta-analysis, two independent researchers searched for LitCOVID and Embase. Studies published before January 1, 2021 include at least 100 patients. The age of the study participants was between 17 and 87 years old. It is estimated that 80% of patients infected with SARS-CoV-2 have one or more symptoms. The five most common symptoms are fatigue, headache, attention deficit, hair loss and breathing difficulties. For a better understanding, studies need to be stratified according to gender, age, previous comorbidities, severity of COVID-19 (including asymptomatic), and duration of each symptom. explain. From a clinical point of view, a multidisciplinary team is essential to formulate preventive measures, rehabilitation techniques, and clinical management strategies. From the perspective of the overall patient, it aims to solve the care problems after COVID-19 [42].

3. SIGNIFICANCE OF RESEARCH

A disaster has struck with the COVID-19 pandemic. It affects a large number of people. In places where there is a severe pandemic, how food, medicines, masks, and other necessities can be delivered in a timely manner can save lives. As there may be city lockdowns, there may also be obstructions to transportation. Next, it is important to protect the health of service agents, volunteers, and logistics workers. We know that COVID-19 virus is highly infectious and spreads rapidly. Keeping the people who help safe is also crucial. The third challenge is how to ensure that healthcare products, vaccines and drugs are delivered safely to the people in need while operating smoothly through the respective supply chains. As such, designing the logistics system in question is a very important research topic, which should be explored in more depth in the future. The use of different vaccine techniques cannot control covid-19, due to its variants, which have been identified from 2019 onwards. A major purpose of this study is to use machine learning techniques to analyze how the covid-19 variants are changing regularly, as well as their effect on humans.

4. METHODS

In this Research article, we have applied these Machine learning techniques to predict the Covid-19 mutation through the classification of the J48 and Linear Regression prediction algorithm. The details about the COVID-19 variant mutation data set from WHO have been included. The classification models J48 and Linear Regression are used to predict further mutation of Covid-19 variants. A significant number of results have been achieved through J48 and REP Tree in areas of representing, utilizing, and learning statistical knowledge. Using trained data, which we have collected from the World Health Organization web portal, is our test option. With this interface, all algorithms could be run at the same time, and the results could be compared. Processing methods for prediction are shown in figure 2. Detailed explanations of the algorithms applied to process Covid-19 variant data sets are given in the section below.

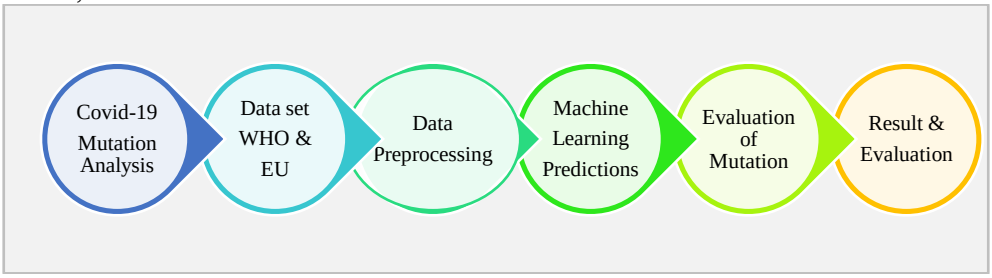

Figure 2. Covid-19 mutation prediction analysis process

The reason for applying the Weka tool is it has the facility of extending features to diagnose and predict different diseases. In addition, medical practitioners and researchers can expand their research activities with cost-effective and time-saving options. It can also help in solving the problems of clinical research using different applications of Weka. Another advantage of using Weka for the prediction of diseases is that it can easily diagnose any kind of disease with the necessary dataset.

4.1 Data Pool

A dataset is a collection of data or an instance of statistical data in which every attribute of the data represents a variable. Every instance has its own description. In order to compare the accuracy of algorithms with Weka tools for prediction of Covid-19 mutations, we used Mutation data for the prediction and classification of algorithms. For classification and accuracy of further Coronavirus mutation prediction, we used 6 attributes and 98282 instances from 120 countries. We have applied different J48, Rep Tree algorithms with the WEKA data mining tool for our data analysis purpose. This study focused on the disease (COVID-19), rather than the virus, as well as the prediction of a vulnerability to further mutation.

The data pool is collected from the WHO web portal and European Union official website, which is mentioned in figure 3 Data set. The data set is named the Covid-19 variant. WHO and EU web portals contain a vast and varied number of datasets relating to Covid-19 worldwide datasets. These data are widely used by Covid-19 Researchers and Medical experts to understand the ongoing Covid-19 situation. Various academic papers and researches have been conducted using this Covid-19 data set.

A1 location

	A	B	C	D	E	F	G
95603	United States	11/1/2021	B.1.620	0	0	103533	
95604	United States	11/1/2021	Beta	0	0	103533	
95605	United States	11/1/2021	Delta	103101	99.58	103533	
95606	United States	11/1/2021	Epsilon	0	0	103533	
95607	United States	11/1/2021	Eta	0	0	103533	
95608	United States	11/1/2021	Gamma	5	0	103533	
95609	United States	11/1/2021	Iota	0	0	103533	
95610	United States	11/1/2021	Kappa	0	0	103533	
95611	United States	11/1/2021	Lambda	0	0	103533	
95612	United States	11/1/2021	Mu	5	0	103533	
95613	United States	11/1/2021	Omicron	0	0	103533	
95614	United States	11/1/2021	S:677H.Robin1	0	0	103533	
95615	United States	11/1/2021	S:677P.Pelican	0	0	103533	
95616	United States	11/1/2021	others	414	0.42	103533	
95617	United States	11/1/2021	non_who	417	0.42	103533	
95618	United States	11/15/2021	Alpha	3	0	114760	
95619	United States	11/15/2021	B.1.1.277	0	0	114760	
95620	United States	11/15/2021	B.1.1.302	0	0	114760	
95621	United States	11/15/2021	B.1.1.519	0	0	114760	
95622	United States	11/15/2021	B.1.160	0	0	114760	
95623	United States	11/15/2021	B.1.177	0	0	114760	
95624	United States	11/15/2021	B.1.221	0	0	114760	

Source :https://github.com/owid/covid-19-data

Figure 3. Covid-19 variant Data set

4.2 Data Preprocessing

The machine learning algorithm can be used to identify the effects of mutations to determine why mutated variants of SARS-CoV-2 spread more rapidly. By doing this, we could identify mutation variants that are concerning before they spread and inform the response by health authorities. Using machine learning models for classification, the study sought to find prognostic factors for predicting Covid-19 mutation. In preprocessing, data are cleaned, missing data is replaced, data is transformed, and data imbalances are reduced. In this study, 98232 records and 120 nations' mutation data were included, as shown in figure 4 and a graphical view of attributes values is shown in figure 5. Preprocessing data plays an influential role in machine learning. For high-quality datasets, imputation and normalization were applied.

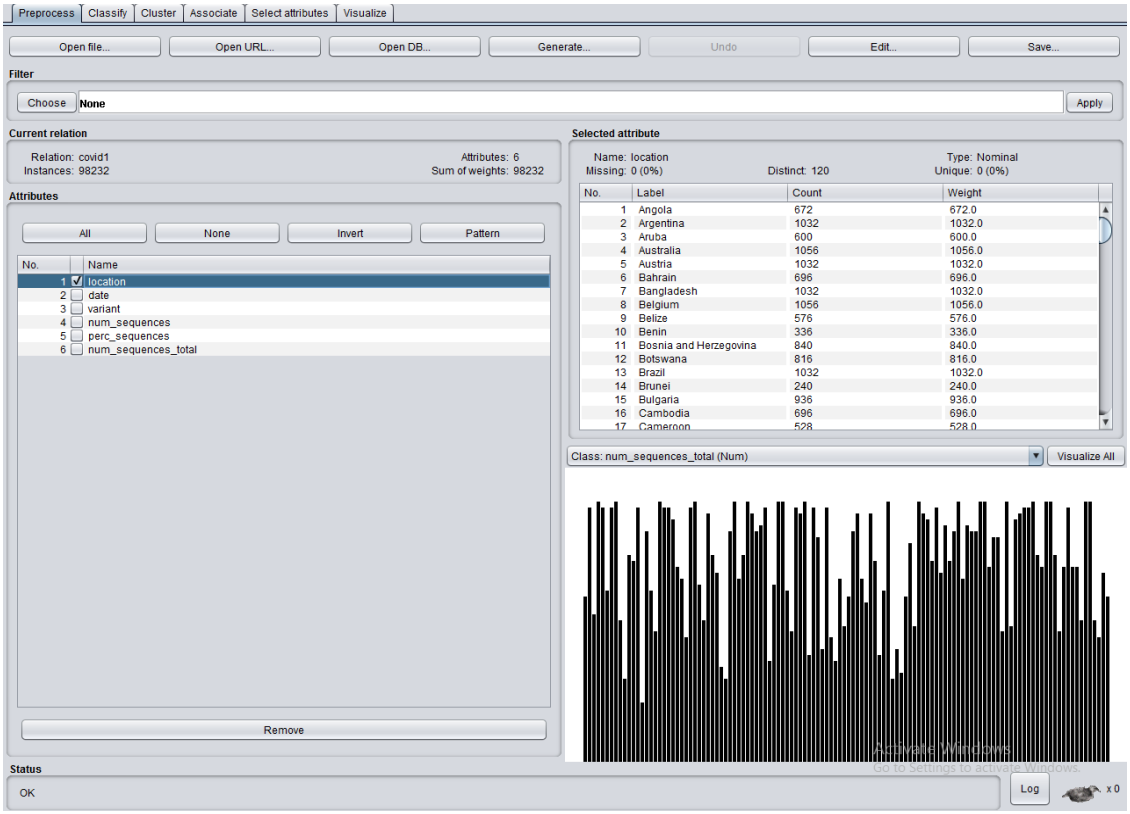

Figure 4. Data preprocessing

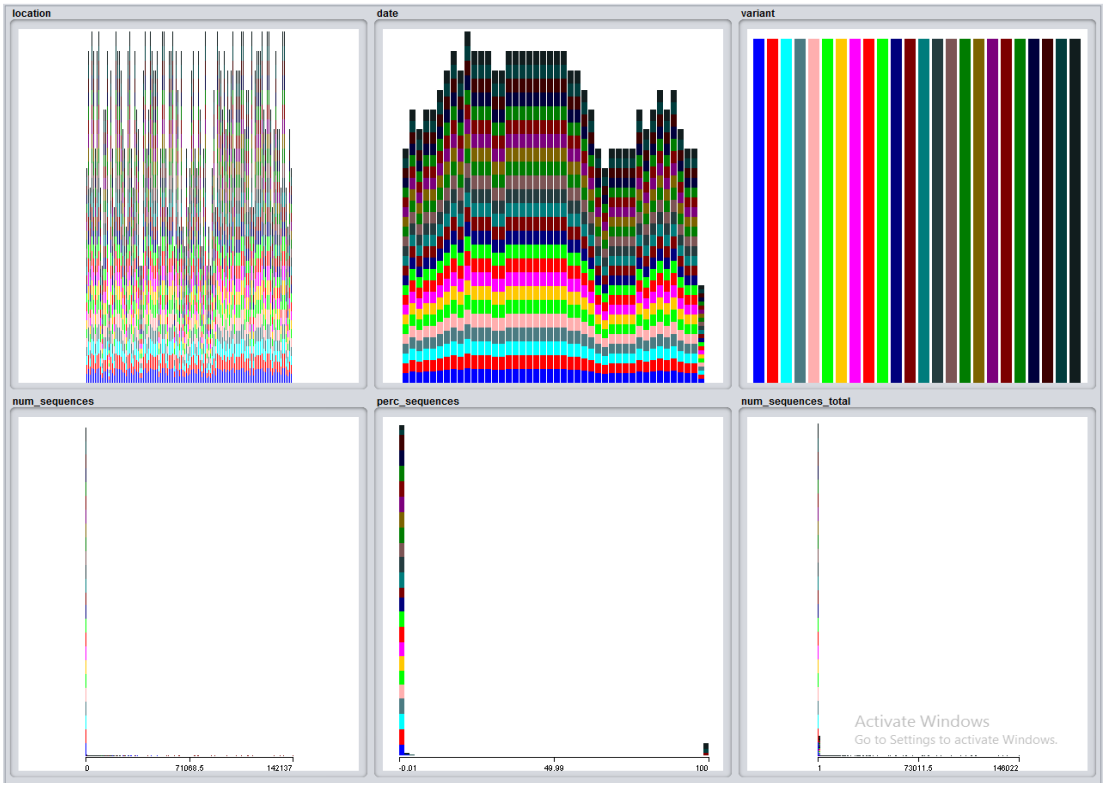

Figure 5. Graphical view of instances

4.3 Choosing Research variables

It is a process of selecting the best subset of the relevant variable for use in the prototypical structure. It selects a suitable classification for prediction accuracy. Perfect variable selection helps to evaluate the effectiveness of included variables in the training dataset. The dataset and its description, which we have derived from the WHO portal are shown in table 1.

Table 1. Covid-19 variant dataset attributes and description

ATTRIBUTES NAME	DESCRIPTION
Location	Name of Country
Date	Mutation date
Variant	Covid-19 Variant
num_sequences	Mutation sequence
Perc_sequences	Percentage of Mutation
num_sequences_total	Total sequence

4.4 Model building

The predictive classifier models were developed to accurately identify Covid-19 variant mutation. The classification model J48 and Linear Regression is used to develop prediction models. We considered these models due to their following characteristics. It is a widely and most commonly used method of empirical analysis in sociology, bio-statistics, clinical medicine, quantitative psychology, econometrics, marketing, and often uses to compare with machine learning studies. It has many advantages including high power and accuracy. The J48 classifier is built independently by applying the general technique of bootstrap aggregating and is a selected sample for the training set. It is yet very powerful model that is used to the prediction with a degree of certainty. The validation results from j48 classifier models were then combined to provide a measure of the overall performance.

Prediction statistical analysis Continuous variables were presented as the mean, standard deviation which is analyzed by the J48 classifier in addition to the Linear Regression. The performance of classification models to predict Covid-19 mutation prediction was measured by the Receiver Operating Characteristic (ROC). We also calculated the accuracy (AC), Kappa statistic, Mean absolute error (MAE) Root mean squared error (RMSE), Relative absolute error (RSE). R software (Version 3.4.2) and Weka (V.3.9.5) were used to construct a model by using classification models. Weka contains a collection of visualization tools and a graphical user interface for easily performing algorithms.

5. RESULTS & DISCUSSION

Machine learning has many applications in healthcare. For example, it can be used to automate time-consuming and complex tasks within this field. Today, the rapid and significant progress in machine learning (ML), designing faster processors, and accessing digital health data has created opportunities to improve the healthcare process. These new technologies reduce costs, accelerate proper drug discovery, and improve therapeutic results. Today, machine learning is attracting investors and major players in healthcare. We used machine learning algorithms to predict further mutation situations and vulnerability statistics based on the Covid-19 variant dataset, which received data from 120 nations and 98282 instances. Our goal was to analyze the data with classification models and define the prediction of further mutation and its effects. The J48 classifier was chosen because it performs better with a large number of datasets with a variety of attributes. Moreover, the J48 algorithm promoted the performance of the data without any obscuring effects on selected attributes. The performance metrics after attribute selection, parameter tuning, and calibration are used because this is a standard process of evaluating algorithms.

The precision of J48 performance using trained data set results is mentioned in table 2. The classification results achieved correctly were 8979 instances, i.e. 9.1406%, and incorrectly classified instances were 89253, i.e. 90.8594%. A total of 98282 instances were involved in the process of classifier results. According to further statistical analysis, kappa statistic output was 0.0519, Mean absolute error was 0.0769, Root mean squared error was 0.01961, Relative absolute error was 96.2989%, and Root relative squared error was 98.132%. table 3 shows the details of accuracy for all J48 instances.

Table 2. J48 algorithm classifier result

Result features	Output
Correctly Classified Instances	8979 i.e. 9.1406 %
Incorrectly Classified Instances	89253 i.e.90.8594 %
Kappa statistic	0.0519
Mean absolute error	0.0769
Root mean squared error	0.1961
Relative absolute error	96.2989 %
Root relative squared error	98.132 %
Total Number of Instances	98232

Table 3. J48 classifier result all instances

TP Rate	FP-Rate	Precision	Recall	F-Measr.	MCC	ROC Area	PRC Area	Class
0.126	0.015	0.262	0.126	0.170	0.158	0.661	0.129	Alpha
0.002	0.000	0.204	0.002	0.005	0.018	0.569	0.049	B.1.1.277
0.000	0.000	0.000	0.000	0.000	0.000	0.569	0.048	B.1.1.302
0.000	0.000	0.000	0.000	0.000	0.000	0.559	0.047	B.1.1.519
0.009	0.002	0.193	0.009	0.016	0.033	0.572	0.058	B.1.160
0.021	0.003	0.213	0.021	0.038	0.055	0.572	0.063	B.1.177
0.000	0.000	0.000	0.000	0.000	0.000	0.562	0.051	B.1.221
0.011	0.001	0.246	0.011	0.020	0.043	0.567	0.059	B.1.258
0.000	0.000	0.000	0.000	0.000	0.000	0.570	0.049	B.1.367
0.000	0.000	0.000	0.000	0.000	0.000	0.567	0.048	B.1.620
0.000	0.000	0.000	0.000	0.000	0.000	0.568	0.055	Beta
0.299	0.007	0.666	0.299	0.413	0.431	0.700	0.350	Delta
0.000	0.000	0.000	0.000	0.000	0.000	0.556	0.047	Epsilon
0.000	0.000	0.000	0.000	0.000	0.000	0.549	0.046	Eta
0.002	0.000	0.281	0.002	0.004	0.022	0.555	0.052	Gamma
0.000	0.000	0.000	0.000	0.000	0.000	0.559	0.047	Iota
0.000	0.000	0.000	0.000	0.000	0.000	0.562	0.047	Kappa
0.000	0.000	0.000	0.000	0.000	0.000	0.558	0.047	Lambda
0.000	0.000	0.000	0.000	0.000	0.000	0.558	0.048	Mu
0.019	0.001	0.500	0.019	0.036	0.091	0.581	0.063	Omicron
0.994	0.855	0.048	0.994	0.092	0.080	0.570	0.048	S:677H.Robin1
0.000	0.000	0.000	0.000	0.000	0.000	0.570	0.048	S:677P.Pelican
0.247	0.034	0.238	0.247	0.242	0.208	0.826	0.315	others
0.465	0.029	0.408	0.465	0.435	0.409	0.840	0.373	non_who
0.091	0.040	0.136	0.091	0.061	0.065	0.597	0.091	Weight.Avg

The purpose of this work is to predict the chances of further mutation and its vulnerability. We have taken datasets with different performance measures using machine learning. All data were preprocessed and used for test prediction. The performance results are shown in figure 6 as a plot matrix and graphical view of choosing the date, location and, variant preprocess shown in figure 7. The highlight of the resulting output is given below.

$$\text{Mean Absolute Error} = 1/n \sum_{j=1}^{n} |y_j - y_j| = 0.0769$$

$$\text{Root Relative Squared Error} = \sqrt{\frac{\sum_{i=1}^{N}(\theta_i-\theta_i).^2}{\sum_{i=1}^{N}(\theta-\theta_i).^2}} = 98/13\%$$

$$\text{Relative Absolute Error} = \frac{\sum_{i=1}^{N}|\theta_i-\theta_i|}{\sum_{i=1}^{N}|\theta_i-\theta_i|} = 96.2989\%$$

$$\text{Root Mean Squared Error} = \sqrt{1/N \sum_{i=1}^{n}(\theta_i - \theta_i).^2} = 0.1961$$

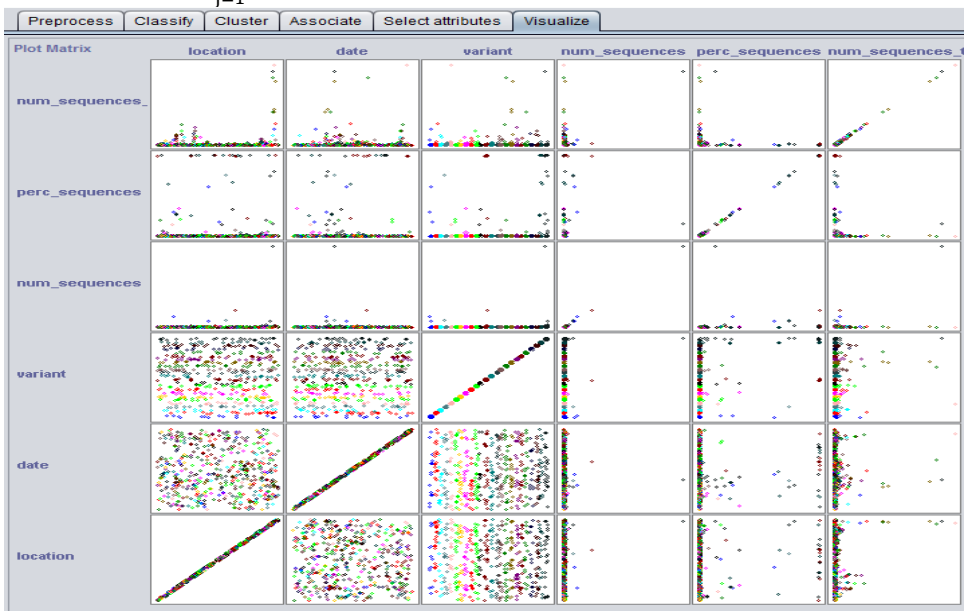

Figure 6. J48 plot matrix

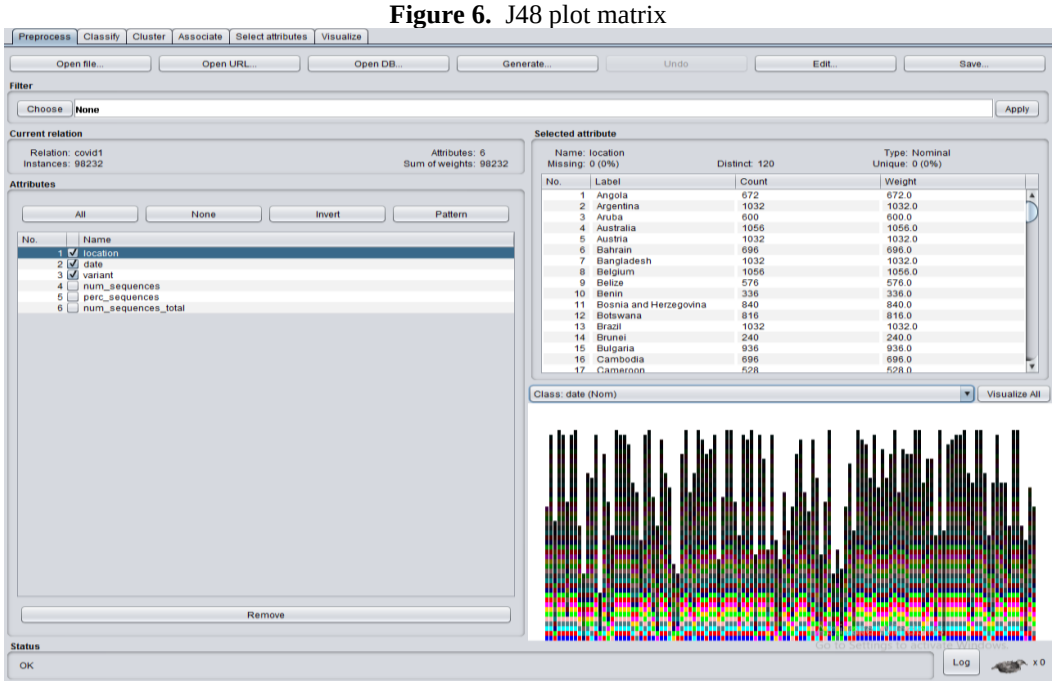

Figure 7. Process of location, date and variant

In this study, the accuracy of Covid-19 variants was classified using J48 and Linear Regression models on different data sets, and the results were compared to arrive at a feasible solution and cross-verification of the predicted result. The above algorithms used by us were applied to a prediction of the covid-19 mutation situation. The data set details are mentioned in figure 3. We performed a 10-fold cross-validation test in order to improve accuracy. For each classification, we selected training and testing samples randomly from the base set to train the model and then test it in order to estimate the classification and accuracy measure for each classifier. The attributes of covid-19 related date, country of origination, and variant types are shown in figure 7.

We have executed the J48 classification algorithm. The overall performances and results are presented in figure 8. Furthermore, the obtained results metrics are displayed in above table 2. Also, we have run 10 fold cross-validation and the results of the output are shown in figure 9.

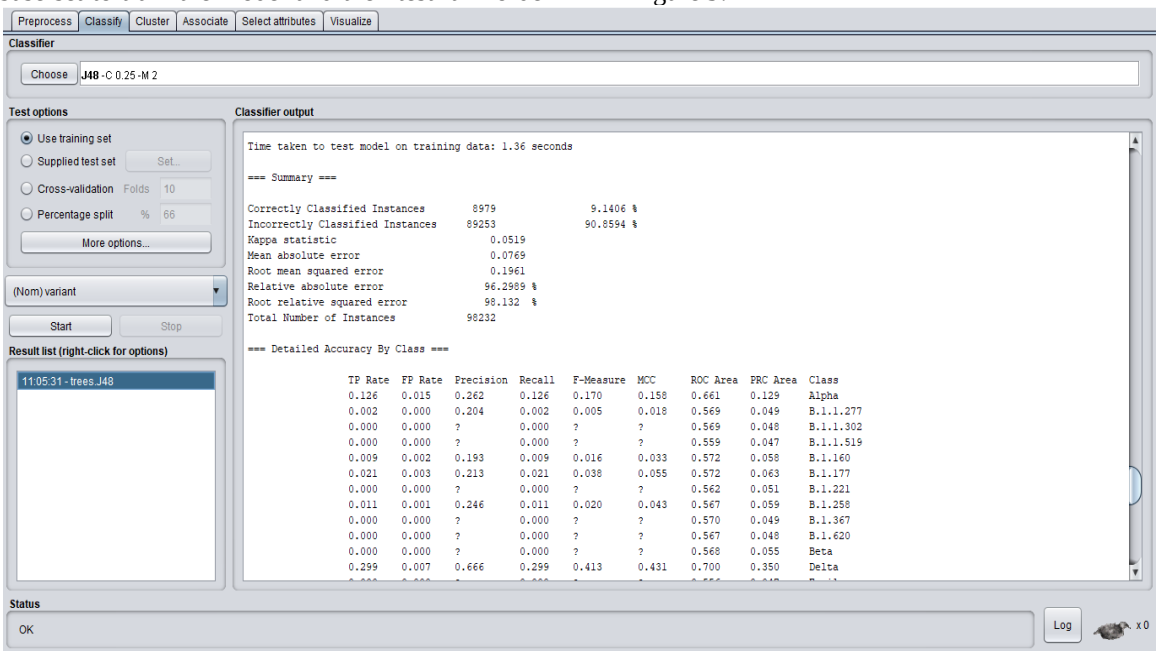

Figure 8. J48 Classification Result

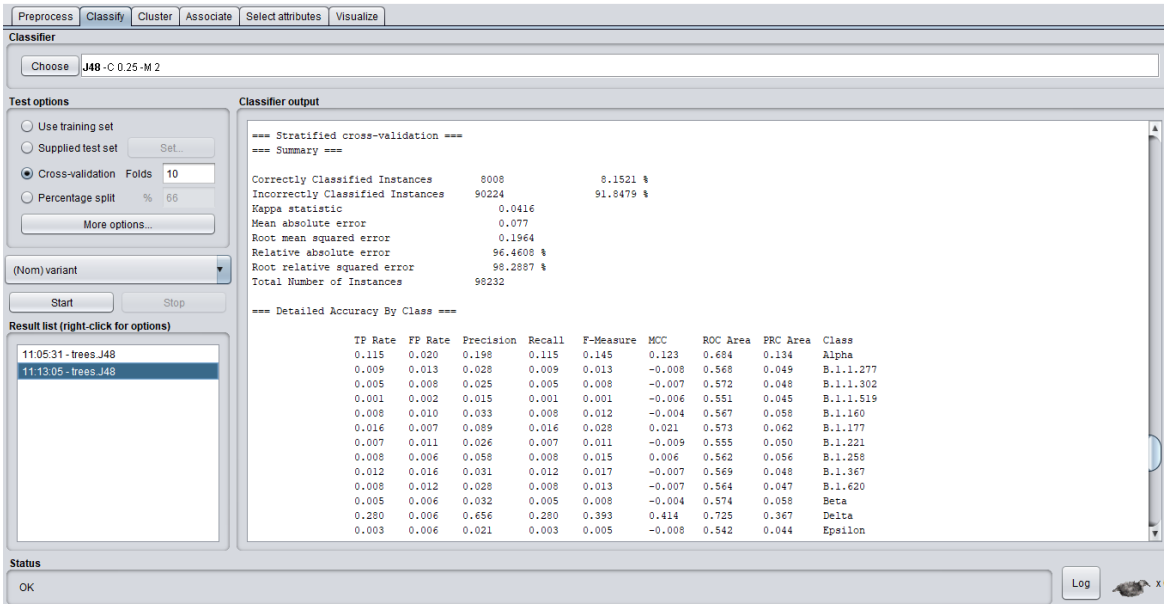

Figure 9. J48 Result using 10-fold cross validation

Next, a Linear Regression algorithm has been used in this paper for randomly choosing the attributes at each variable to allow the estimation of class probabilities. Running the algorithm, we analyze the classifier output with statistics-based output by using 90 percentage split training and 10% reminder test to make a prediction of each instance of a dataset are shown in figure 10.

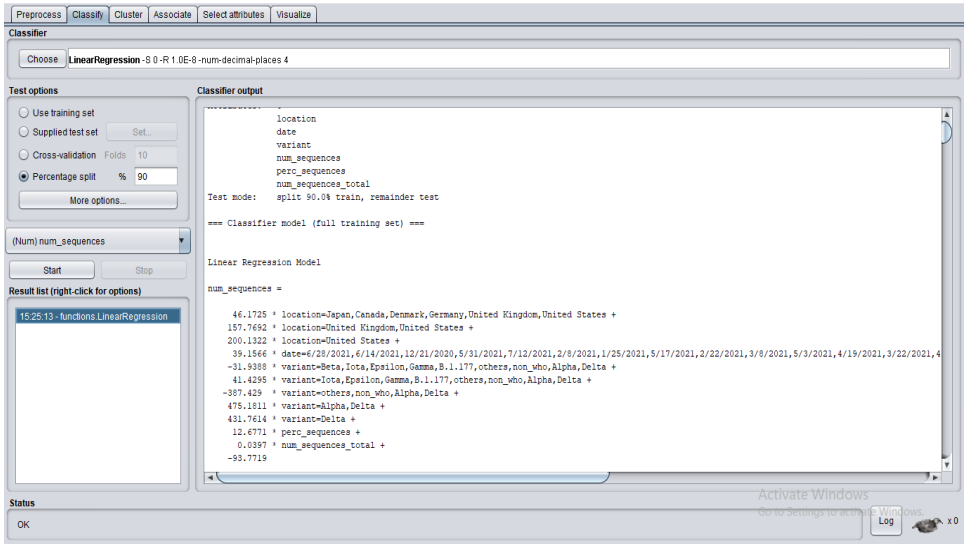

Figure 10. Linear regression classifier Result

As shown in figure11, it shows the results of the linear regression prediction process with metrics of the instances, actual value, predicted value, and error report.

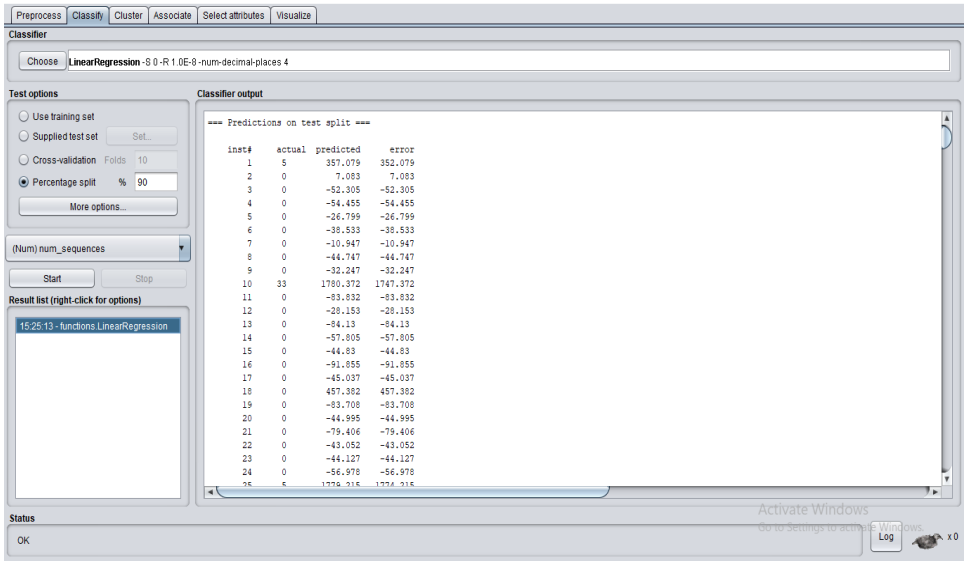

Figure 11. Linear regression prediction output

The output of the Linear regression classifier result is shown in figure 12, From the result, we have observed the value of 0.27 is correlation coefficient result, Mean absolute error (MAE)result is 216.809, Root mean squared error (RMSE) result is 1395.4758, Relative absolute error (RAE) value is 165.2355% and Root relative squared error (RRSE) value is 96.4624% from the 10% instances i.e 9823 from 98232 total instances.

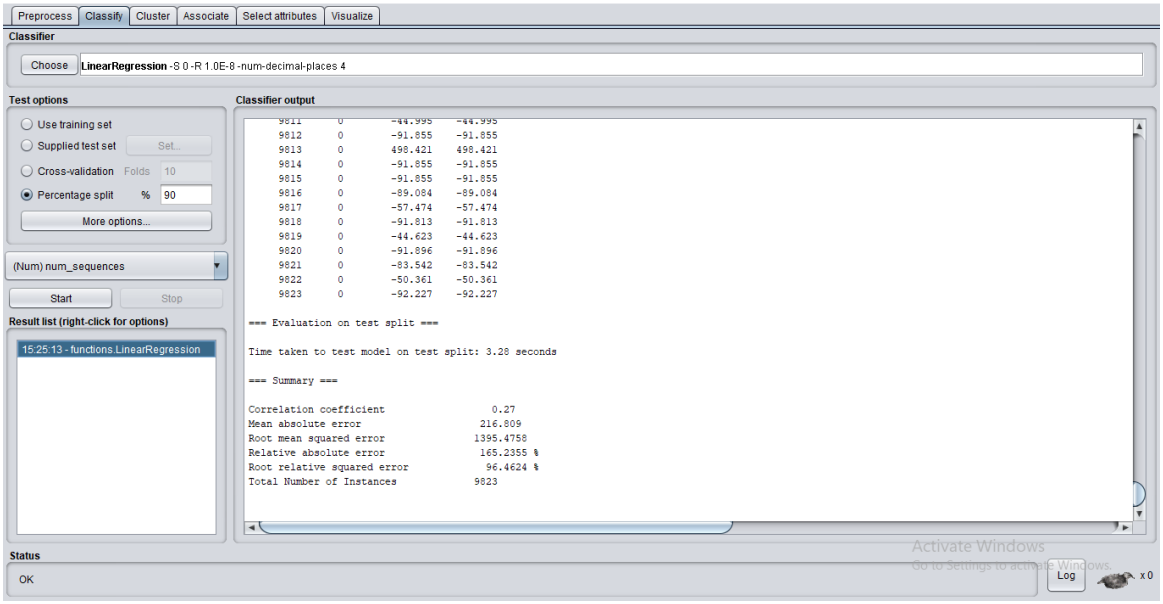

Figure 12. Linear regression prediction result

We have selected the J48 classifier once again to identify and cross verify with the linear regression to predict any other further mutations. J48 Tree has been used in this paper to decide the target value based on the time to build the model, accurately classified instances and incorrectly classified instances, as well as the accuracy of all classifiers. figure 13 shows the process results based on various attributes of the dataset used to predict machine learning models and classify their accuracy.

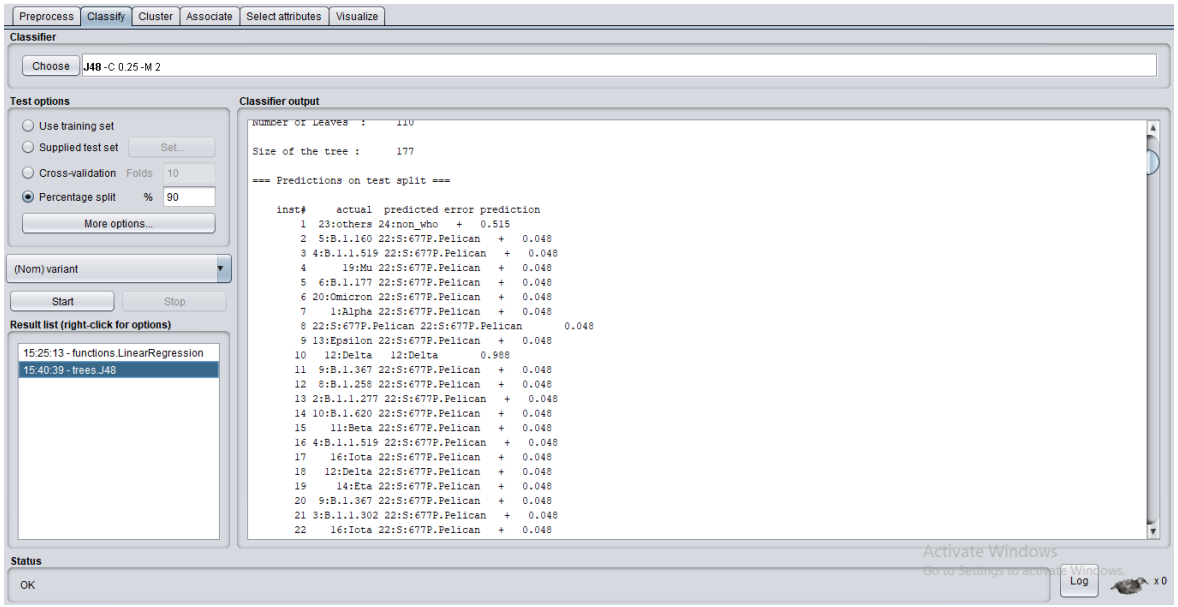

Figure 13. J48 Classification Result

After the J48 prediction classifier, we analyzed results from the obtained classifier. The output gave several statistics based on a percentage split of 90% to make a prediction of each instance of the dataset. figure 14 shows the classification accuracy achieved from this algorithm i.e. 8.134 % is the correctly classified accuracy for a batch of 9823, incorrectly classified instances 9024, Kappa statistics 0.432, Mean absolute error 0.0771, Root mean squared error 0.1966, Relative absolute error 96.5718% Root relative squared error 98.3706%, Coverage of case 96.6609%, and mean region size 83.7053. It has taken for execution duration 3.45 seconds.

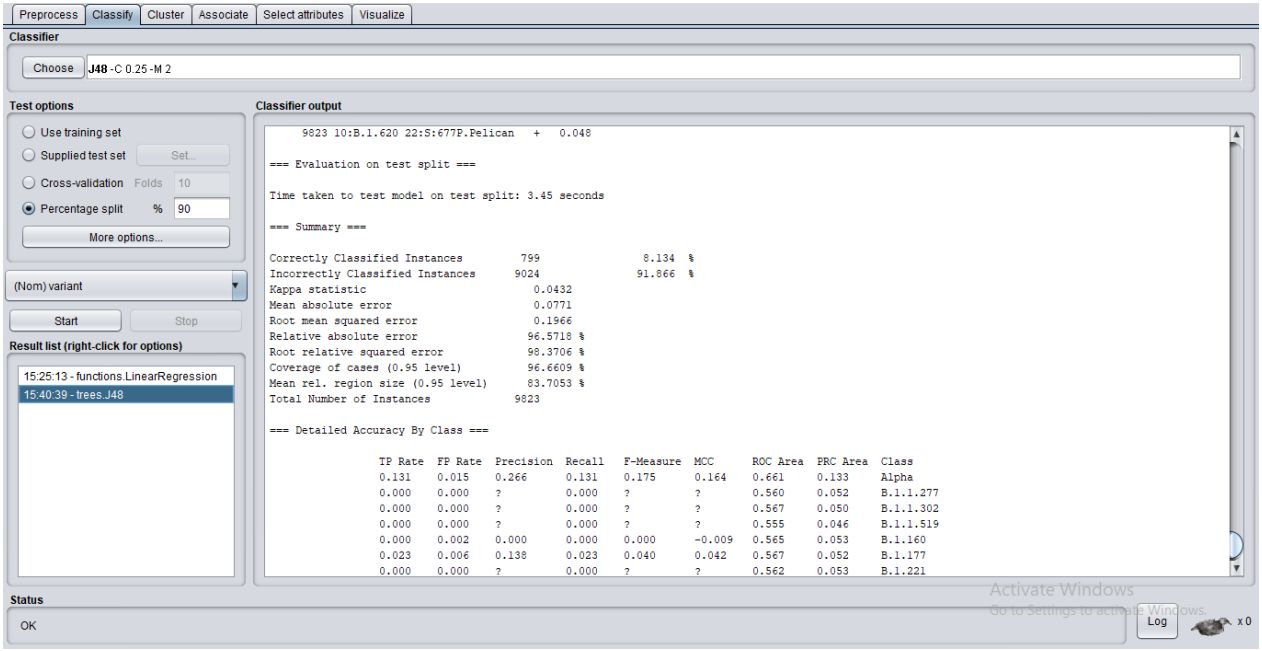

Figure 14. J48 Prediction Result

In figure 15 shows the output of detailed accuracy by class value with the metrics of TP rate, FP Rate, Precision, Recall, F-Measure, MCC, ROC area, PRC area, and class details are given. Observing the values from the below table the value of Receiver Operating Curve and Area (ROC), almost all types of covid-19 variants value more than 0.5. The weighted average of TP Rate is 0.81, FP Rate 0.038, ROC is 0.591, PRC value is 0.085 are shown. The other value of Precision, F-Measure, MCC returns an indeterminable value.

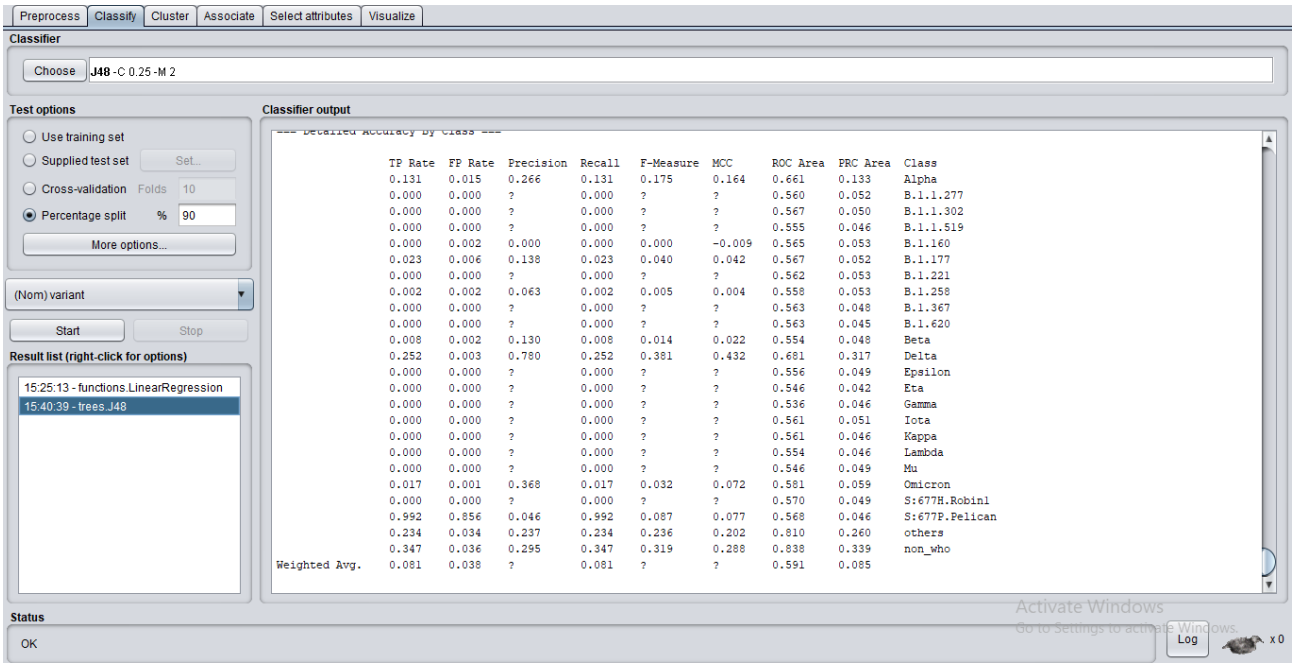

Figure 15. Detailed accuracy by class

The confusion matrix for variants of Covid-19 is shown in figure 16.

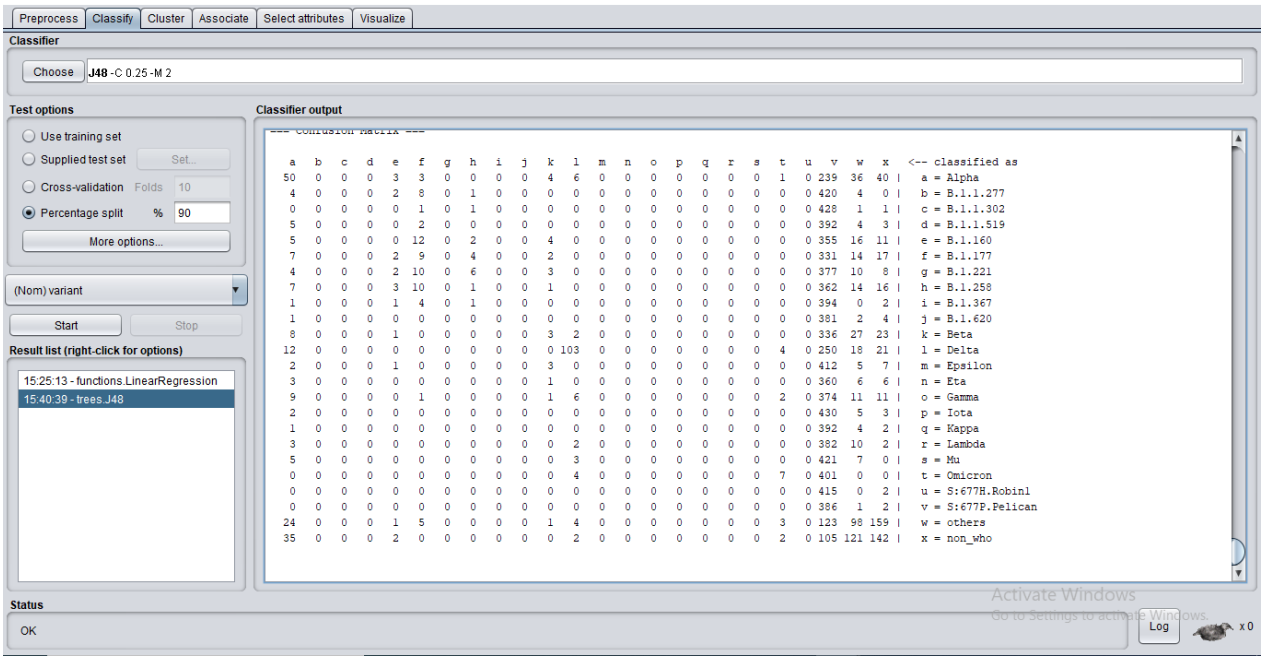

Figure 16. J48 Confusion Matrix Result

In figure 17, we have validated the result in a graphical view. The Y-axis consists of the total number of sequences, and the X-axis consists of the value of covid-19 mutation data. It shows, that Alpha, Delta and unspecified variants are spreading more than another category of viruses.

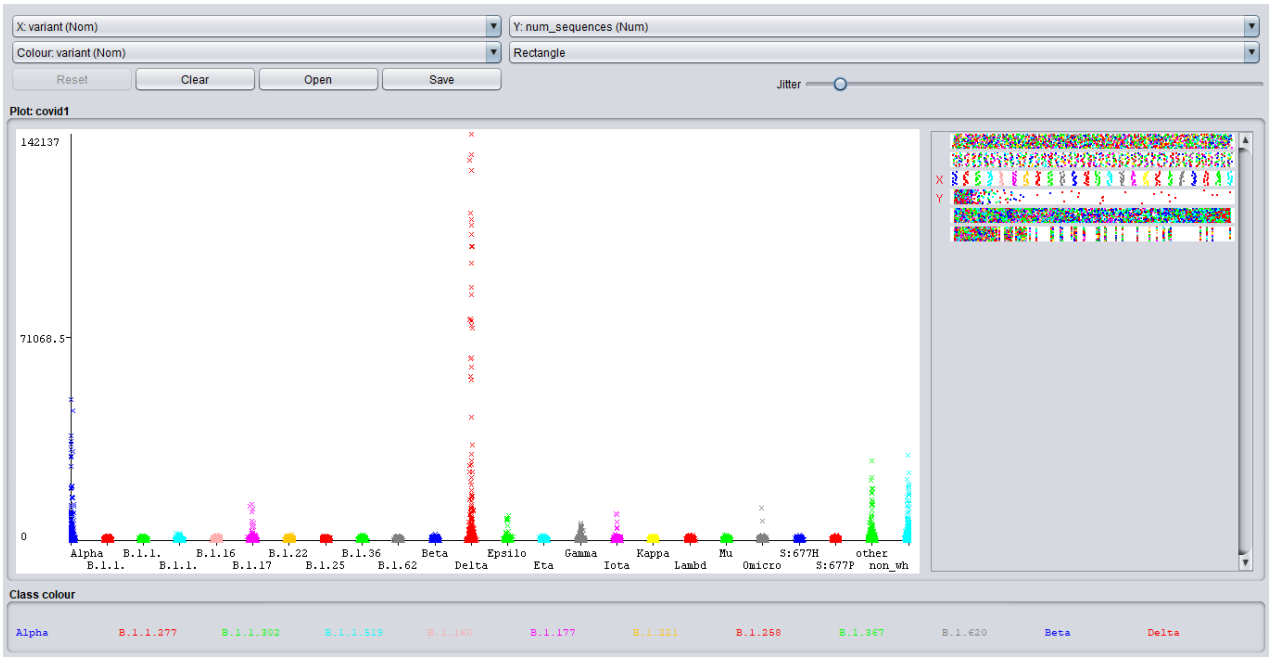

Figure 17. Covid-19 mutation

We performed experiments on different numbers of the selected attributes of the covid-19 mutation dataset. In addition, we have verified using various preprocess setups such as supervised and unsupervised with various metrics. Next, we have applied different classifiers using supervised learning, unsupervised learning, cross-validation and percentage split etc.,

6. RESEARCH FINDINGS

Healthcare data has expanded rapidly in recent years, and machine learning makes it possible to analyze massive amounts of data quickly [33]. Therefore, it is an opportunity to apply machine learning models to the care of individual patients in medical practice. There is an alarming spread of this new type of Coronavirus, which largely evades immunity. Within just two weeks of time, between Dec 25, 2021 and Jan 7, 2022, it has affected 24,151,332 people in the world [44]. The really bad news is there is every indication that more dangerous variations are on the way. Coronaviruses can gradually acquire mutations as they spread from person to person. That is how beta and delta variants arise. Unfortunately, all this might not prevent variants such as alpha and omicron and another type of virus from forming [43]. The covid-19 mutations seem to be appearing in all parts of the world. We found several significant points in the covid-19 mutation data set derived from WHO and processed using a machine learning algorithm. In the case of the J48 algorithm, its predicted ROC value is 0.591, not a favorable value. It implies that the given dataset of processed elements has a higher mutation rate, that there are further mutation possibilities. Somehow, the covid-19 death rate is better than previous covid-19 variants such as Alpha and Delta. figure 18 shows the Covid-19 mutation period and table 4 shows the death of Covid-19 at intervals.

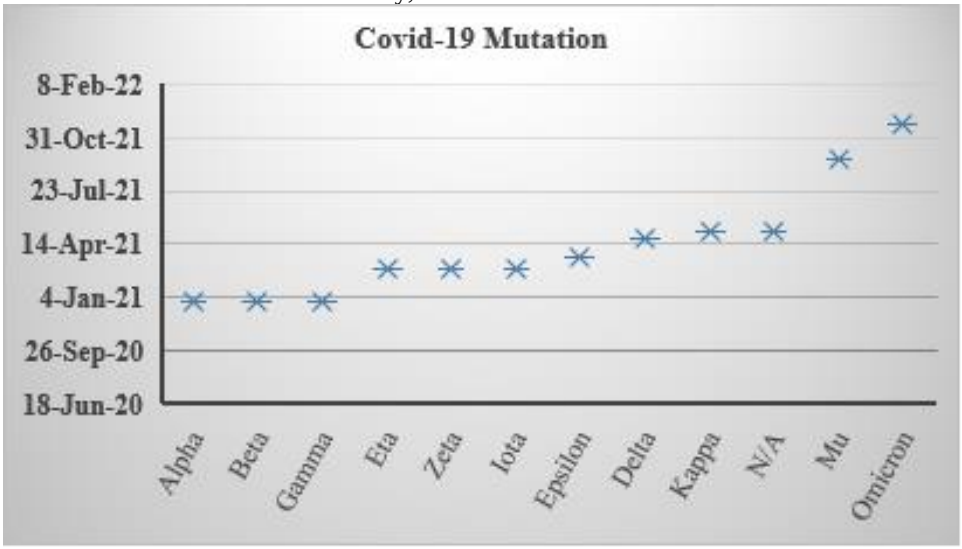

Figure 18. Mutation period

Table 4. Covid-19 death analysis table

Period	No of Deaths	Total deaths	Average(Monthly)
Upto 31 January 2020	266	266	-
01-2-2020 to 31-6-2020	573125	573391	95565
01-7-2020 to 31-12-2020	1368371	1941762	228061
01-1-2021 to 31-6-2021	2045731	3987493	340955
01-7-2021 to 31-12-2021	1465986	5453479	244331
Source: https://www.worldometers.info/coronavirus/worldwide-graphs/#total-deaths			

7. LIMITATION

The present study has several limitations that need to be addressed. In the first instance, the information was collected from the WHO with the novelty of different viruses of Covid-19. SARS-CoV-2 is still causing a lot of discussion and confusion. Nevertheless, World Health Organization itself is unable to identify how and from where the SARS-CoV-2 virus originated. In the next step, a Weka data mining tool was used to predict further mutation and vulnerability, however, the result differed from clinical and laboratory-based data experiments. Scientists, researchers and clinicians would be able to produce better outcomes due to using physical datasets instead of the available datasets to make therapeutic decisions. The experiment that we have conducted fully relies on computer-based evaluation utilizing J48 and linear regression algorithms. We used a huge dataset of 98232 instances with six attributes, which suggests that a machine learning algorithm may be able to provide the best results. When we have a huge number of data, it is difficult to validate. We analyzed a huge number

of variables that were considered a sample size, even though most of the variables are not statistically significant because they are repeated in the dataset. Due to the large dataset, there is a possibility of producing nearly unbiased estimates of the prediction results. Finally, we used a classification approach for automatic Machine Learning variables integration, but a deep learning approach would have improved prediction.

8. CONCLUSION

The purpose of this paper is to predict further mutations of Covid-19 and determine its vulnerability. For the verification and validation of prediction results, we have used the J48 algorithm and the Linear Regression algorithm. These algorithms were then analyzed based on the accuracy they provided after running them in the output window using WEKA data mining. We analysed the results based on accuracy after running these algorithms. On the basis of a given attribute value, model building time, mean absolute error, and ROC Area, J48 and linear regression algorithms were used to analyze the Covid-19 variant data set.

Our results show that the J48 algorithm predicted ROC value of 0.591 and the mean region size was 83.7053. In contrast, linear regression predicted a correlation value of 0.27, and Root mean squared result was more than 1. We have concluded that 73% of further mutations are possible based on the prediction (1-0.27 = 0.73), it is likely to spread to more than 83% of the country (Mean Region size 83.7053) and has some favorable results regarding death rates. According to the WHO report, the total number of deaths as of 31/12/2021 is 5453479. A WHO report shows that the number of dead people is falling. The highest number of deaths was 2045731 between January-2021 and June-2021, the average was 340955. There were 212542 deaths recorded in December 2021, showing a better causality rate than previous years. Our prediction result is based on the WHO database attribute only, and it has been evaluated by the prediction algorithm. We are not evaluating the data on a clinical basis, and some ambiguity may arise from computer-based data analysis. As a result, the machine learning prediction algorithm predicts that a further mutation is likely to have fewer casualties.

In the future, the research article should lead to more clinical-oriented experiments to optimize the predictive performance of these classifiers for Covid-19 virus-related diagnoses using other feature selection algorithms and optimization techniques. In addition, it may provide physicians with a better understanding of real-world clinical practice, which would let them better identify the vulnerabilities of novel Coronavirus diagnoses and preventative measures.

REFERENCES

[1] Andreas Alga, Oskar Eriksson, Martin Nordberg, “Analysis of Scientific Publications During the Early Phase of the COVID-19 Pandemic: Topic Modeling Study”, Journal of Medical Internet Research, Vol.22, issue 11, 2020.

[2] Alexandra L Phelan, Rebecca Katz , Lawrence O Gostin , “The novel coronavirus originating in Wuhan, China: challenges for global health governance”, Pub. Med, National Library of Science, Vol.8, Feb. 2020.

[3] Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, “Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia” New England Journal of Medicine, 382, 1199-1207, 2020.

[4] WHO announces COVID-19 outbreak a pandemic. Word Health Organization - Regional Office for Europe. Copenhagen, Denmark URL: http://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/news/news/2020/3/ who-announces-covid-19-outbreak-a-pandemic [accessed 2021-25-12].

[5] Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, “Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study”, The Lancet, vol.28, pp.1054-1062, March 2020.

[6] Thorlund K, Dron L, Park J, Hsu G, Forrest JI, Mills EJ, “A real-time dashboard of clinical trials for COVID-19”, Lancet Digital Health, 6, 286-287, 2020.

[7] Global Coronavirus COVID-19 Clinical Trial Tracker. URL: https://www.covid-trials.org/ [accessed 2021-12-26].

[8] Choi TM, Risk analysis in logistics systems: A research agenda during and after the COVID-19 pandemic Transportation Research Part E 145 (2021), www.elsevier.com/locate/tre 102190, Received 24 Dec 2020.

[9] Ivanov. D, “Predicting the impacts of epidemic outbreaks on global supply chains: A simulation-based analysis on the coronavirus outbreak (COVID-19/SARSCoV-2) case”, Transportation Research Part E: Logistics and Transportation Review, Vol.136, 2020.

[10] Rachel E, Peymane Adab, K K Cheng, “Covid-19: risk factors for severe disease and death” BMJ 2020, doi: 10.1136/bmj.m1198.

[11] Hannah Ritchie, Edouard Mathieu, Lucas Rodés-Guirao, Cameron Appel, Charlie Giattino, Esteban Ortiz-Ospina, Joe Hasell, Bobbie Macdonald, Diana Beltekian,Max Roser,

"Coronavirus Pandemic (COVID-19)", Published online at OurWorldInData.org, 2020.
[12] Wolkewitz M, Cooper BS, Bonten MJ, Barnett AG, Schumacher M, "Interpreting and comparing risks in the presence of competing events", BMJ, 2014.
[13] Stockman LJ, Bellamy R, Garner P, "SARS: systematic review of treatment effects", Journal of PubMed, Vol.3, 2006.
[14] Vetter P, Eckerle I, Kaiser L, "Covid-19: a puzzle with many missing pieces", BMJ, 2020.
[15] Zumla A, Hui DS, Azhar EI, Memish ZA, Maeurer M, "Reducing mortality from 2019-nCoV: host-directed therapies should be an option", Lancet, Vol.20, 2020.
[16] Del Rio C, Malani PN, "2019 novel coronavirus important information for clinicians", JAMA Network, Vol.11, pp.1039-1040,2020.
[17] Sheahan TP, Sims AC, Leist SR, "Comparative therapeutic efficacy of remdesivir and combination lopinavir, ritonavir, and interferon beta against MERS-CoV", Nat Communication, Vol.11, 2020.
[18] Russell CD, Millar JE, Baillie JK, "Clinical evidence does not support corticosteroid treatment for 2019-nCoV lung injury", Lancet, 395(10223), 2020.
[19] Shang L, Zhao J, Hu Y, Du R, Cao B. "On the use of corticosteroids for 2019- nCoV pneumonia", Lancet, 395(10225):683–4, 2020.
[20] Bouadma L, Lescure F-X, Lucet J-C, Yazdanpanah Y, Timsit J-F, "Severe SARS-CoV-2 infections: practical considerations and management strategy for intensivists", Intensive Care Med. February 2020.
[21] Yang X, Yu Y, Xu J, "Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study", Lancet Respir Med, S2213-2600(20)30079-5, 2020.
[22] Xu X-W, Wu X-X, Jiang X-G, "Clinical findings in a group of patients infected with the 2019 novel coronavirus (SARS-Cov-2) outside of Wuhan, China: retrospective case series", BMJ. 2020.
[23] Chen N, Zhou M, Dong X, "Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study", Lancet, 395, 2020.
[24] Young BE, Ong SWX, Kalimuddin S, "Epidemiologic features and clinical course of patients infected with SARS-CoV-2 in Singapore", Vol. 15, 1488-1494, JAMA Network 2020.
[25] Huang C, Wang Y, Li X, "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China", Lancet, https://doi.org/10.1016/S0140-6736(20)30183-5, 2020.
[26] Wang D, Hu B, Hu C, "Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China", JAMA Network, 2020. https://doi.org/10.1001/jama.2020.1585.
[27] Guan W, Ni Z, Hu Y, "Clinical characteristics of coronavirus disease 2019 in China", New England Journal of Medicine, 2020.
[28] Arabi YM, Shalhoub S, Mandourah Y, "Ribavirin and interferon therapy for critically ill patients with Middle East respiratory syndrome: a multicenter observational study", Clinical Infectious Diseases, June 2019.
[29] Arabi YM, Mandourah Y, Al-Hameed F, "Corticosteroid therapy for critically ill patients with Middle East respiratory syndrome", American journal of respiratory critical care med, 197(6):757–67, 2018.
[30] Delaney JW, Pinto R, Long J, "The influence of corticosteroid treatment on the outcome of influenza a(H1N1pdm09)-related critical illness", Critical Care, 2016.
[31] Munn Z, Peters MDJ, Stern C, Tufanaru C, McArthur A, Aromataris E, "Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach", BMC Med Res Methodol, 19;18(1), 2018.
[32] Kumar A, Gupta PK, Srivastava A, "A review of modern technologies for tackling COVID-19 pandemic, Diabetes and Metabolic Syndrome Clinical Research and Reviews 14(4), May 2020.
[33] Li X, Guan B, Su T, Liu W, Chen M, Bin Waleed K, "Impact of cardiovascular disease and cardiac injury on in-hospital mortality in patients with COVID-19: a systematic review and meta-analysis", Heart journal Vol.15, pp.1142-1147, May 2020.
[34] Parasa S, Desai M, Thoguluva Chandrasekar V, Patel HK, Kennedy KF, Roesch T, "Prevalence of gastrointestinal symptoms and fecal viral shedding in patients with coronavirus disease 2019: A systematic review and meta-analysis", JAMA Network, Open 3(6)-e2011335, 2020.
[35] Cortegiani A, Ingoglia G, Ippolito M, Giarratano A, Einav S, "A systematic review on the efficacy and safety of chloroquine for the treatment of COVID-19", Journal of Critical Care, 57, pp. 279-283, 2020.
[36] Magazine, "covid-19-guidance-on-social-distancing-and-for-vulnerable-people", Public

Health England. Guidance on social distancing for everyone in the UK. 2020.
[37] Kobayashi T, Jung SM, Linton NM, "Communicating the risk of death from novel coronavirus disease (COVID-19)", Journal of Clinical Medicine, 9, 2020.
[38] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, Paul Mooney, Dewey Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, Sebastian Kohlmeier, "CORD-19: The COVID-19 Open Research Dataset", PMC Lab, US National Library of Medicine National Institutes of Health, April 2020.
[39] Andreas Alga, Oskar Eriksson, Martin Nordberg, "Analysis of Scientific Publications During the Early Phase of the COVID-19 Pandemic: Topic Modeling Study", Journal of Medical Internet Research Vol.22, Issue 11, 2020.
[40] Li Tan, Qi Wang, Duanyang Zhang, Jinya Ding Qianchuan Huang, Yi-Quan Tang, Qiongshu Wang, Hongming Miao, "Lymphopenia predicts disease severity of COVID-19: a descriptive and predictive study", Signal Transduction and Targeted Therapy 5-33, 2020.
[41] Katherine R. Tuttle, "Impact of the COVID-19 pandemic on clinical research", NephROlOgy Nature Reviews, volume 16, October 2020.
[42] Sandra Lopez-Leon, Talia Wegman-Ostrosky, Carol Perelman, Rosalinda Sepulveda, Paulina A Rebolledo, Angelica Cuapio, Sonia Villapol, More than 50 long-term effects of COVID-19: a systematic review and meta-analysis", *Sci Rep* 11, 16144, 2021.
[43] Michael Le page, "Stopping corona Virus", New Scientist, Volume 252, Issues 3365–3366, December 2021.
[44] Magazine, "Centers for Disease control and prevention", United state https://covid.cdc.gov/covid-data-tracker/#variant-proportions accessed on 9/1/2022.

4/22/2023