Post-COVID-19 Health Awareness Survey and Predicting Health Awareness Through Machine Learning Techniques

Md. Samiul Islam
Assistant Professor
Department of Computer Science and Engineering
State University of Bangladesh

Md. Ashikuzzaman
Software Engineer
BSc. in CSE
Department of Computer Science and Engineering
Stamford University Bangladesh

Joy Mojumdar
BSc. in CSE
Department of Computer Science and Engineering
Stamford University Bangladesh

ABSTRACT
For the people of the 21st century, COVID-19 is a name of fear. So many people have surrendered to this disease before realizing anything. After so much has happened, many people still take it less seriously and don't follow the standard hygiene rules given by the WHO. The purpose of this research work was to predict health awareness among the people of Bangladesh in post-COVID-19 era. To perform this work, we needed a huge amount of data. For this reason, we conducted a survey through online and offline medium. Eventually, we were able to collect 1,023 data from two divisions (Dhaka and Khulna) in a very small range. The amount of data is too less. But we were not able to collect more than this because of some unwanted problems. So, ultimately, we had to proceed with this much number of data. We used 20% data for testing and the rest 80% of data for training the models. We applied 6 different types of machine learning algorithms such as Logistic Regression, Decision Tree Classifier, KNN (K-Nearest Neighbor), SVM (Support Vector Machine), Random Forest Classifier and Gradient Boosting Classifier. Random Forest Classifier outperformed all the other machine learning algorithms achieving the highest accuracy of 96.58%, highest F1 score of 96.42%, highest precision of 95.83% and highest recall of 97.26%.

General Terms
COVID-19, Survey, Machine Learning Algorithms.

Keywords
COVID-19, Health awareness, Survey, Machine Learning, Logistic Regression, Decision Tree Classifier, KNN (K Nearest Neighbors), Random Forest Classifier, SVM (Support Vector Machine), Gradient Boosting Classifier.

1. INTRODUCTION
The name of the most terrible and deadly epidemic of the 21st century is COVID-19. On 31 December 2019, China reported to the WHO cases of pneumonia with unknown causes. On 12 January, 2020, the world was first introduced to this epidemic. Bangladesh confirmed the first corona virus case on March 2, 2020. Till November, 2022, Total number of cases have reached to 2,036,268 and the total number of deaths to 29,430 [3]. To reduce the general risk of transmission, individuals were advised to abide some important and logical measures. The country faced its first challenge in this phase. It was very tough to make the individuals follow those hygiene rules as they didn't take the pandemic seriously at the very beginning. Even, when Bangladesh first got corona virus vaccine, most the common people refused to take that. After, 2 years of the pandemic, there is a large number of people who haven't taken the all 3 doses of corona virus vaccine yet. Keeping all these things in mind, we thought to start a survey on 'post-COVID-19 health awareness among the population of Bangladesh'. Our main motive was to find if the people of Bangladesh are health aware as they were in the beginning of the pandemic and predicting health awareness using various machine learning techniques on that collected data. And eventually, we applied 6 different machine learning algorithms before feature engineering and after feature engineering the data. After analyzing their performances in terms of different performance metrics such as precision, recall, f1-score and classification accuracy, we have achieved some satisfactory results Random Forest Classifier outperformed all the other machine learning algorithms achieving the highest accuracy of 96.58%, highest F1 score of 96.42%, highest precision of 95.83% and highest recall of 97.26%. The rest of the paper is arranged as follows:

The section (2) will be the literature review on breast cancer detection using machine learning techniques. The section (3) will explain the fundamental concept of the 6 machine learning algorithms being experimented. The section (4) will describe the workflow of our whole work. The section (5) will describe methodology that we have used to improve the performance of the proposed ML techniques. The section (6) will present the comparative study of the proposed algorithms. And finally, section (7) will conclude the paper.

2. LITERATURE REVIEW
COVID-19 is not the only pandemic world has ever faced. In the prior era, a lot of pandemics had occurred. But the difference between the prior and the current world is huge. With time, a lot of reliable methods have been introduced in health sector, the newest and the most trending of those is machine learning. In this research work, our main motive was to find if the people of Bangladesh are enough careful about the most recent pandemic named COVID-19. And for verifying this thing we have taken the help of different machine learning technologies. We wanted create more awareness among them showing the real picture of the current situation. As a developing country, Bangladesh is always trying its level best to distinguish itself to the world. In this case, we need to take care about the very basic fields such as health sector. To convey the importance of this particular thing a lot of researchers are trying their level best. As a successor of them, we have also tried to do something in this field. As, the current problem is about COVID-19, we have taken this disease into account and

have tried to contribute through a manual survey with some trending technologies.

A lot of researches have been done in recent times for technological advancements in healthcare. Machine learning technologies occupy a large place in this field of research. In our research work, we have collected a large number of data at first and then applied different machine learning techniques on that data to predict if an individual is health aware or not after COVID-19 pandemic.

In [1] the researchers examined the impact of health literacy on COVID-19 awareness and protective behaviors of university students in Pakistan. An online questionnaire was used to collect data from students at three universities in Punjab. The results demonstrated an urgent need for planning a needs-based literacy programme focusing specially on COVID-19 literacy in Pakistan.

In [2] a questionnaire-based survey was carried also. For the measurement of responses 5-point Likert scale was used. Age, sex, location, and education were the demographic characteristics of the study population. An open-ended question was also added at the end of the questionnaire to record the general opinion of participants on COVID-19. Results indicate that more than 50% of opinions are suggestive. People in urban are strongly opinionated that serious predictive measures are needed, some of them are satisfied with the health-protective behavior and few shared their concerns and appear to be panicked. Some of the respondents have compared the behavior of the general public with government policies while others have shared concerns about the future strategic development of COVID-19.

3. MACHINE LEARNING ALGORITHMS

Machine learning algorithm is such a method through which an AI system performs its task, basically predicting output from given input data. There are four types of machine learning algorithms:

1. Supervised.
2. Semi-supervised.
3. Unsupervised.
4. Reinforcement.

In this paper, we have proposed 6 different popular machine learning algorithms:

A. Logistic Regression: Logistic Regression falls in the category of Supervised machine learning algorithm. Its main work is to predict the probability of a binary target class. It works very well on categorical data. Logistic Regression uses the sigmoid function to convert the independent variable into the expression of probability that ranges between 0 and 1 with respect to the dependent variable.

B. KNN (K-Nearest Neighbors): The KNN (K- nearest neighbors) is a supervised machine learning algorithm that can be used to solve both classification and regression problems. In this technique, a query data point is given. And, we have to calculate the distance between the query data point and all the data points of the dataset. The distance is calculated usually through Euclidean distance. After calculating all the distances, we have to select the nearest distances based on the value of K. The value of K is usually an odd number. In this research work, we have found the K value of 3 to be the best since KNN algorithm gives the best result at this K value in terms of our dataset.

C. Decision Tree Classifier: Decision Tree is also a supervised machine learning technique that can be used for both classification and regression problems. But, in most of the cases, it is usually preferred for solving classification problems. When the Decision Tree algorithm is used on the training data, it generates a tree-structured model/classifier. When the model is generated, we provide a test data point to the model. Then the model tells us what class the data point belongs to. A generated decision tree contains two types of node: (1) Decision node and (2) Leaf node

D. Gradient Boosting Classifier: Gradient Boosting Classifiers are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model.

E. SVM (Support Vector Machine): SVM is a supervised machine learning algorithm. It can be used for both classification and regression problems. In classification problem, SVM works drawing a Hyperplane between the target classes.

F. Random Forest Classifier: Random Forest Classifier is kind of an ensemble classifier which uses Decision Tree algorithm in a randomized way. In the very first step, Random Forest Classifier generates a Bootstrap dataset taking random data points from the original dataset. In the Bootstrap dataset, having duplicate data points is possible. From the bootstrap dataset a decision tree is generated taking a subset of variables at each step of the tree.

4. WORKFLOW

Figure 1 shows the workflow of our health awareness prediction system in a concise way.

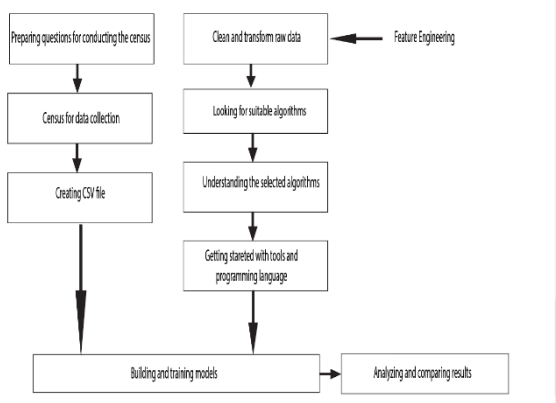

Figure 1. Workflow of our health awareness prediction system

Here, firstly we are preparing questions for conducting the survey. Then we are conducting the survey for collecting data. After collecting data, we are creating a CSV file that is needed for training and testing our proposed models. Then we are cleaning and transforming the raw data using different data mining techniques such as feature engineering. Then we are looking for some appropriate algorithms that are suitable for our prediction work. After that, we are understanding how those selected algorithms works behind. Then we are selecting the required tools and programming language. Then we are building and training the models. And at the very last phase, we

are analyzing the results that our models provide and comparing their accuracy and efficiency.

5. METHODOLOGY

A. Preparing questionnaire for survey: The first thing we had to do was create a fairly perfect questionnaire through which we would interview people for collecting data. To do this we did some researches and found some basic but important problems through which we could determine if an individual is health aware or not in the post-COVID-19 era. We selected about 14 questions such as:

1. Gender?

2. What is your educational qualification?

3. Your age?

4. Have you received the COVID-19 vaccine?

5. How many doses of COVID-19 vaccine have you taken?

6. Do you maintain a physical distance of at least 1 meter from others in public places?

7. Do you wear a mask regularly?

8. What kind of mask do you use?

9. Do you clean the cloth mask you use daily?

10. Do you change the surgical mask you use daily?

11. Do you wash your hands regularly with soap or sanitizer?

12. How many times a day do you wash your hands with soap or sanitizer?

13. Do you wash your hands with soap or sanitizer before eating?

14. Did you follow the above hygiene rules before the start of the COVID-19 pandemic?

B. Different sectors for collecting data: We conducted our survey in two divisions (Dhaka and Khulna) in a short range. We collected data from 5 different sectors:

1. School.

2. College.

3. Madrasah.

4. University.

5. Common population.

C. Method of collecting data: In this survey or more precisely in this research, we have adopted two methods for data collection:

1. Offline / Physical Survey.

2. Online Survey.

In Offline/Physical survey method, we conducted our survey in different sectors like school, college, madrasah and common population as mentioned in the previous part of this chapter. Altogether, we have been able to collect about 400 data through this method. In Online survey method of collecting data, we created a google form and shared it to the students of different universities and to some common people. In this method we have been able to collect more data than the offline survey. We have been able to collect 623 data through this method.

D. How we determined whether a person health aware or not: We prepared 14 questions for the Survey purpose. But all of those questions are not that impactful in determining the final result (whether a person is health aware or not). Out of these 14 questions we have selected 11 most impactful questions. The selected questions are given below:

1. Have you received the COVID-19 vaccine?

2. How many doses of COVID-19 vaccine have you taken?

3. Do you maintain a physical distance of at least 1 meter from others in public places?

4. Do you wear a mask regularly?

5. What kind of mask do you use?

6. Do you clean the cloth mask you use daily?

7. Do you change the surgical mask you use daily?

8. Do you wash your hands regularly with soap or sanitizer?

9. How many times a day do you wash your hands with soap or sanitizer?

10. Do you wash your hands with soap or sanitizer before eating?

11. Did you follow the above hygiene rules before the start of the COVID-19 pandemic? (optional)

Now, with these 11 questions we will make a condition through which we can easily determine whether a person is health aware or not in post-COVID-19 era. The condition is given below:

1. If the answer of question no. 1 is ‘Yes’.

2. If the answer of the question no. 2 is greater than or equal to 2.

3. If the answer of the question no. 3 is ‘Yes’.

4. If the answer of the question no. 4 is ‘Yes’.

5. If the answer of the question no. 5 is ‘Surgical mask’ and the answer of the question no. 7 is ‘Yes’. Or if the answer of the question no. 5 is ‘Cloth mask’ and the answer of the question no. 6 is ‘Yes’.

6. If the answer of the question no. 8 is ‘Yes’

7. If the answer of the question no. 9 is greater than or equal to 2.

8. If the answer of the question no.10 is 'Yes'

9. If the answer of the question no. 11 is 'Yes' or 'No'.

If all of the above conditions are true, we will determine a person as health aware in post-COVID-19 era.

D. Renaming the columns of the dataset:

we rename the all the columns for better code readability. The aliases of all the columns are given in table 1.

Table 1. Renaming columns in dataset.

Original Column Name	Renamed Column
Gender?	Q0
What is your educational qualification?	Q1
Your age?	Q2
Have you received the COVID-19 vaccine?	Q3
How many doses of COVID-19 vaccine have you taken?	Q4
Do you maintain a physical distance of at least 1 meter from others in public places?	Q5
Do you wear a mask regularly?	Q6
What kind of mask do you use?	Q7
Do you clean the cloth mask you use daily?	Q8
Do you change the surgical mask you use daily?	Q9
Do you wash your hands regularly with soap or sanitizer?	Q10
How many times a day do you wash your hands with soap or sanitizer?	Q11
Do you wash your hands with soap or sanitizer before eating?	Q12
Did you follow the above hygiene rules before the start of the COVID-19 pandemic?	Q13

E. Dataset description: Table 2 describes our dataset in a concise manner.

Table 2. Dataset description.

Feature	Values
Q0	Male/Female
Q1	Primary/ Secondary/ Higher Secondary/ Graduation or above/ Not studied/illiterate
Q2	10-15/15-20/20-25/25-30/30-35/35-40/40-45/45-50/50-55/55-60/Above 60
Q3	Yes/No
Q4	0/1/2/3
Q5	Yes/No
Q6	Yes/No
Q7	Surgical mask/Cloth mask/I don't use any kind of mask
Q8	Yes/No
Q9	Yes/No
Q10	Yes/No
Q11	0/1/2/3/ More than 3 times
Q12	Yes/No
Q13	Yes/No

After creating the dataset, we divided the rest of the work in two different categories. The two categories are:

1. Applying all the machine learning techniques without any feature engineering (i.e. applying on noisy data) and analyzing the performances.

2. Applying all the machine learning techniques with feature engineering (i.e. applying on cleaned data) and analyzing the performances.

F. Transforming data from string to numeric: we transformed all the data from string to numeric version because of easy and quick coding facility. Table 3 shows the data after transforming to numeric.

Table 3. Dataset description after transforming data from string to numeric.

Feature	Values	Values after transforming
Q0	Male/Female	1/0
Q1	Primary/ Secondary/ Higher Secondary/	0/1/2/3/4/5

	Graduation or above/ Not studied/illiterate	
Q2	10-15/15-20/20-25/25-30/30-35/35-40/40-45/45-50/50-55/55-60/Above 60	0/1/2/3/4/5/6/7/8/9/10
Q3	Yes/No	1/0
Q4	0/1/2/3	0/1/2/3
Q5	Yes/No	1/0
Q6	Yes/No	1/0
Q7	Surgical mask/Cloth mask/I don't use any kind of mask	0/1/2
Q8	Yes/No	1/0
Q9	Yes/No	1/0
Q10	Yes/No	1/0
Q11	0/1/2/3/ More than 3 times	0/1/2/3/4
Q12	Yes/No	1/0
Q13	Yes/No	1/0

G. Creating target feature: As our final result we needed to identify whether a person health aware or not. For this reason, we needed create an extra target feature named 'Class'. This feature was created programmatically using the condition discussed before at the very first of this section. Let's take a look at that condition again:

if Q3 == 1 and Q4 >= 2 and Q5 == 1 and Q6 == 1 and ((Q7 == 0 and Q9 == 1) or (Q7 == 1 and Q8 == 1)) and Q10 == 1 and Q11 >= 2 and Q12 == 1 and (Q13 == 1 or Q13 == 0):

print(1) [Here, 1 means 'the person is health aware']

else:

print(0) [Here, 0 means 'the person is not health aware']

This condition is based on some rules provided by WHO [4].

After this step our dataset was looking as shown in figure 2

	Q0	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Class
0	1	3	2.0	1	3	1	1	0.0	1	1	1	4	1	0	1
1	0	2	2.0	1	3	1	1	1.0	1	1	1	4	1	1	1
2	1	3	2.0	1	3	1	1	1.0	1	0	1	4	1	1	1
3	1	3	3.0	1	3	1	1	0.0	1	1	1	4	1	0	1
4	1	3	2.0	1	3	1	1	0.0	1	1	1	4	1	1	1

Figure 2. Dataset after adding target attribute 'Class'.

After this step, we have 680 people who are not health aware and 343 people who are heath aware. Figure 3 shows the countplot for 'Class' attribute of initial dataset.

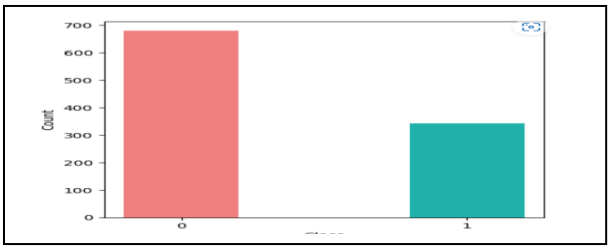

Figure 3. Countplot for 'Class' attribute.

H. Handling missing values: There were some missing values in the 'Q7' attribute. Basically, handling missing is the part of feature engineering. But, we had to do it in very first stage, because some machine learning algorithms like Logistic Regression has no possibility to reasonably deal with missing values. In this part, we filled the missing values with the mean value of that column.

I. Feature Engineering: In machine learning, feature engineering is basically the pre-processing step. Through this process, we are able to extract important features from raw data. Now, we will also see how the feature engineering techniques affected the overall performance of the all the models and how we got a reliable result.

J. Plotting Boxplot to identify outliers: In machine learning, outlier is a datapoint that is markedly differs from the rest of the data points. It basically shows the variables that should not be considered when we train our model on our dataset. There are several ways of identifying this outlier. One of them is Boxplot. Figure 4 shows the boxplot for each of the attributes of our dataset.

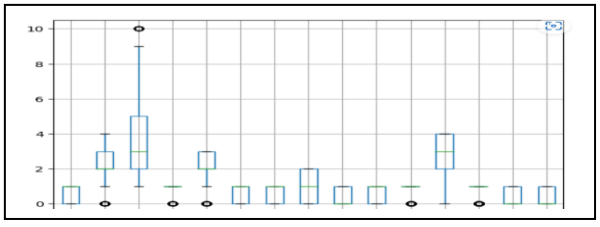

Figure 4. Boxplot for all attributes to identify outliers.

0Figure 4 shows that, 'Q1', 'Q2', 'Q3', 'Q4', 'Q10' and 'Q12' attributes contain some outliers. After removing all of these outliers using 3-standard deviation, we were able to reduce 126 instances from the initial 1,023 instances. Now, we will perform the rest of the works on (1,023 - 126) = 897 instances out of 1,023.

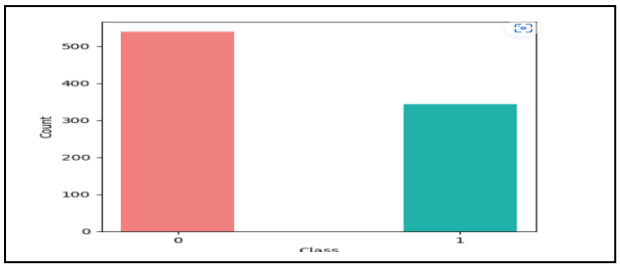

Figure 5. Countplot for 'Class' attribute after removing outliers

From figure 5, we can see that we have now 343 health aware people and 554 people who are not health aware.

After removing the outliers, we used corr() function to find the correlation among the columns in the Dataframe using the 'Pearson' method. We used the threshold value greater than 0.4. And we found the most important features such as Q5, Q6, Q7, and Q8. Figure 6 shows a sample of the result.

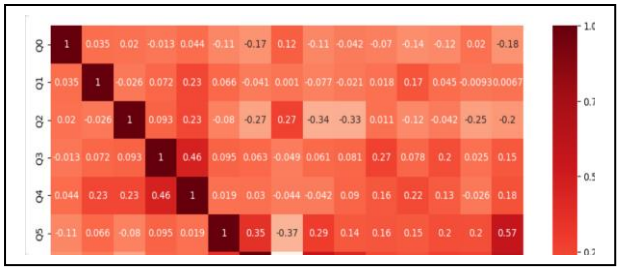

Figure 6. Finding Correlation of Other Variables with Output Variable to Find Out the Most Important Features

K. Applying PCA (Principal Component Analysis) to find the most important features: In machine learning, PCA (Principal Component Analysis) is a technique used to reduce dimensions. When we work on real-life machine learning problems, there may be a thousand columns or features. In this circumstance, we need to find the most important features to correctly classify a particular problem. PCA plays an important role in this case. PCA helps to reduce the number of columns and as a result the overall complexity reduces. It reduces the number of columns that doesn't mean it removes the actual columns. It simply produces new variables that are constructed as the mixtures of the initial variables. We set the parameter 'n_components' as 4 to produce 4 new variables that are the most important. Then we applied all of our algorithms on the dataset with these new important variables.

6. COMPARATIVE STUDY

A. Performance analysis of all the machine learning algorithms almost without any feature engineering: Before applying the algorithms we splitted our dataset using train_test_split() function. We used 20% data for testing and the rest 80% of data for training the models. Table 4 shows the performances of all the machine learning techniques after applying on the raw data which is almost without any feature engineering.

Table 4. Performance analysis of all the techniques without feature engineering.

Algorithms	Accuracy	F1 score	Precision	Recall
Logistic Regression	90.73%	89.25%	88.50%	90.15%
Decision Tree Classifier	90.73%	89.73%	88.21%	92.89%
KNN	90.24%	88.91%	87.70%	90.72%
Random Forest Classifier	**93.17%**	**92.35%**	**90.78%**	**95.10%**
SVM	91.70%	90.46%	89.46%	91.77%
Gradient Boosting Classifier	92.19%	91.13%	89.84%	93.03%

Here we can see that, without any feature engineering, Random Forest Classifier outperforms all the other machine learning algorithms achieving the highest accuracy of 93.17%, highest F1 score of 92.35%, highest precision of 90.78% and highest recall of 95.10%.

B. Performance analysis of all the machine learning algorithms after feature engineering: Table 5 shows the performances of all the machine learning techniques after applying on the dataset which is almost with feature engineering.

Table 5: Performance analysis of all the techniques with feature engineering.

Algorithms	Accuracy	F1 score	Precision	Recall
Logistic Regression	95.60%	95.37%	94.91%	95.96%
Decision Tree Classifier	96.41%	96.26%	95.60%	97.20%
KNN	90.24%	89.70%	89.36%	90.11%
Random Forest Classifier	**96.58%**	**96.42%**	**95.83%**	**97.26%**
SVM	95.12%	94.89%	94.28%	95.83%
Gradient Boosting Classifier	95.60%	95.37%	94.91%	95.96%

Here we can see that, after feature engineering, Random Forest Classifier again outperforms all the other machine learning algorithms achieving the highest accuracy of 96.58%, highest F1 score of 96.42%, highest precision of 95.83% and highest recall of 97.26%.

7. CONCLUSION

Our motive was to predict health awareness in post-COVID-19 era among the people of Bangladesh. For this purpose, we conducted a survey to collect data. We conducted the survey online and offline combinedly. We reached too different types of people such as students and common people. After collecting data, we created a dataset. On that dataset, we applied different machine learning algorithms before feature engineering and after feature engineering the data. We could not rely on the result obtained using data which was raw and not feature engineered because of the possibility of overfitting. But, we got a reliable and efficient result after applying feature engineering techniques such as removing outliers using 3-standard deviation and PCA (Principal Component Analysis). Random Forest Classifier outperformed all the other machine learning algorithms achieving the highest accuracy of 96.58%, highest F1 score of 96.42%, highest precision of 95.83% and highest recall of 97.26%.

In this research work, we have used Jupyter Notebook (version 6.4.5) as our work environment. The PC configuration was:

- Processor: AMD Ryzen 5 3500U with Radeon Vega Mobile Gfx 2.10 GHz.
- Ram: 8.00 GB.
- System type: 64-bit operating system, x64-based processor.
- Operating System: Windows 11 Home Single Language Edition.

8. ACKNOWLEDGMENTS

We, the authors of this paper are really grateful to Mr. Adnan Ferdous Ashrafi, Senior Lecturer at the Department of Computer Science and Engineering, Stamford University Bangladesh for his enormous support and valuable guidance. We would also like to show our appreciation to the anonymous reviewers for their invaluable comments and suggestions.

9. REFERENCES

[1] Naveed, Muhammad Asif, and Rozeen Shaukat. "Health literacy predicts COVID-19 awareness and protective behaviours of university students." *Health Information & Libraries Journal* 39.1 (2022): 46-58.

[2] Qazi, Atika, et al. "Analyzing situational awareness through public opinion to predict adoption of social distancing amid pandemic COVID-19." *Journal of medical virology* 92.7 (2020): 849-855.

[3] Covid19.who.int – region-wise COVID information: https://COVID19.who.int/region/searo/country/bd.

[4] Novel-coronavirus-2019 advice for people: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public