Supervised Learning Inspired Fast Forecasting Model of 2019-nCoV Outbreak using Small Dataset

ArijitChakraborty1, SajalMitra2, Dipankar Das1, Debashis De3,Sankar Prasad Mondal4**\***, Anindya J. Pal5

1Bachelor of Computer Application Department, The Heritage Academy, Kolkata, India.

2Department of Computer Science and Engineering, Heritage Institute of Technology, Kolkata, India.

3Department of Computer Science and Engineering, MaulanaAbulKalam Azad University of Technology, West Bengal, India.

4Department of Applied Science, MaulanaAbulKalam Azad University of Technology, West Bengal, India.

5University of Burdwan, Burdwan, India.

[\*](mailto:*arijit.chakraborty@heritageit.edu)[[sankar.mondal02@gmail.com](mailto:*arijit.chakraborty@heritageit.edu)](mailto:sankar.mondal02@gmail.com)

Abstract:A rapid spread of the 2019-novel Corona Virus (2019-nCoV) epidemic imposes a threat to society and the global economy. The epidemic induced by the contagious coronavirus resulted in the suspension of day to day activities such as education, tourism, and community services in provinces of China and its neighboring countries. The real impact of this virus on a society largely depends on its outbreak momentum. Therefore, it is imperative to formulate a robust and accurate prediction model to approximate its disastrous repercussions on human lives. Limited understanding of the 2019-nCoV outbreak with the imprecision involved induces an extraordinary challenge in framing a prudent forecasting model. This publication elucidates a collaborative framework consisting of Machine Learning (ML) and Statistical prediction methods to estimate the adversity of this virus.The suggested framework offers a high degree of accuracy in evaluating the rise in the 2019-nCoV pandemic in Chinese provinces, with a reasonably small Root Mean Square Error (RMSE) on a small dataset rendered by the World Health Organization (WHO).

1. Introduction

Coronavirus (**CoV**) is a class of microorganisms that produce diverse human ailments such as common cold to more severe conditions like Middle-East Respiratory Syndrome (**MERS-CoV)**and Severe Acute Respiratory Syndrome (**SARS-CoV**).The 2019-nCoV is a new stretch in the CoV family that has not been found in humans until recent times. Diverse phylogenetic experiments concluded that the CoV family keeps an indivisible strand of the RNA genome of varying lengths between 26 to 32 kilobases [1]. The adversity of this virus is not only limited to humans but also identified in several avian hosts [2-3]. Lately, The CoV Study Group of the International Committee for Taxonomy of Viruses suggested three families of Cov, namely, ***α-CoV***, ***β-CoV****,*and ***γ-CoV*** [4]. Most of the coronavirus types that are pathogenic to humans resulted in moderate symptoms except a few notable exceptions, such as SARS-CoV, fundamentally a β-CoV with first evidence reported in Guangdong, southern China in 2002 and **MERS-CoV** [5] first detected in Saudi Arabia in 2012, both of them resulted in massive fatalities. Lately, in December 2019, numerous cases with symptoms of viral pneumonia discovered to be epidemiologically linked with the Wuhan city, the Hubei province of China. To date, several suspected cases of 2019-nCoV infection reported across Chinese provinces. Besides, 2019-nCoV has now been notified in many countries worldwide. The Chinese economy increased at a pace of 6% last year and became a global leader in trade. However, the outbreak of the 2019-nCoV epidemic adversely influenced the Chinese economy with a slower growth pace of 4.5%, in the first financial quarter of 2020. This economic fallout threatens global economic growth with shutdowns of factories and crude oil prices. Therefore, it becomes critical for administrators to estimate human loss and the socio-economic impact caused by 2019-nCoV.

Coronavirus Disease-2019 (**COVID-19**) mainly circulated among persons through using the facial and oral pathways [6]. A significant no. of cases with mild symptoms of COVID-19 not presented to hospitals, thereby limitingus with only a handful of information on this epidemic. Therefore, it becomes incredibly challenging and labour intensive for us to design a faster, accurate, and robust forecast model from the smallest clues for estimating the fatality of the 2019-nCoV outbreak.

In thiscontext, Q. Guo et al. [7] proposed a prediction model to estimate the infectivity pattern of 2019-nCov in Wuhan province of China, using the deep-learning algorithm. [7] observed that 2019-nCoV has a resembling infectivity association with SARS-CoV. By examining the infectivity patterns of all viruses hosted on vertebrates, [7] concluded that the infectivity pattern of mink viruses are strictly proximate to 2019-nCov. In another work, J.M. Read et al. [8] developed an early estimation model based on epidemiological parameters to predict 2019-nCoV in Wuhan, China. By employing the model, [8] estimated that the mean ascertainment rate in Wuhan between 1 and 22 January 2020 was 0.05. However, building a prediction model from a tiny dataset consisting of no. of effected and no. of deaths caused by the 2019-nCoV in Chinese provinces imposes a challenge in using **ML**-based estimation of human causalities as no single model can be considered as the best fit to estimate the adversities caused by 2019-nCoV. In the context of the seriousness of the situation, a new forecasting approach should be examined as an option besides other conventional methods.

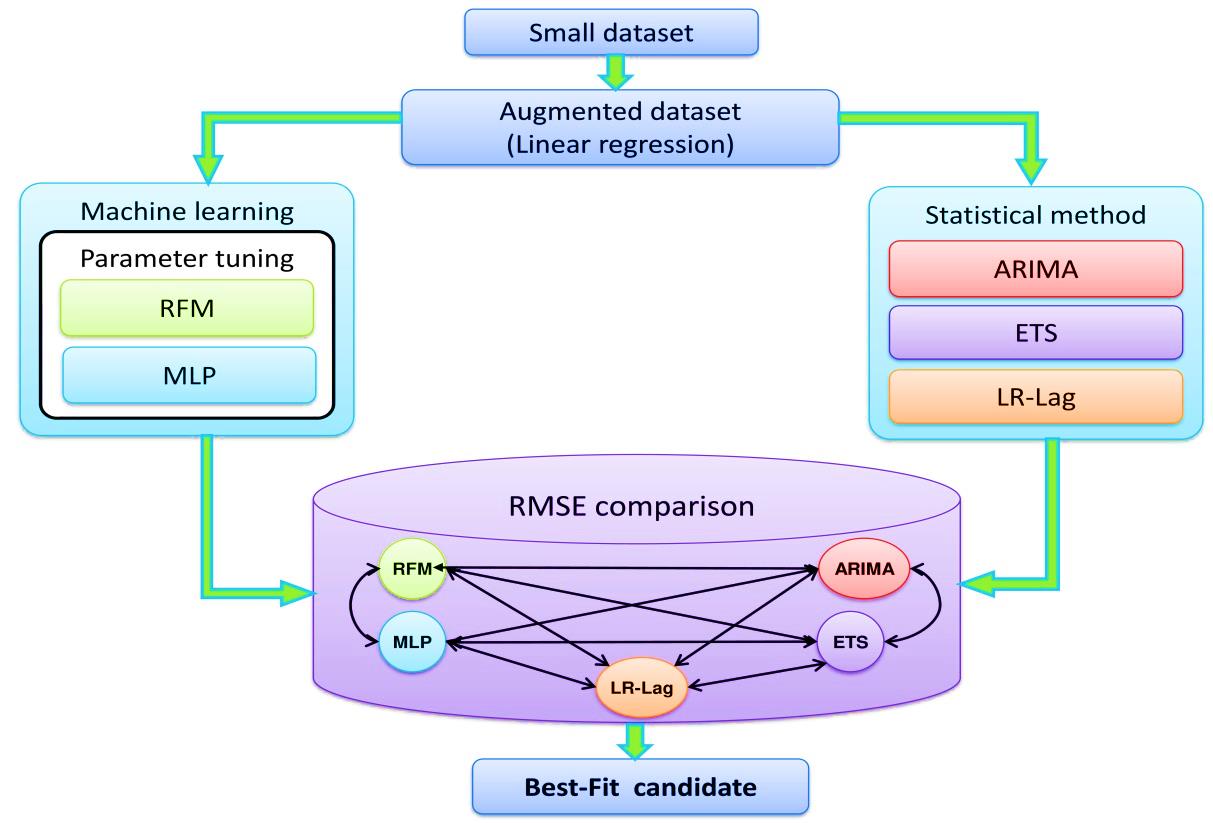
This publication presents a collaborative approach consisting of the merits of statistical and ML-based prediction methods to determine the best-fit candidate in terms of RMSE, to estimate the death toll induced by the 2019-nCov outbreak and therefore named as a Collaborative Framework for Prediction using Small Dataset (**CFPSD**).A layout of our proposed CFPSD model is shown in Fig. 1.

RMSE is a measure of the variations between actual and predicted values offered by a forecasting model, and it represents the degree of error in the forecasting process. Since a small RMSE implies better accuracy of the model, we considered RMSE as a metric to evaluate the performance of the participating units of the proposed collaboration.

The CFPSD approach embraces the following:

1. Augmentation of a small in-scale dataset by using linear regression.
2. The augmented dataset partitioned into a proportion of 70:30 for creating the training set and test set.
3. The machine classifiers, such as Random Forest Model (RFM) [9] and Multi-Layer Perceptron (MLP) [10], employed, and the RMSE values obtained.
4. The parameter settings are sensitive to the performances of [9-10]. Consequently, parametric fine-tuning performed to minimize the RMSE score.
5. The CFPSD model also incorporates the traditional methods such as, The Auto-Regressive Integrated Moving Average **(ARIMA)**, **ETS**, and Linear Regression-Lag (**LR-lag**), and the corresponding RMSE values noted.

Finally, all the RMSE values compared to conclude the performance of the proposed model.



***Fig. 1.****CFPSD model*

1. Methodology

At the initial stage of an international outbreak of the nCoV-2019 epidemic, limited data is available. Therefore, our CFPSD model has to encounter a high degree of contingency. Considering that the virus is novel and needs human expertise for the evaluation of catastrophe induced by nCoV-2019, a hybrid and collaborative forecast model is warranted. Therefore, an ML approach on a small dataset demanded with the following objectives:

1. The forecasting model should be most competing among its peers, i.e., the one with the most moderate RMSE, and highest performance in terms of accuracy.
2. The best forecast candidate must be optimized to maximize the performance by fine-tuning of parameters.
3. The proposed model must be designed to offer reasonable classification accuracy under the constraints such as inadequate availability of data and knowledge.

Our principal research design is based on collaboration between two sets, i.e., a set of machine classifiers, and a set of a traditional forecast approach. Both of these sets accept regression-based augmented data. A detailed description of CFPSD model as follows:

1. **Dataset**

Our analysis is based on the public report given by **WHO** [11], consisting of small observations about no. of people effected and died due to the 2019-nCoV epidemic in China from **21st January 2020** to **14th February 2020**.

1. **Augmenting small dataset using Linear Regression**

**Linear Regression** method used to estimate the linear association between the independent variable and the dependent variable, given in (1).

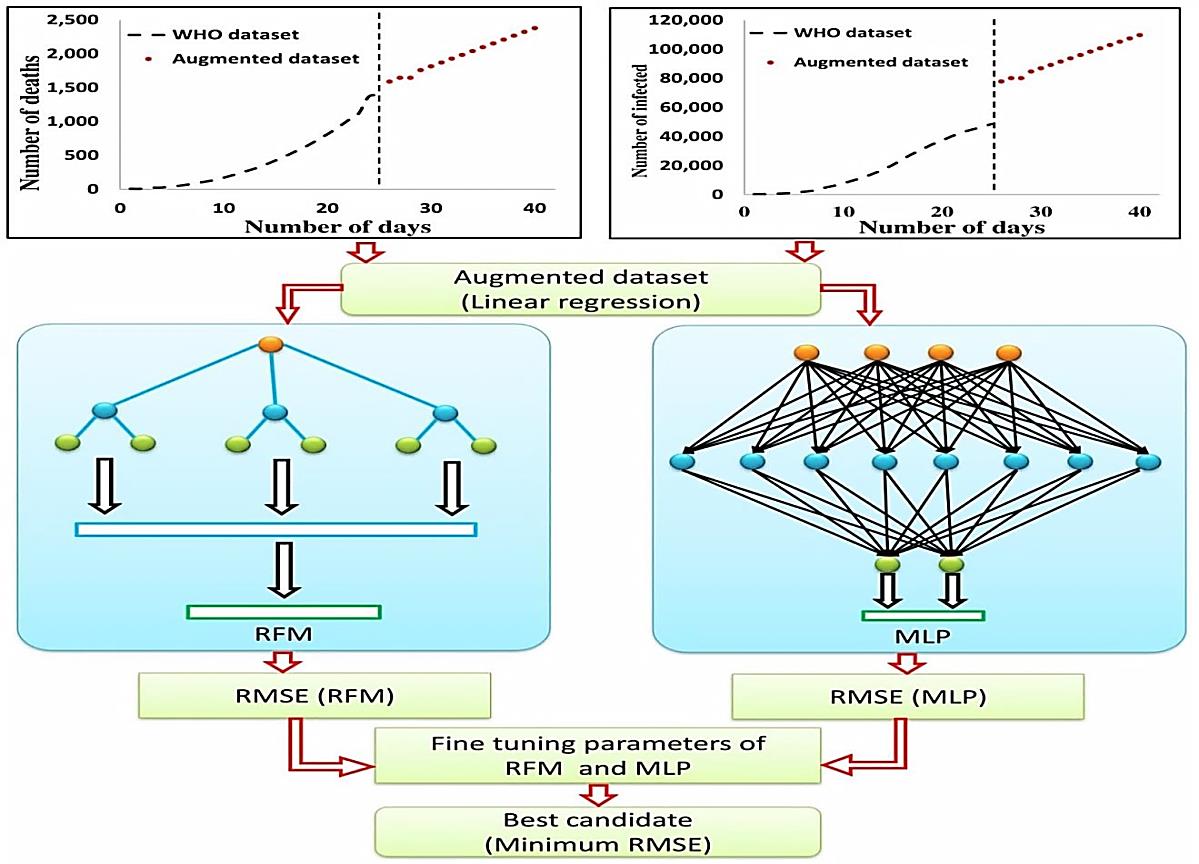
In (1), is the slope of the regression line Y on X, C is the intercept, X is the independent variable, and Y is the dependent variable. The regression coefficient is given in (2).

In (2), N is the no. of observations. We augmented the dataset rendered by [11] of 25 days of the epidemic into a moderate dataset of 40 no. of days, i.e., **up to 29th February 2020**, using the regression line given in (3)-(4).

In (3)-(4), is the independent variable, and, areestimated no. of deaths and effected respectively.

1. **Complex Machine Classifiers**

We employed two machine classifiers [9-10]. [9] is a learning algorithm, where the entire dataset is partitioned into more than one subtree during training and produces a class label for each subtree, whereas [10] mimics the functionality of the human brain through weighted interconnections between artificial neurons. A three-layer MLP consists of the input layer, hidden layer, and output layer, where link weights are adjusted during training, and an appropriate activation function produces class labels, i.e., output layer for a given set of inputs. The regression augmented dataset fed into [9-10], and the corresponding RMSE values obtained, shown in Fig. 2.



***Fig. 2****. Calculating RMSE of RFM and MLP from an augmented dataset*

1. **Fine Tuning of Parameters**

We have used the **WEKA** package [12] to implement [9-10]. The no. of trees, i.e., α is a parameter of the RFM method referred to as no. of iterations in [12]. Similarly, the learning rate,i.e., Ƞ, is the MLP parameter. The default values of these parameters in [12], listed in Table 1.The default parameter of [9-10] does not always produce the best result. Therefore we have fine-tuned the parameters of [9-10] to obtain their highest performances, presented in section 3.

1. **Statistical Prediction Models**

We employed the forecast package of **R** for ARIMA and ETS model development [13], [14], and [15].**ARIMA**[16] is a widely accepted scheme for univariate time series forecasting. The standard notation of ARIMA is ARIMA (p, d, q), where p is the lag order, d is for the degree of differencing, and q is the order of moving average [17]. The ARIMA (p, d, q) model has many combinations. Therefore, identification of the best ARIMA (p, d, q) model for our dataset is a challenging and labor-intensive task. Astepwise approach of ARIMA modelling by referring [18] as follows:

1. Make the data stationary by differencing.
2. Identification of the possible candidate models by carefully observing the Auto Correlation function (**ACF**) and Partial Autocorrelation Function (**PACF**) plots of the data obtained in Step1, i.e., differenced data.
3. Searching for a superior model by using the corrected Akaike Information Criterion (**AICc**).
4. Performing the residual diagnostics to test the excellent fit. If a good fit is obtained, then perform forecast else go to Step2.

This publication incorporates an automated algorithm, namely **auto.arima()**, for the convenience of fast forecasting from forecast package of R to obtain the ARIMA model development and subsequent forecasting using [13], [14] and [15].

The ETS method [18] is a member of the exponential smoothing-state space model family that uses three parameters – **Error**, **Trend**, and **Seasonality**. Each of these three parameters, described as follows:

1. Error = {Additive, Multiplicative}
2. Trend = {None, Additive, Additive-Damped, Multiplicative, Multiplicative-Damped}
3. Seasonality = {None, Additive, Multiplicative}

**Table 1** Parameter table with default value in [12]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Serial no. | Method | Symbol | Description | Default value in WEKA |
|  |  |  |  |  |
| 1 | RFM | α | No. of Iterations/Trees | 100 |
| 2 | MLP | Ƞ | Learning rate | 0.3 |
|  |  |  |  |  |

The ETS scheme, with its three parameters – Error, Trend, and Seasonality, can provide us many combinations. Our primary objective is to select the best-fitted model from the set of ETS combinatorics. Developing individual models of ETS is wearisome. Thus, we have used the **ets()** of the forecast package of R to identify the best ETS candidate model [13], [14] and [15]. The AICc used to select the appropriate model automatically.

The ARIMA and ETS modelling methods have multifaceted applications in the field of medical informatics, wherein A. Zheng et al. [19]used the ARIMA model to forecast the health expenditure such as Total, Government, Social, and Out of pocket expenses in China. In another work, L. Wang et al. [20]employed the ARIMA model to predict the Brucellosis pandemic in Jinzhou, China, whereas, H. Liu et al. [21]used a combination of ETS and SARIMA model for forecasting haemorrhagic conjunctivitis drifts in China. In their paper, Y. J. Tseng and Y. L. Shih [22] used time series models, namely, ETS, ARIMA, and Prophet, to perform the forecasting of influenza scourge in Taiwan. M. Ordu et al. [23] employed ARIMA, ETS, Multiple Linear Regression, and Seasonal and Trend decomposition using Loess Function (**STLF**) methods in generating a complete forecast framework to determine the demand of health care services in hospitals.

We have also employed another statistical forecasting method, namely, LR-lag, to estimate dependent variable Y (Death) based on independent variable X (no. of Days), given in (5).

In (5), Yt is the dependent variable (observed response), d is no. of Days, Yt-1 and Yt-2 are the independent variables, is the random disturbance, α0, β1, and β2 are regression coefficients.

Therefore, Yt-k is the lagged values of the observed response variable Deatht. The final LR-lag is given in (6).

1. ****Results and Discussion****

The Mortality Rate (**MR**) or the Death Ratio (**DR**)due to the nCoV-2019 outbreak is a measure of significant human casualties in a given period, i.e., 21st January - 29th February 2020. The Death Ratio (DR), also known as Fatality Rate (**FR**), is quantified as a fraction of people who died out of effected people in Chinese provinces. Consequently, we developed a feature, namely, mean DR represented as, is the average DRof our considered timeline of 40 no. of days. The DR is given in (7). By using (7), we derived, defined in (8).

In (8), N is the no. of samples.

The augmented samples labeled with two classes, class 1, consist of all samples having DR greater or equal to, and the remaining samples labeled as class 0.The class labeling strategy is givenin (9).

In (9), is the class label of the ith instance, where i∈ [1, 40].

The entire dataset partitioned into a training set and test set in a proportion of 70:30. A subset of 28 no. of samples used to train RFM and MLP classifiers, this is critical as it enables these classifiers to identify a pattern, whereas the remaining 12 no. of samples used to create a test set to perform an impartial performance evaluation of classifiers. By employing the ten-fold cross-validation technique, we noted the RMSE values for default parametric values of [9-10] of [12], listed in Table 2.

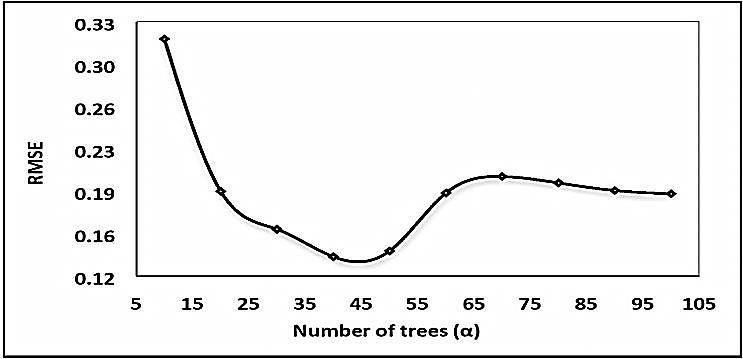
We adjusted the parameter α of the RFM method and noted the RMSE values for each variation of α, and observed that the RMSE of RFM declines till α=40, after that it rises sharply with the increase in α. Consequently, the lowest RMSE = 0.136 for 40 no. of subtrees, i.e., α = 40 achieved listed in Table 3 and shown in Fig. 3. Similarly, we adjusted the ƞ value of the MLP method and observed that the RMSE of MLP rises sharply by increasing the learning rate, i.e., ƞ, listed in Table 4, and shown in Fig. 4.Therefore, it is evident from Table 4 and Fig. 4 that as ƞ increases the divergence between actual casualties due to 2019-nCoV, and the estimated victims also increase. Thus, we considered the lowest RMSE = 0.3066 for ƞ = 0.01 for the MLP method. Additionally, a performance comparison between these two machine classifiers shows that the RFM method outperforms the MLP method, and become the best-fit component of our proposed CFPSD model.

**Table 2** RMSE of RFM and MLP

|  |  |  |
| --- | --- | --- |
| Serial no. | Method | RMSE |
|  |  |  |
| 1 | RFM | 0.1880 |
| 2 | MLP | 0.3274 |
|  |  |  |

**Table 3** Lowest RMSE obtained for α = 40 of RFM

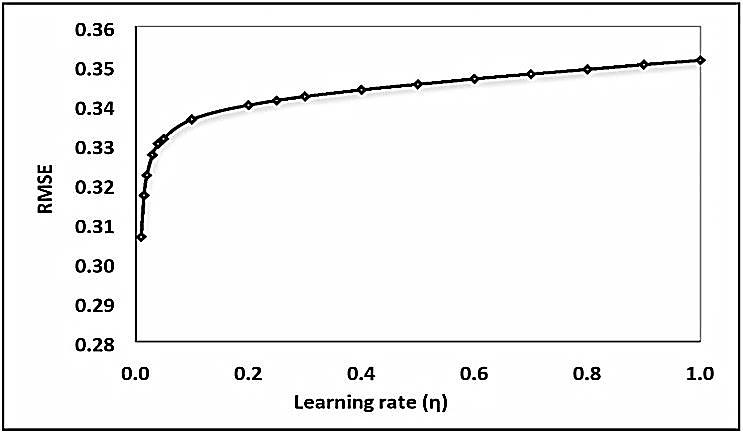
|  |  |  |
| --- | --- | --- |
| Serial no. | No. of trees(α) | RMSE |
|  |  |  |
| 1 | 10 | 0.3160 |
| 2 | 20 | 0.1900 |
| 3 | 30 | 0.1587 |
| ***4*** | ***40*** | ***0.1360*** |
| 5 | 50 | 0.1408 |
| 6 | 60 | 0.1887 |
| 7 | 70 | 0.2023 |
| 8 | 80 | 0.1970 |
| 9 | 90 | 0.1907 |
| 10 | 100 | 0.1880 |
|  |  |  |



***Fig. 3.****Lowest RMSE obtained by the fine-tuning α=40 of RFM*

**Table 4** Lowest RMSE obtained for ƞ =0.01 of MLP

|  |  |  |
| --- | --- | --- |
| Serial no. | Learning Rate(ƞ) | RMSE |
|  |  |  |
| ***1*** | ***0.010*** | ***0.3066*** |
| 2 | 0.015 | 0.3171 |
| 3 | 0.020 | 0.3222 |
| 4 | 0.030 | 0.3274 |
| 5 | 0.040 | 0.3302 |
| 6 | 0.050 | 0.3315 |
| 7 | 0.100 | 0.3364 |
| 8 | 0.200 | 0.3400 |
| 9 | 0.250 | 0.3412 |
| 10 | 0.300 | 0.3422 |
| 11 | 0.400 | 0.3439 |
| 12 | 0.500 | 0.3453 |
| 13 | 0.600 | 0.3467 |
| 14 | 0.700 | 0.3479 |
| 15 | 0.800 | 0.3491 |
| 16 | 0.900 | 0.3503 |
| 17 | 1.000 | 0.3514 |
|  |  |  |



***Fig. 4.****Lowest RMSE obtained by the fine-tuning ƞ =0.01 of MLP*

The predicted no. of deaths obtained using traditional methods such as [16], [18], and LR-lag in China due to the 2019-nCOV outbreak shown in Fig. 5.

RMSE values obtained for [16], [18], and LR-lag in the estimation of death due to the 2019-nCOV outbreak listed in Table 5.The best fit ARIMA and ETS models for the data set as follows:

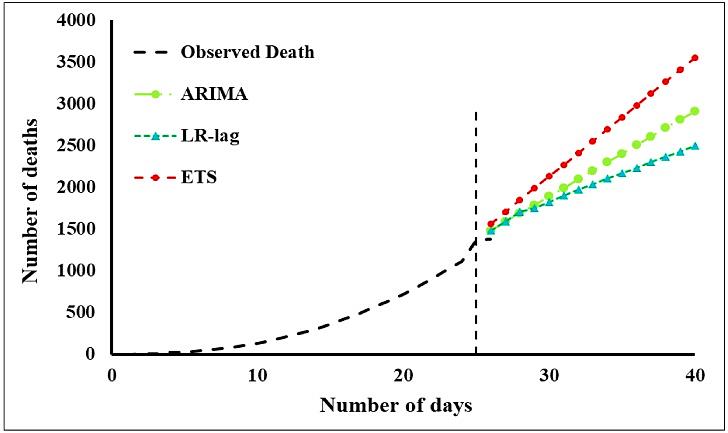
* ARIMA(0, 2, 1),i.e., p = 0, d = 2, and q = 1
* ETS(Additive, Additive, None), i.e., Error = Additive, Trend = Additive, and Seasonality = None

A comparison of RMSE values obtained using ARIMA, ETS, LR-lag, RFM, and MLP is shown in Fig. 6.It is apparent from Fig. 6 that the minimum RMSE of the conventional forecasting method set, i.e., {ARIMA, ETS, LR-lag} is 35.50, and for the ML set, i.e., {RFM, MLP}, the minimum RMSE is 0.14. Therefore, the minimum (35.50, 0.14) = 0.14 is the RMSE achieved for [9] in the proposed CFPSD model.

We investigated the performance of our proposed CFPSD model with two state-of-art time series-forecasting methods, namely, MLP-lag and Exponential smoothing state-space model with Box-Cox transformation, ARMA errors, Trend and Seasonal components (**BATS**) [24].

The MLP-lag is a widely accepted alternative to traditional statistical approaches for addressing the time series forecasting problems because of its ability to learn and generalize accumulated knowledge and model nonlinear trends. In this context, S. Pryima et al. [25]had shown that the MLP with lag inputs yields accurate prediction. In another work, C. ZhiYuan et al.[26] selected the lag values using a differential evolution algorithm to combine with the artificial neural network to achieve prudent forecasting.

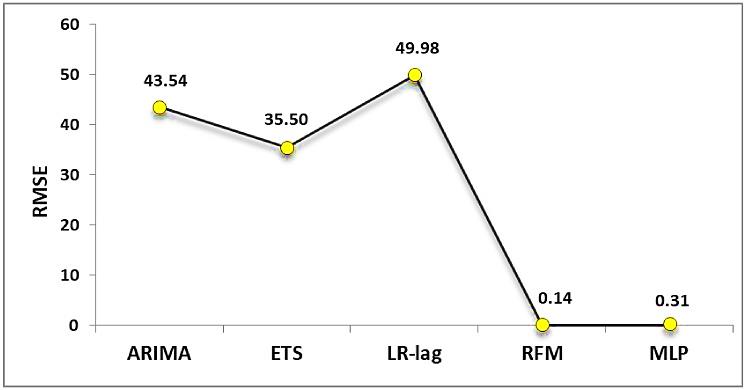
Consequently, we predicted no. of deaths due to the 2019-nCoV epidemic using MLP-lag shown in Fig. 7. The RMSE obtained for the MLP-lag method is 8.14, much higher than the RMSE of RFM. Therefore, the RFM based prediction also outperformed MLP-lag.



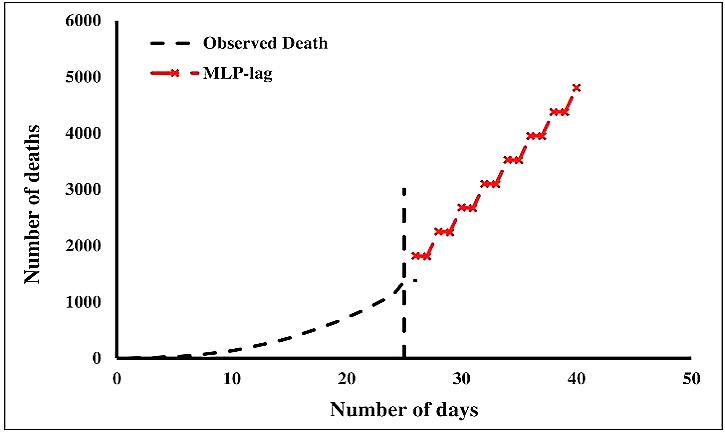
***Fig. 5.****Predicted no. of deaths using ARIMA, ETS, and LR-lag*

**Table 5** RMSE of ARIMA, ETS, and LR-lag

|  |  |  |
| --- | --- | --- |
| Serial no. | Method | RMSE |
|  |  |  |
| 1 | ARIMA | 43.54 |
| 2 | ETS | 35.50 |
| 3 | LR-lag | 49.98 |
|  |  |  |

****

***Fig. 6***.*RMSE of ARIMA, ETS, LR-lag, RFM, and MLP*



***Fig. 7.****Predicted no. of deaths using by MLP-lag*

It is evident from Fig. 5 and Table 5 that all the three models, namely ARIMA, ETS, and LR-lag, exhibited trends upward. The ETS model showed the sharpest rise while the LR-lag has the lowest rising-rate amongst the three. Although the actual data showed a non-linear pattern, evident from the dotted line, the trends are moderately linear. The forecasting results of all the three models suggest that the number of deaths rises sharply with time with a lower bound estimate of approximately 2498 on Day 40 (LR-lag model) and an upper bound estimate of approximately 3553 on day 40 (ETS model).

It can be observed from Fig. 7 that the trend exhibited by the MLP-lag model is upward with a steep rise; however unlike the ARIMA, ETS, and LR-lag models, it portrays a stepwise pattern. The MLP-lag model suggests that the number of deaths reaches 4808 approximately on the 40th day, i.e., 29th February 2020.

The mathematical basis of the BATS model is given in (10).

(10)

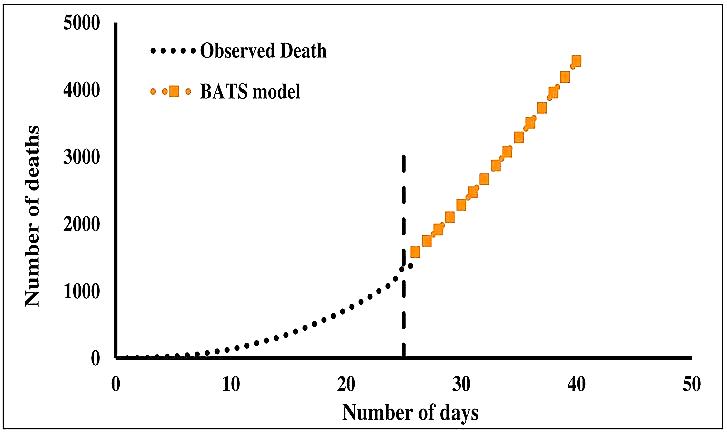
In (10), is the Box-Cox transformed time-series with parameter, denotes level component or local level, denotes growth component, denotes trend damping, denotes the ith seasonal component, M denotes overall seasons, denotes seasonal frequencies, and denotes irregular component or ARIMA (*p, q*) for errors.

The BATS model has varied applications [27], as follows:

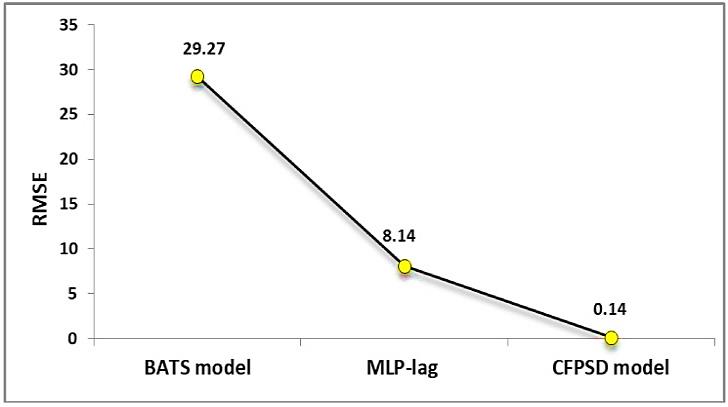
1. It offers an effective means to deal with non-linear data by using the Box-Cox transformation.
2. It can be prudently used to achieve a solution to the autocorrelation problem by introducing the ARMA model on errors.
3. It exhibits better performance than uncomplicated state-space models.

I. Naim et al. [28]used BATS and Trigonometric BATS (**TBATS**) models to forecast daily time series data of BHEL, India. In a similar work, N. Phumchusri and P. Ungtrakul[29]had used BATS, TBATSmodels to forecast daily hotel demands in Phuket, Thailand.We have used the **bats ()** of the forecast package of R software to develop the BATS model [13], [14] and [15]. The BATS model has been developed using the Box-Cox transformation, ARMA errors, and damped trends.We achieved an in-sample RMSE of 29.27 for the BATS model. The death estimates using the BATS model is shown in Fig. 8 where it is evident that the BATS model exhibits a very sharp rise in no. of deaths with time. The trend of the forecast values is approximately linear. The number of death on the 40th day is approximately 4428, lower than the death estimates of the MLP-lag model.We also compared the RMSE achieved by our proposed CFPSD model with the MLP-lag and BATS model, shown in Fig. 9.

A majority of the COVID-19 remains unrecognized and interpreted as common influenza [30]. This gross under-detection of mild cases of COVID-19 inevitably and adversely effects in estimation of actual no. of effected due to the 2019-nCoV epidemic, which may distort the correlation between infected and died, twisting an epidemiologic actuality. Therefore, determination of infection status and confirm cases through a positive Polymerase Chain Reaction(**PCR**) test [31] required in the initial phase of the epidemic. However, the dataset of [11] reports a day-wise summary about the no. of affected and no. of deaths only, without incorporating the underreporting of mild cases of COVID-2019, which may render a gap between prediction and actual death trolls in China. Despite considering a minimal public dataset of [11], we estimated 2380 no. of deaths by 29th February 2020 in China with a reasonably small RMSE of 0.136, using the proposed CFPSD model.



***Fig. 8***.*BATS model for death estimation*



***Fig. 9***.*RMSE of CFPSD, MLP-lag, and BATS model*

Recently, S.J. Fong et al. [32] developed an early forecasting model using a small experimental dataset of the 2019-nCoV epidemic collected from the archive of Chinese health authorities between 21st Jan – 3rd Feb 2020, i.e., a total 14 no. of observations. The proposed model, named Group of Optimized and Multi-source Selection (**GROOMS**) consisting of ARIMA, Exponential, Holt-Winters Addictive and, Holt-Winters Multiplicative methods for conventional prediction, Linear regression, Support Vector Machine with Regression (**SVR**) for regression-based estimates, Fast decision tree learner, MP5 decision tree leaner methods for ML predictions, and Polynomial Neural Network (**PNN**), Polynomial Neural Network with correction (**PNN+cf**) for PNN-forecasting. TheGROOMS is designed to converge a collection of optimized forecasting models. In GROOMS, [32] augmented an existing small dataset to a relatively large set for 6 days advance forecast. This extended dataset was then applied to the panel selection mechanism for determining the best-fit forecast candidate among several models that participated in GROOMS. There is a total of five no. of groups of data analytics in GROOMS used to develop the proposed forecasting model. Authors of [32] observed that the *PNN+cf* methods offer minimal error, i.e., a relatively low RMSE, and become the best-fit candidate to forecast the 2019-nCoV outbreak in China.In conventional forecasting, [32] achieved the lowest RMSE of 695.98 for the Holt-Winters Addictive model. However, for the ML-based forecasting, such as the SVR method, a minimum RMSE of 228.39 achieved, and lastly, for in the PNN+cf model produced a comparatively smaller RMSE of 136.55. Therefore, the PNN+cf produced the lowest RMSE of all methods, i.e., min. {695.98, 228.39, 136.55} = 136.55 and becomes the best-fit forecasting candidate.

The group collaboration approach of estimation has varied applications and not limited in predicting epidemics only. In this context, S. Janani et al. [33] predicted heart disease by using the Highly Co-Related Practical Swarm Optimization (HCR-PSO) feature. A total of 583 no. of heart patient data samples consisting of 10 no. of independent and one dependent variable collected from the UCI repository, to analyze and determine an active heart disease prediction method using an array of ML algorithms, namely KNN (K-Nearest Neighbor), RFM, Support Vector Machine (SVM), Bayesian network and MLP. Authors of [33] predicted heart disease at an early stage with a satisfactory outcome in terms of classification RMSE and accuracy. In [33], the Bayesian network found to produce the lowest RSME of 0.40 and reasonably high accuracy of 90.33. Therefore, the Bayesian network outperformed the other classification algorithms in the proposed prediction model. To evaluate the performance of our CFPSD model, we compared the RMSE achieved by us with [32-33] and observed that our CFPSD model performed satisfactorily in terms of RMSE, listed in Table 6.

**Table 6**Comparison of RMSE with [32-33]

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Authors | Model used | Best-Fit Candidate | Dataset size | Forecast | RMSE |
|  |  |  |  |  |  |
| S.J. Fong et al. | GROOMS | PNN+cf | Small | 2019-nCoV Outbreak | 136.55 |
| S. Janani et al. | ML | Bayesian network | Medium | Heart disease | 0.40 |
| Ourmodel | CFPSD | RFM | Small | 2019-nCoV Outbreak | 0.14 |
|  |  |  |  |  |  |

1. ****Conclusion****

A mutational impact on the protein molecule alters the biophysical characteristic, such as Protein-Protein Interactions (PPI)-thermodynamics. Therefore, assessment of any such alteration due to mutation between a pair of interacting proteins can usually be accomplished using faster computational methods such as Machine Learning. The widespread use of machine classifiers in handling complicated biological information as well as in the energetic or architectural features on proteins analysis offers sound forecast estimates in a reasonable time-frame. Lately, the pandemic outbreak of a 2019-nCoV poses a threat of international interest resulted in a shutdown of public services. The worldwide research highlighted some critical aspects of this epidemic, such as a structural property of 2019-nCoV, which binds to ACE2 with a higher affinity than SARS-CoV. The exponential increase in a 2019-nCoV tantrum, leading to the deaths of thousands, warrants the use of a new estimation mechanism that can be efficiently used on a small imprecise dataset. In this context, the proposed CFPSD model can facilitate an early evaluation of 2019-nCoV adversity, as it offers a relatively small RMSE of 0.136. The proposed model is useful for a scarce dataset as it combines the merits of ML-predictions and statistical predictions, thereby offers a time-critical estimation mechanism for administrative counteracting induced by this unfortunate infectious outbreak.

The outcome of the proposed collaboration is very encouraging. Notwithstanding, with the result, we compared the performance of the CFPSD model with MLP-lag and BATS method, where it performed admirably. We also examined some similar collaborative strategies used on medium and small datasets against our model and observed that despite using small scale data, our proposed model performed satisfactorily. The proposed work incorporated a three-layer approach consisting of augmentation, a mixture of conventional and ML-based forecasting methods, and performance validation with two widely used time-series forecasting techniques. However, the deviation in predictions offered by different estimation methods used in this publication warrants an in-depth examination further. The RMSE outcomes achieved using the methods participated in CFPSD may help in probing new algorithms breeds that fit well in the small scale dataset. The minimum RMSE of the participating members of CFPSD remains the sole consideration to evaluate its performance. Though, the 2019-nCoV pandemic escalates or plunges with time are complicated and depends on multiple socio-economic factors. The use of the WHO dataset in estimating human casualties is essential. Still, CFPSD can be more productive by considering non-technical components that can aid authorities in making reliable decisions.

1. References
2. Su, S., Wong, G., Shi, W., Liu, J., Lai, A.C.K., Zhou, J., Liu, W., Bi, Y., Gao, G.F.: ‘Epidemiology, genetic recombination, and pathogenesis of coronaviruses’, Trends in Microbiology, 2016, 24, (6), pp. 490–502
3. Cavanagh, D.: ‘Coronavirus avian infectious bronchitis virus’, Veterinary Research, 2007, 38, (2), pp. 281–97
4. Ismail, M.M., Tang, A.Y.,Saif, Y.M.: ‘Pathogenicity of turkey coronavirus in turkeys and chickens’, Avian Diseases, 2003, 47, (3), pp. 515–522
5. Woo, P.C.Y., Huang, Y., Lau, S.K.P., Yuen, K.Y.: ‘Coronavirus Genomics and Bioinformatics Analysis’, Viruses, 2010, 2, (8), pp. 1804–1820
6. Zaki, A.M., Boheemen, S.V., Bestebroer, T.M., Osterhaus, A.D.,Fouchier, R.A.: ‘Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia’, N. Engl. J. Med., 2012, 367, (19), pp. 1814–1820
7. Gale, J.: ‘Coronavirus May Transmit Along Fecal-Oral Route’, Xinhua Reports, February 2020
8. Guo, Q., Li, M., Wang, C., Wang, P., Fang, Z., Tan, J., Wu, S., Xiao, Y., Zhu1, H.: ‘Host and infectivity prediction of Wuhan 2019 novel coronavirus using deep learning algorithm’, bioRxiv2020.01.21.914044, 2020
9. Read, J.M., Bridgen, J.R.E., Cummings, D.A.T., Ho, A., Jewell, C.P.: ‘Novel coronavirus 2019-nCoV: early estimation of epidemiological parameters and epidemic predictions’, medRxiv2020.01.23.20018549, 2020
10. Ho, T.K.: ‘The random subspace method for constructing decision forests’, IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20, (8), pp. 832-844
11. Rosenblatt, F.: ‘Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms’, Spartan Books, Washington DC, 1961
12. ‘WHO | Novel Coronavirus – China’, Situation report archived from WHO, https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports, accessed 14th February 2020
13. Smith, A., Tony, C.: ‘Introducing machine learning concepts with WEKA’, Statistical Genomics Methods in Molecular Biology, Humana Press, New York, 2016, 1418, pp. 353-378
14. Hyndman, R.J., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E.,Yasmeen, F.: ‘forecast: Forecasting functions for time series and linear models’, R package version 8.4, http://pkg.robjhyndman.com/forecast, 2018
15. Hyndman, R.J., Khandakar, Y.: ‘Automatic time series forecasting: the forecast package for R’, Journal of statistical software, 2008, 26, (3), pp. 1-22
16. R Core Team, ‘R: A language and environment for statistical computing’, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org, 2019
17. Box, G.E.P.,Jenkins, G.M.: ‘Time series analysis: Forecasting and control’, Holden-Day, San Francisco, 1970
18. Zhang, M.: ‘Time Series: Autoregressive Models AR, MA, ARMA, ARIMA’, University of Pittsburgh, October 2018
19. Hyndman, R.J., andAthanasopoulos, G.: ‘Forecasting: principles and practice’(OTexts: Melbourne, Australia, 3rd edn. 2019)
20. Zheng, A., Fang, Q., Zhu, Y., Jiang, C., Jin, F., Wang, X.: ‘An application of ARIMA model for predicting total health expenditure in China from 1978-2022’, Journal of Global Health, 2020, 10, (1), pp. 1-8
21. Wang, L., Liang, C., Wu, W., Wu, S., Yang, J., Lu, X., Jin, C., Cuihong, J.: ‘Epidemic Situation of Brucellosis in Jinzhou City of China and Prediction Using the ARIMA Model’, Canadian Journal of Infectious Diseases and Medical Microbiology, 2019, pp. 1-9
22. Liu, H., Li, C., Shao, Y., Zhang, X., Zhai, Z., Wang, X., Qi, X., Wang, J., Hao, Y., Wu, Q., Jiao, M.: ‘Forecast of the trend in incidence of acute hemorrhagic conjunctivitis in China from 2011–2019 using the Seasonal Autoregressive Integrated Moving Average (SARIMA) and Exponential Smoothing (ETS) models’, Journal of Infection and Public Health, 2020, 13, (2), pp. 287-294
23. Tseng, Y.J., Shih, Y.L.: ‘Developing epidemic forecasting models to assist disease surveillance for influenza with electronic health records’, International Journal of Computers and Applications, 2019, pp. 1-6
24. Ordu, M., Demir, E., Tofallis, C.: ‘A comprehensive modelling framework to forecast the demand for all hospital services’, The International journal of health planning and management, 2019, 34, (2), pp. e1257-e1271
25. ‘BATS Model – R Documentation’, https://www.rdocumentation.org/packages/forecast/versions/8.11/topics/bats, accessed 19 March 2020
26. Pryima, S., Vovk, R., Vovk, V.: ‘Using Artificial Neural Networks to Forecast Stock Market Indices’, 2019 XIth International Scientific and Practical Conference on Electronics and Information Technologies (ELIT), Lviv, Ukraine, September 2019, pp. 108-112
27. ZhiYuan, C., Khoa, L.D.V., Boon, L.S.: ‘A Hybrid Model of Differential Evolution with Neural Network on Lag Time Selection for Agricultural Price Time Series Forecasting’, Advances in International Visual Informatics Conference, Springer, Cham, November 2017, pp. 155-167
28. ‘BATS and TBATS Model’, https://yintingchou.com/posts/bats-and-tbats-model, accessed 19 March 2020
29. Naim, I., Mahara, T., Idrisi, A.: ‘Effective short-term forecasting for daily time series with complex seasonal patterns’, Procedia computer science, 2018, 132, pp. 1832-1841
30. Phumchusri, N., Ungtrakul, P.: ‘Hotel daily demand forecasting for high-frequency and complex seasonality data: a case study in Thailand’, Journal of Revenue and Pricing Management, 2020, 19, (1), pp. 8-25
31. Battegay, M., Kuehl, R., Tschudin-Sutter, S., Hirsch, H.H., Widmer, A.F., Neher, R.A.: ‘2019-novel coronavirus (2019-nCoV): estimating the case fatality rate – a word of caution’, Swiss Medical Weekly, February 2020
32. Johns Hopkins Center for Health Security, ‘Daily updates on the emerging novel coronavirus from the Johns Hopkins Center for Health Security’, 9th February 2020
33. Fong, S.J., Li, G., Dey, N., Crespo, R.G., Herrera-Viedma, E.: ‘Finding an Accurate Early Forecasting Model from Small Dataset: A Case of 2019-nCoV Novel Coronavirus Outbreak’, International Journal of Interactive Multimedia and Artificial Intelligence, 2020
34. Janani, S., Tamilselvi, R.: ’HCR-PSO Feature Selection for Heart Disease Prediction’, International Journal of Scientific Research and Reviews, 2018, 7, (4), pp. 1857-1864