COVID-19 Future Forecasting using supervised Machine learning

Divyashree.G[1], Mrs.Meghashree.A.C[2] and Dr.G.F.Ali Ahammed[3]
[1]*PG Scholar,* [2]*Assistant Professor,*[3]*Associate Professor*
[1,2,3]*Visvesvaraya Technological University Centre for Post Graduate Studies, Mysuru*

Abstract*: Machine learning (ML) is the study of computer algorithm that improves automatically through experience and by the use of data, and it is based on forecasting mechanisms have proved their signification to outcomes to improve the decision making on the future course of actions. The ML has proved itself as a prominent field of study over the last decade by solving many very complex and sophisticated real world problems such as health care, autonomous vehicle (AV), business applications, natural language processing (NLP), and intelligent robots. This study demonstrates the capability of ML models to forecast the number of upcoming patients affected by covid-19 which is presently consider as a potential threat to mankind. In particular, four standard forecasting models, such as liner regression, least absolute shrinkage and selection operator, support vector machine, and exponential smoothing is used for study to forecast the threatening factors of COVID-19. This study aims to provide an early forecast model for the spread an early novel disease corona virus. Three types of prediction are detected by the model such are [1] the number of newly infected cases, [2] the number of deaths, and [3]the number of recoveries in the next 10days. This are the three results detected from this model this is most important application in the pandemic covid year.*

Keywords: **Corona virus dieses-19 (covid-19), Machine learning (ML), AV, NLP, LR, ES, SVM**.

I. INTRODUCTION

Machine learning is a method of data analysis that automates analytical model building.ML is generally considered to be a subfield of artificial intelligence, and even a subfield of computer science in some perspectives.[1].Corona viruses are the group of viruses that cause and effect the disorders in human respiratory system.[2] This paper presents the analysis and prevention of corona viruses, corona viruses are the microscopic that replicate only inside the living cells of organisms,[3].Forecasting of spread of corona virus is one of the challenges in this pandemic year. ML model is used for modeling real-world problems.[4Machine learning model are the most significant in forecasting,[5].The Algorithm are used to guide the future forecasting actions.[6]In ML there are lots of studies done to prevent the different diseases,[7] For example diabetic disorder,[8].Breast cancer prediction.[9]And focused on prevention of corona viruses future forecasting techniques, this prediction system is very useful in the decision making of COVID-19 prevention. This paper aims to focus on future forecasting of corona virus spread and prevention.[10]In 2019 December corona virus first found in Wuhan city in china, and made an very serious threat on human life. [11] COVID-19 effect on human body like respiratory system breathing problems, and multi organ failure and etc.[12]The most important challenging aspect to control the spread of corona virus, it will spread from person to person and have some symptoms like cold, fever and cough.[13] To develop the contribution for prevention of corona virus, this study develop the future forecasting of

COVID-19.[14] The study gives the three prediction are newly effected cases, death cases, recoveries.[14] This results will give the next upcoming 10 days results. Here we are using supervised machine learning regression models such as linear regression (LR), and support vector machine (SVM), and exponential smoothing (ES), and least absolute shrinking and selection operation. This machine learning has a trained data set of corona virus affected patients.

This study has some the key points which are listed below.

- ES has time-series datasets are very limited entries.
- Different ML algorithms are perform better in prediction
- ML algorithm require most large amount of data sets to predict the future.

In our proposed system we use the efficient data science algorithm for corona virus prediction and future forecasting.

The methodology of this study gives the comparisation of different studies. Different machine learning models gives different limitations and the data sets for the researches work for the future forecasting of covid -19 diseases. and advantage over existing models. The related analysis of our scheme against some different common possible algorithm in ML techniques.

1. Dataset collection: it includes data collection an understanding the data to study the pattern and trends which helps in prediction and evaluating the result.
2. Data preprocessing: this phase of model handles inconsistence data in order to get more accurate and précised result this data set contains missing values. so we imputed missing values for few selected attributes like glucose level, blood pressure etc.
3. ML classification: in this phase we have implemented random forest and K-nearest neighbor classification on the data set to classify each patient. Before performing ML algorithms, highly correlated attributes were found which has glucose and age. After implementation of this algorithm will get class labels of each record.

1) Algorithm 1:Naïve Bayes algorithm:

The Naïve Classifier algorithm can be implemented as shown in Algorithm. (Naïve Classifier has the ability to predict class membership probabilities of Diabetes as normal or abnormal such as the probability that a given sample belongs to a particular class. Through create Likelihood by finding the probabilities based on the Bayes theorem. The most widely used type of Bayesian Network for classification is the Naïve Bayesian's, which has the highest accuracy value of up to 99.51% respectively. The Bayesian Network applies the Naïve Bayes theorem which firmly assumes that the occurrence of any particular attribute in a class is not related to the presence of any other attribute, making it much more advantageous, efficient and independent. The Naïve Bayesian is based on the conditional probability (given a set of features, the probability of occurrence of certain results

2) K-nearest neighbor algorithm**:**

K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new the similarity. This means when new data appears then case into the category that is most similar to the available categories. K-NN algorithm stores all the

available data and classifies a new data point based on it can be easily classified into a well suite category by using K- NN algorithm. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. KNN algorithm at the training phase just stores the dataset and when it gets new data, and then it classifies that data into a category that is much similar to the new data.

2. System design

The purpose of the design phase is to plan a solution of the problem specified by the requirements document. This phase is the first step in moving from the problem domain to the solution domain. In other words, starting with what is needed; design takes us toward how to satisfy the needs. The design of a system is perhaps the most critical factor affecting the quality of the software; it has a major impact on the later phases particularly testing and maintenance.

The design activity often results in three separate outputs

- Architecture design.
- High level design.
- Detailed design.

According to Software Engineering the approach adopted to develop this project is the Iterative waterfall Model. The iterative waterfall Model is a systematic approach that begins at the feasibility study phase and progress through analysis, design, coding, testing, integration and maintenance. Feedback paths are there in each phase to its preceding phase as show in the fig to allow the correction of the errors committed during a phase that are detected in later phase.

The context-level data flow diagram first, which shows the interaction between the system and external agents which act as data sources and data sinks. On the context diagram (also known as the 'Level 0 DFD') the system's interactions with the outside world are modeled purely in terms of data flows across the system boundary. The context diagram shows the entire system as a single process, and gives no clues as to its internal organization. This context-level DFD is next "exploded", to produce a Level 1 DFD that shows some of the detail of the system being modeled. The Level 1 DFD shows how the system is divided into sub-systems (processes), each of which deals with one or more of the data flows to or from an external agent, and which together provide all of the functionality of the system as a whole. It also identifies internal data stores that must be present in order for the system to do its job, and shows the flow of data between the various parts of the system.

.

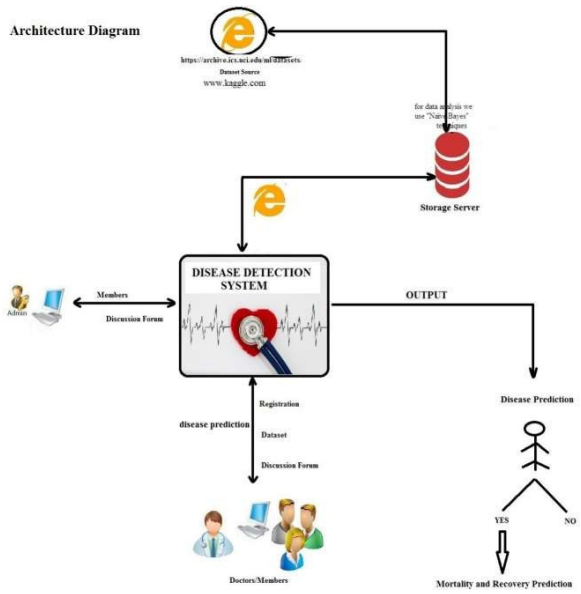

Figure 1: Architecture diagram

Nowadays, health care industries are providing several benefits like fraud detection in health insurance, availability of medical facilities to patients at inexpensive prices, identification of smarter treatment methodologies, and construction of effective health-care policies, effective hospital resource management, better customer relation, improved patient care and hospital infection control. Stroke type detection is also one of the significant areas of research in medical. There is no automation for diabetes disease prediction.

In context dataflow diagram contains some following information.

- Staff Creation Module (Admin)**:** Administrator of the system creates the staffs (specialist, receptionist) and manages the staffs and sets the unique id and password for each staff.
- Data-set Module**:** Admin manages the data-set required for the covid prediction. Here admin uploads the old data into server which includes covid disease patients data with related constraints/parameters and results.
- Prediction Module (Doctor): This is the core module of the project where system accepts the input given by the disease specialist. This module predicts the final output weather patient is classified to “Yes” or “No”. We make use of classification rules techniques for the output prediction which is one the efficient technique which works fine for small data-set as well as huge data-set.
- Treatment Module**:** This module maintained by the Specialist where specialist uploads the treatment details for the patients and patients can view the treatment details.
- Account Module (Admin and member): This is a common module of all actors where they can manage their profile by updating, changing passwords etc.

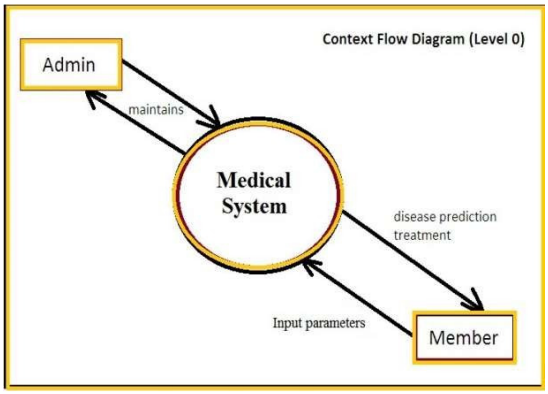

Figure 2: context flow diagram

3. Materials and methods

1. DATASET

This paper is aim to study the future forecasting of COVID-19 spread and focused on the newly affected cases, deaths, and recoveries.

The dataset used to obtain the result and the data set provide by the central for system and science of engineering

TABLE 1: COVID-19 patient death cases time-series worldwide.

Province/State	Country/Region	last	long	1/22/21	1/23/21	3/22/21
Northern territory	Australia	-12.46	130.84	0	0	0
Diamond	Canada	0.000	0.000	0	0	1
Princess Nan	Algeria	28.03	1.65	0	0	19

TABLE 2: COVID-19 new confirmed cases time-series worldwide

Province/ State	Country/Region	Last	long	1/22/21	1/22/21	3/22/21
Northern territory	Australia	33.00	65.00	0	0	74
Diamond	Canada	-37.81	144.96	0	0	411

2. SUPERVISED MACHINE LEARNING MODELS

A supervised learning model is used to prediction and provides with an unknown input .then this learning technique, takes the dataset with input instances.

TABLE 3: COVID-19 recovery cases time-series worldwide

Province/state	Country/Region	last	long	1/22/21	1/23/21	3/22/21
Colombia	Canada	49.28	-123.1	0	0	4
Victoria	Australia	-37.81	144.96	0	0	70
Princess Nan	Algeria	28.03	1.65	0	0	65

Four regression model have used in this COVID-19 future forecasting

- Linear Regression
- LASSO Regression
- Support Vector machine
- Exponential smoothing

4. Evaluation parameters

In this study we evaluate the performance of the each learning models in the terms of R-squared score ($R^{2)}$. Adjusted R-square. Mean square error (MSE) means absolute error. And root mean square error

1) R-squared score

R-square score is a statistical measure used to evaluate the performance of regression models

R-square score ($R^{2)}$ =Variance explained by model

Total variance eq (1)

2) Adjusted R-squared score

This is modified form of $R^{2,}$ which also show the well data points. The primary different between the R^2 and $R^2_{adjusted}$ is the later adjusted number for features in prediction model

Adjusted R-square score ($R^2_{adjusted}$) $=1-(1-R^2)\ \frac{n-1}{n-(k+1)}$ eq(2)

3) Mean absolute error

The mean absolute error is the average magnitude error in the model prediction

$$\mathbf{MAE} = 1/n \sum_{j=1}^{n} [y_j - \hat{y}_j] \qquad \text{eq (3)}$$

4) Mean square error (MSE)

The mean square error is a way to measure the regression models. Its take the distance of data points of regression line and squaring of them.

$$\mathbf{MSE} = 1/n \sum_{i=1}^{n} [y_j - \hat{y}_j]^2 \qquad \text{eq (4)}$$

5) Root mean square error (RMSE)

This root mean square error can be defined as the deviation of prediction errors. This root mean square error is given by

$$\mathbf{RMES} = \sqrt{1/n \sum [y_j - \hat{y}_j]^2} \qquad \text{eq (5)}$$

TABLE 4: Day wise total death cases sample data

Day 1 deaths	Day 2 deaths	. . . . .	Day 66 Death
0	4	. . . . .	20

TABLE 5: Day wise total recoveries rate sample data

Day 1 recoveries	Day 2 recoveries	. . . . . .	Day 66 recoveries
0	6	. . . . .	139

TABLE 6: Day wise total new confirmed cases sample data

Day 1 new	Day 2 new	……	Day 66 new

cases	cases		cases
0	21	. . . . .	749

The forecasting is done by using the machine learning approaches. The dataset is used to give the daily time series summary table. After the initial data is preprocessing step, the data is trained for the 56 days to train the models and testing set.

5. RESULTS AND DISCUSSION

This proposed workflow is for future forecasting of COVID-19 is having the data set and it will be preprocessing by data splitting and data are tested by testing sets. The data are splitting and is training set are then models are used for forecasting and then it is given for the trained models and the evaluation parameters are given the result for the forecasting of future days.

A. DEATH RATE FUTURE FORECASTING

The study performs the death predication on the model of LR, LASSO, SVM, ES forms in the better and equally well achieve in all different models.

TABLE 7: Models performance on future forecasting for death rate

Model	**R^2 score**	**R^2_{adjust}**	**MSE**	**MAE**	**RMSE**
LR	0.96	0.95	840240.11	723.11	916.64
LASSO	0.85	0.81	3244066.79	1430.2	1801.1
SVM	0.53	0.39	16016210.98	3129.7	813.77
ES	0.98	0.97	662228.72	406.08	4002.02

This table shows the performance of the models and also plotted in the form of graphically.

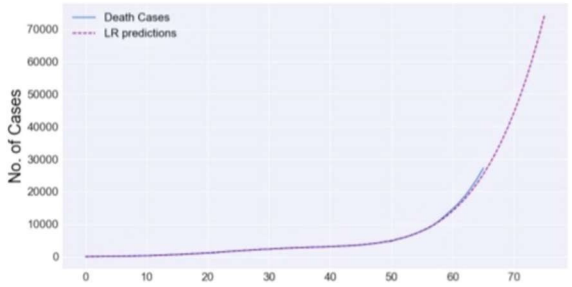

Figure 3: Death prediction by LR for the upcoming 10 days.

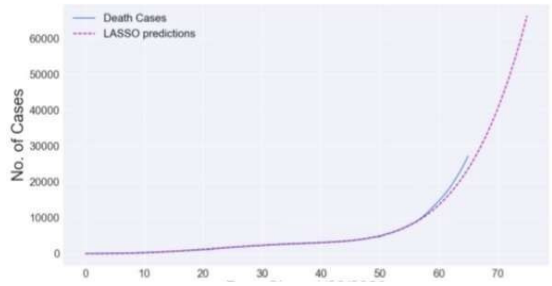

Figure 4: Death prediction by LASSO for the upcoming 10 days.

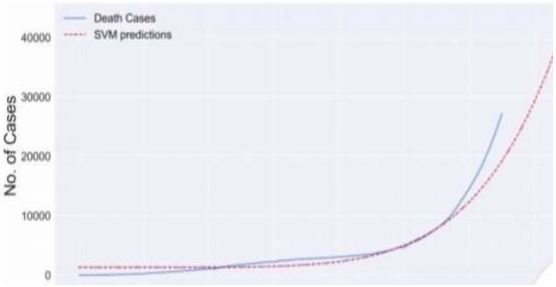

Figure 5: Death prediction by SVM for the upcoming 10 days

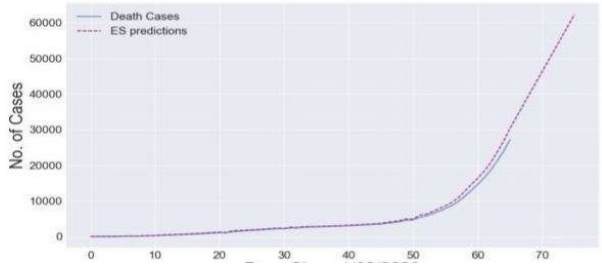

Figure 6: Death prediction by ES for the upcoming 10 days.

B. NEW INFECTED CONFIRM CASES FUTURE FORECASTING

The new confirmed cases of covid-19 are increased day by day.

TABLE 8: Models performance on future forecasting for new infected and confirmed cases

Model	R^2 score	R^2_{adjust}	MSE	MAE	RMSE
LR	0.83	0.79	14729860.99	723.11	38390.51
LASSO	0.98	0.97	3244066.79	1430.2	15322.11
SVM	0.59	0.47	16016210.98	3129.7	813.77
ES	0.98	0.97	662228.72	406.08	4002.02

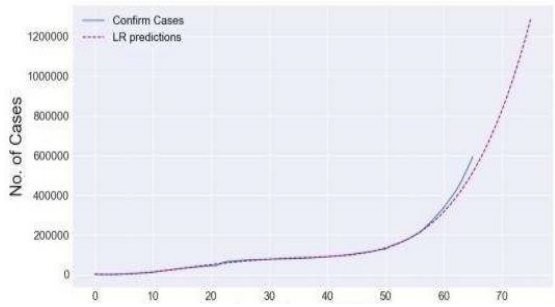

Figure 7: New infected confirm cases prediction by LR for upcoming 10 days

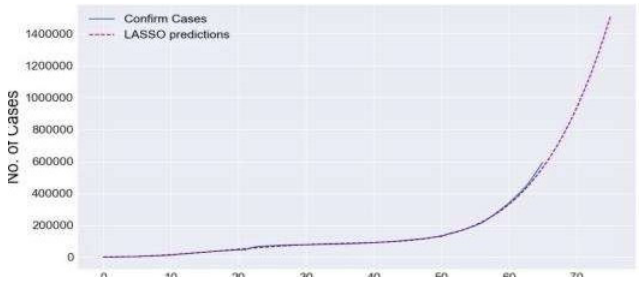

Figure 8: New infected confirm cases prediction by LASSO for the upcoming 10 days.

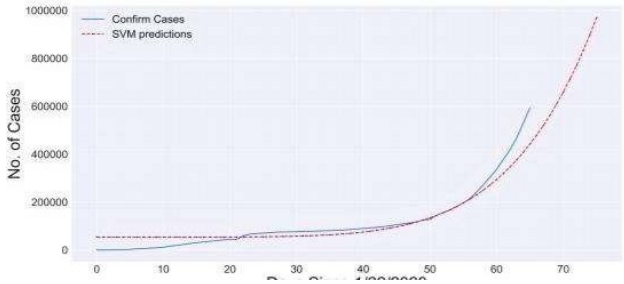

Figure 9: New infected confirm cases prediction by SVM for the upcoming 10 days

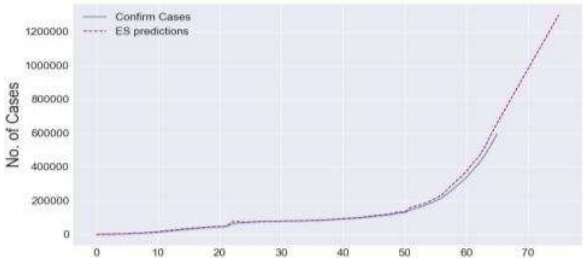

Figure 10: New infected confirm cases prediction by ES for the upcoming 10 days.

C. RECOVERY RATE FUTURE FORECASTING

TABLE 9: Models performance on future forecasting for recovery rate.

Model	R^2 score	R^2_{adjust}	MSE	MAE	RMSE
LR	0.39	0.21	14729860.99	723.11	38390.51
LASSO	0.29	0.08	3244066.79	1430.2	15322.11
SVM	0.24	0.02	16016210.98	3129.7	813.77
ES	0.99	0.99	662228.72	406.08	4002.02

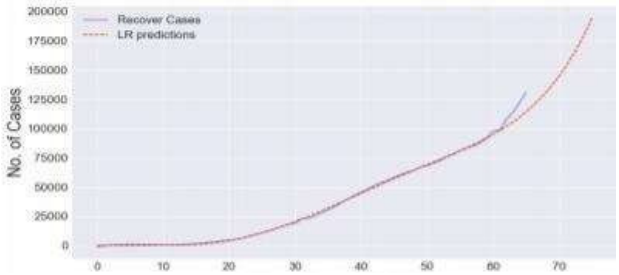

Figure 11: Recovery rate predication by LR for upcoming 10 days

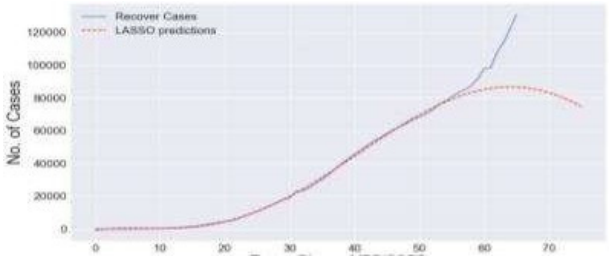

Figure 12: Recovery rate predication by LASSO for upcoming 10 days

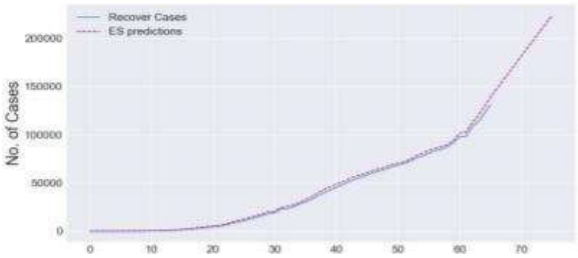

Figure 13: Recovery rate predication by ES for upcoming 10 days

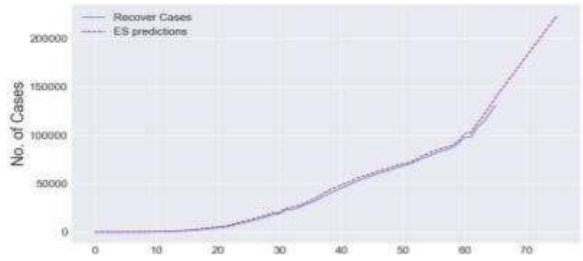

Figure 14: Recovery rate predication by SVM for upcoming 10 days

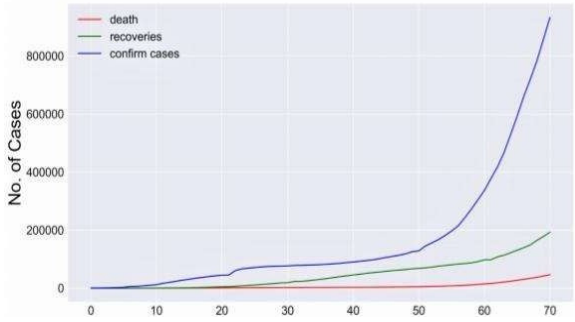

Figure 15: Comparison between death rate and recovery rate and confirm case rate after the 5-days.

6. CONCLUSION AND FUTURE WORK

The prevention of covid-19 pandemic have a massive global crisis. Some the researcher are seriously trying to prevent the covid pandemic, and it is affected the human life. Our government and researchers are cancerously working on it. In this study of future forecasting is made by Machine learning algorithm and proposed to predicting the risk of covid-19.

This system analyses made the collection of data sets and gives the accurate results. Next we plan to explore the predication methodology to fight against covid-19.

REFERENCES

[1] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, ‘‘Statistical and machinelearningforecastingmethods:Concernsandwaysforward,’’PLoS ONE, vol. 13, no. 3, Mar. 2018, Art. no. e0194889.

[2]G.Bontempi,S.B.Taieb,andY.-A.LeBorgne,‘‘Machinelearningstrategiesfortimeseriesforecasting,’’inProc.Eur. Bus.Intell.Summer School. Berlin, Germany: Springer, 2012, pp. 62–77.

[3] F. E. Harrell Jr, K. L. Lee, D. B. Matchar, and T. A. Reichert, ‘‘Regression models for prognostic prediction: Advantages, problems, and suggested solutions,’’ Cancer Treat. Rep., vol. 69, no. 10, pp. 1071–1077, 1985.

[4]P.Lapuerta,S.P.Azen,andL.Labree,‘‘Useofneuralnetworksinpredicting the risk of coronary artery disease,’’ Comput. Biomed. Res., vol. 28, no. 1, pp. 38–52, Feb. 1995.

[5] K. M. Anderson, P. M. Odell, P. W. Wilson, and W. B. Kannel, ‘‘Cardiovascular disease riskprofiles,’’Amer.heartJ., vol.121,no.1,pp. 293–298, 1991.

[6] H. Asri, H. Mousannif, H. A. Moatassime, and T. Noel, ‘‘Using machine learning algorithms for breast cancer risk prediction and diagnosis,’’ Procedia Comput. Sci., vol. 83, pp. 1064–1069, Jan. 2016.

[7] F. Petropoulos and S. Makridakis, ‘‘Forecasting the novel corona virus COVID-19,’’ PLoS ONE, vol. 15, no. 3, Mar. 2020, Art. no. e0231236.

[8] G. Grasselli, A. Pesenti, and M. Cecconi, ‘‘Critical care utilization for the COVID-19outbreakinlombardy,italy:Earlyexperienceandforecastduring an emergency response,’’ JAMA, vol. 323, no. 16, p. 1545, Apr. 2020.

[9] WHO. Naming the Corona virus Disease (Covid-19) and the Virus That Causes it. Accessed:Apr.1,2020. [Online]. Available: https:// www.who.int/emergencies/diseases/novel-coronavirus-2019/technic% alguidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virusthat-cau%ses-it

[10] C. P. E. R. E. Novel, ‘‘The epidemiological characteristics of an outbreak of 2019 novel corona virus diseases (Covid-19) in China,’’ Zhonghua Liu Xing Bing Xue Za Zhi= Zhonghua Liuxingbingxue Zazhi, vol. 41, no. 2, p. 145, 2020.

[11] L. van der Hoek, K. Pyrc, M. F. Jebbink, W. Vermeulen-Oost, R. J. Berkhout, K. C. Wolthers, P. M. Wertheim-van Dillen, J. Kaandorp, J. Spaargaren, and B. Berkhout, ‘‘Identification of a new human Coronavirus,’’ Nature Med., vol. 10, no. 4, pp. 368–373, 2004.