Forecasting COVID-19 Outbreak Through Fusion of Internet Search, Social Media, and Air Quality Data: A Retrospective Study in Indian Context

Sankhadeep Chatterjee, Kushankur Ghosh, Arghasree Banerjee, and Soumen Banerjee, *Senior Member, IEEE*

***Abstract*— This article proposes a machine learning augmented technique to predict the coronavirus disease (COVID-19) outbreak in India by combining Internet search trends along with social media data retrieved from Twitter. A comprehensive list of suitable search words has been used to select a large collection of Tweets, and the Internet search trends of the same keywords have been fetched. First, a lag correlation analysis is conducted to find the number of days, ahead of the current time, required to make an accurate prediction of COVID-19 cases. Second, both shallow and deep learning methods are engaged to predict the number of COVID-19 cases in a specific geospatial location in India. Thereafter, statewise air pollution data collected from the Central Pollution Control Board, Government of India, are amalgamated to understand the effect of air pollution in spreading of COVID-19 disease. The air pollution monitoring parameters have been combined to understand their effects in the prediction of COVID-19 cases in the Indian context. Experimental results reveal that accurate predictions can be made 85 days ahead of the current time using the proposed method ($r > 0.85$), thereby establishing its ingenuity in the prediction of COVID-19 spread in advance.**

***Index Terms*— Air quality, coronavirus disease (COVID-19), Internet search, SARS-CoV-2, social media.**

I. Introduction

CORONAVIRUS disease (COVID-19), the current ongoing pandemic caused by the SARS-CoV-2 virus, has been declared as a Public Health Emergency of International Concern by the World Health Organization on January 2020 [1]. From the first detected case of COVID-19 in Wuhan, China [2], the virus has spread its wings rapidly and engulfed the entire world with no regions untouched. In India, the first case of COVID-19 was reported in the state of Kerala on January 2020 [3]. With its initial rate of infection being as low as 1.7, its outspreading seems to be significantly lower in comparison to estimates made earlier [4]. Despite that preventive measures have been taken by authorities, the infection rate rose sharply due to a lack of self-precautions among a majority part of the Indian population [5]. Currently, the number of active cases has reached 656 026 as of October 25, 2020, with cases being reported from all states and union territories except Lakshadweep [6]. The mortality rate, considered to be one of the lowest in the world, raises serious concern over the nonavailability of an accurate death reporting system in India, thereby leading to miscommunication of actual mortality rate, which could be higher than the official proclamation [5], [7]. In a quick response to immediately curb the menace, the Government responded by imposing several lockdowns in phases in order to prevent the rapid spread of COVID-19. All nonessential services, including business, supply chain, educational institutes, and interstate transportations, remain closed during this lockdown, which extended for the longest duration in the world to date. However, in the absence of a well-planned lockdown, the true purpose of despreading of the disease could not be materialized [7]; rather it slowed down the economic activities of the nation. The inadvertent effect of this led to a mass community mobilization, thereby triggering a high risk of coronavirus spread across the country [8]. Moreover, it has been found that a major number of COVID-19 cases were reported from certain states, such as Maharashtra, Tamil Nadu, Delhi, Gujarat, and West Bengal. Thus, a countrywide lockdown seemed to be too harsh for the economy. Subsequently, after May 2020, the lockdowns were revoked gradually keeping the "containment" zones under restriction. The uneven distribution of COVID-19 cases in India added with poor healthcare infrastructure and improper management of contact tracing have made the isolation of containment zones nearly impossible [9]. Hence, accurate prediction of local COVID-19 outbreaks becomes essential to combat the spread of this contagious disease.

Initial attempts to predict COVID-19 cases have already been reported earlier [10]–[13]. LSTM-based models are trained using daily COVID-19 case data of the USA and India to predict the number of new cases in advance [14]. As an alternative to such models, primarily useful in capturing an overall picture of the entire country, Internet search trends, social networks, and so on have emerged as a strong source of near-real-time data, capable of reflecting community disease spread in a specific geographical location [15]. During the

Manuscript received March 31, 2021; revised August 8, 2021 and November 1, 2021; accepted December 26, 2021. *(Corresponding author: Sankhadeep Chatterjee.)*

Sankhadeep Chatterjee is with the Department of Computer Science and Technology, University of Engineering & Management, Kolkata 700160, India (e-mail: chatterjeesankhadeep.cu@gmail.com).

Kushankur Ghosh and Arghasree Banerjee are with the Department of Computer Science and Engineering, University of Engineering & Management, Kolkata 700160, India (e-mail: kush1999.kg@gmail.com; banerjeearghasree@gmail.com).

Soumen Banerjee is with the Department of Electronics and Communication Engineering, University of Engineering & Management, Kolkata 711103, India (e-mail: prof.sbanerjee@gmail.com).

Digital Object Identifier 10.1109/TCSS.2022.3140320

lockdown, social media platforms (SMPs) have become a medium for disseminating scientific information and news related to the pandemic, a mode of communication to those who were less accustomed to the use of such platforms. Among several SMPs, Twitter has been used extensively for this purpose across the world [16]. In addition to SMPs, search trends of web browsers have been found useful in revealing meaningful insights about the COVID-19 pandemic [17]. Specific keyword search trends in Google have emerged as an indicator of the spread of the SARS-CoV-2 virus in several countries. In [18], a study conducted using the Google trends of search words "wash hands" and "face masks" revealed a lower speed of spread with an increasing number of searches. A high correlation has been reported for search words "COVID," "COVID pneumonia," and "COVID heart" against the daily number of cases in the United States for a delay of 12–14 days [19]. Moreover, Google trends have been used in Europe [20], Iran [21], Taiwan [22], and Columbia [23] to correlate the search trends with the number of COVID-19 cases. Combined search trends of multiple search engines or platforms have been used to establish a correlation with the number of COVID-19 cases in China [17] with a lag of six days. Another study classified social media posts of the Weibo platform into different categories, and it has been revealed that, with the category of "sick posts," a spike in the number of COVID-19 cases can be predicted 14 days ahead of the current time [24]. A massive surge in social media use has been observed in India, even before the COVID-19 pandemic, with more than 50% of the entire population using some type of SMP for various purposes starting from forming communities to advertisement and business. In order to comply with Government directives in maintaining social confinement in preventing the spread of COVID-19, social interactions among family, friends, and coworkers have reduced dramatically. A good number of social media users have started using SMPs as an alternative to physical communication, thereby increasing the amount of time spent on social media to a greater extent. In addition, various posts, sharing news, and information regarding COVID-19 have significantly increased during the lockdown, thus creating psychological effects among the users [25]. Yet, predictions made solely on the basis of social media data might not appropriately reflect the entire scenario, arising mostly out of imperfections in predicting pandemic-related parameters made from earlier attempts [24]. Also, searching patterns of popular Internet search engines might reveal valuable insights about pandemic [26].

Apart from data acquired through digital platforms, air quality-related information has also been studied extensively in the context of the pandemic. The focus of most of these studies is to understand the variation in the quality and correlate it with the lockdown situation in a country. The complete shutdown has resulted in prominent variation in air quality index in various regions of countries, such as the United States of America [27], Bangladesh [28], Ecuador [29], Italy [30], and India [31]. In [30], a high correlation between air quality and frequency of COVID-19 cases in 71 regions in Italy is reported. The same has been reported in India [31], where the authors established statistically the sustainable improvement in air quality as an impact of nationwide lockdown. This motivated the authors to explore further and identify the constructiveness of the established works in building machine learning and deep learning models. In this study, social media data along with Internet search engine trends of popular search engines and air quality data have been considered for predicting daily cases of COVID-19 in advance. This study took into account a large and comprehensive collection of Tweets relevant to the COVID-19 pandemic in India. Geotagged Tweets have been picked up through the use of a carefully selected pool of keywords highly associated with the COVID-19 pandemic for various states and union territories of India. The air quality data within the range of nationwide lockdown in India is extracted from real-time data published by the Indian Government [32]. This work demonstrates a robust and reliable method to predict COVID-19 cases in India without relying upon traditional theoretical frameworks.

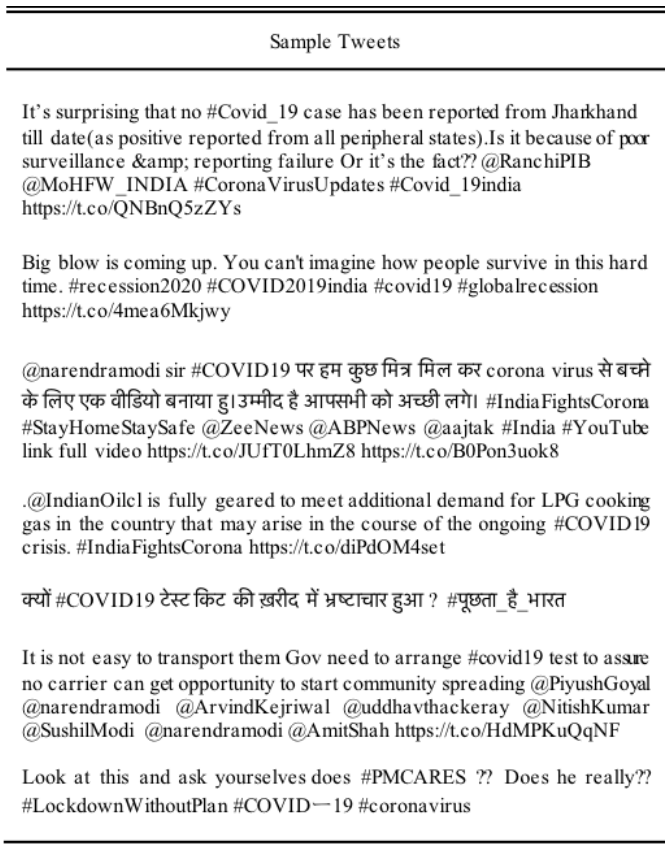

Fig. 1. Some typical COVID-19 tweets originating from India.

The rest of this article is arranged as follows. Section II reports the methodologies that are deployed to collect and preprocess data, as well as the methods to perform lag correlation analysis. In addition, the prediction models used to predict daily COVID-19 cases are discussed as well. Finally, Section III reports experimental results. This section is split into two primary subsections, namely, the first one in which the results are reported for data collected during the lockdown phase of the first wave of COVID-19 outbreak in India and the second one that reports the results for data of the entire pandemic period of the year 2020. In addition, a special study is also conducted that encompasses the second wave

of COVID-19 outbreak in the country and also predicts the number of possible new cases in the near future.

II. Methodology

The experimental methodology used by the authors to analyze and formulate a web-data-based prediction model of new COVID-19 cases consists of four primary segments. To analyze the natural deviation in public reaction with the daily increase in the number of newly identified cases in India, the authors have considered the count of daily social media posts and web searches to analyze the effects of the pandemic [33]. The web data used for the experiment are collected from both sources. The motivation for using social media data is obtained from recent studies that have successfully correlated search trends with an increase in the number of new cases. Few examples of Indian tweets are presented in Fig. 1. Fig. 2 describes the workflow diagram of the proposed methodology. The experiment starts with an initial step of collecting and preprocessing the data from multiple unlinked sources. The data are then segregated accordingly to be competent with the required sets of experiments. These experiments are designed to comprehensively express the elaborated study conducted in 2020. This step is then followed by two separate sets of statistical studies to justify the effectiveness of the proposed idea. The study is associated with a regional analysis to identify the socially active regions during the nationwide lockdown in India. The concluding layer of the experiment applies deep learning and machine learning algorithms to predict the count of future cases by utilizing the extracted diurnal data. The final layer includes a special prediction analysis by using air quality data during the lockdown. It is evident that the nationwide lockdown in the country resulted in less traffic during the earlier months of 2020. The air quality data-based prediction is executed to understand its intensity to determine the number of new cases. Herein, the investigation is further enhanced by documenting the impact of the second wave of COVID-19 in India through a supplemental analysis conducted via social media posts and web searches.

A. Data Collection, Preprocessing, and Segregation

The study carried out by the authors encapsulates the application of various factors that affected the new COVID-19 cases during the nationwide lockdown and afterward. The first phase of nationwide lockdown was imposed on March 25, 2020. Almost all nonessential services were suspended, and the strict ban was imposed on all modes of transportation, offices, educational institutes, shopping malls, marketplaces, and so on. The second phase of lockdown started on April 15, 2020, and allowed restricted activities based on the severity of the spread of the pandemic. This was implemented by the conversion of classified districts into containment zones. In the third phase of lockdown, imposed from May 4, 2020, the entire country was split into regions namely "Red," "Orange," and "Green" depending on the severity of the new cases of COVID-19. Few relaxations in daily activities were allowed in the "Green" and "Orange" regions. In the final phase of nationwide lockdown that started on May 18, 2020, and lasted till

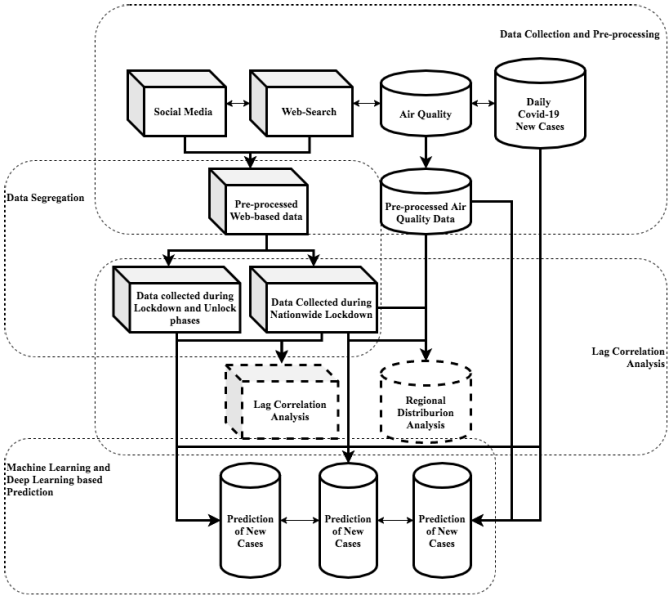

Fig. 2. Proposed methodology.

May 31, 2020, more authority was given to local government bodies to classify regions into "Buffer" and "Containment" zones. A statewise analysis of the COVID-19 pandemic has been depicted in Fig. 5. The authors observed and collected the daily frequency of posts related to the virus outbreak from Twitter [34]–[38]. Recent additions to the literature have effectively used Twitter as a source to assess the public sentiment and opinion progression during the COVID-19 pandemic [39]–[41]. The diurnal posts from January 22 to December 9, 2020, are evaluated by using the Twitter API by adjusting the country name and choosing the hashtags appropriate to the Indian context. The significance of this time period lies in the extensive coverage of the entire lockdown and unlock periods in the country. To conduct the special study on the second COVID-19 wave in India, tweets were collected between February 1 to July 2, 2021 [42]. The collection of tweets is mainly by avoiding retweets and is mostly based on various trending hashtags [43], such as #covid19, #covid_19, #IndiaFightsCorona, #coronavirus, #CoronaVirusUpdate, #COVID-19, #CoronavirusOutbreak, #workfromhome, #lockdown, #stayhomestaysafe, #socialdistancing, and other keywords, and phrases, such as "wash your hands," "lockdown," "coronavirus," "wear a mask," "vaccine," "self-isolating," and "quarantine."

In addition to social media, the authors analyzed and documented web-search trends of search terms, such as "coronavirus," " COVID," and "pneumonia," from Google Trends [44] within the same timelines. To ensure the significance of the air quality data during the four phases of lockdown in India, the authors have collected the quotidian data up to the fourth lockdown in India from the Central Pollution Control Board, Government of India [32]. Each of the extracted data is used with the intention to build a multicontext dataset to support the proposed intelligent models. In the next step, the presence of irregular instances is removed, and missing data are replaced. The extracted web data are then

divided into two parts according to the lockdown timelines to support the two segments of the experiment. One segment of the experiment utilizes data up to May 31, 2020 (ending date of the fourth lockdown in India), while the other exploits the entire extracted timeline. The results have indicated that both web searches and Twitter posts follow a positive correlation with the deviation in the number of detected cases.

B. Lag Correlation Analysis

Statistical analysis is carried out for each of the extracted web-based data with the deviation in frequency of laboratory-confirmed COVID-19 cases in India. This is to analyze the shift in the Indian COVID-19 cases and establish its associative nature with the diurnal Indian social media posts and web searches. The lag correlation analysis is conducted separately on the segregated datasets, and the lags up to 110 and 302 days are analyzed for the three separate studies. The study is conducted to provide an extensive analysis of the propagation of social activities with the drift in cases by applying three popular rank correlation algorithms: 1) Pearson correlation coefficient; 2) Kendall's tau correlation coefficient; and 3) Spearman correlation coefficient. Each of the web-search frequencies from Google Trends and social media post frequency is separately analyzed [45]–[47].

C. Predictive Models

The predictive models are built at the final segment of the proposed method. The primary goal of this segment of the study is to predict the count of COVID-19 cases by taking the web data and air quality data as independent elements. The predictive layer of the experimental framework is formed by five regression techniques: 1) polynomial regression (PolynomialR); 2) support vector regression (SVR); 3) multilayer perceptron regression (MLPR); 4) ElasticNet; and 5) deep neural network (DNN). PolynomialR has been experimented with in several studies in context to COVID-19 case prediction. One study [48] analyzed and predicted COVID-19 cases by using PolynomialR. This approach was found to be successfully established within the time period of early 2020 by reaching an RMSE score of 1.72 for the model. The mathematical illustration of PolynomialR is [48], [49]

$$\gamma = \sum_{i=0}^{N} \Phi_i \sigma_i \tag{1}$$

where Φ is the partial coefficient, σ is the independent variable, and N is the polynomial degree. In this study, a linear model on a polynomial feature matrix of degree 2 has been employed.

Herein, the experiment is followed by the application of SVR. The regression model has also been applied in the virus outbreak prediction in Indian data [50]. This motivated the authors to incorporate the model in the experiment. The SVR [51], [52] is a nonparametric architecture that is formulated to make predictions based on a hyperplane function and support vectors resembling the closest points to the function. The probability of a correct prediction can be stated to be directly proportional to its distance from the hyperplane. The mathematical definition of the plane can be stated as

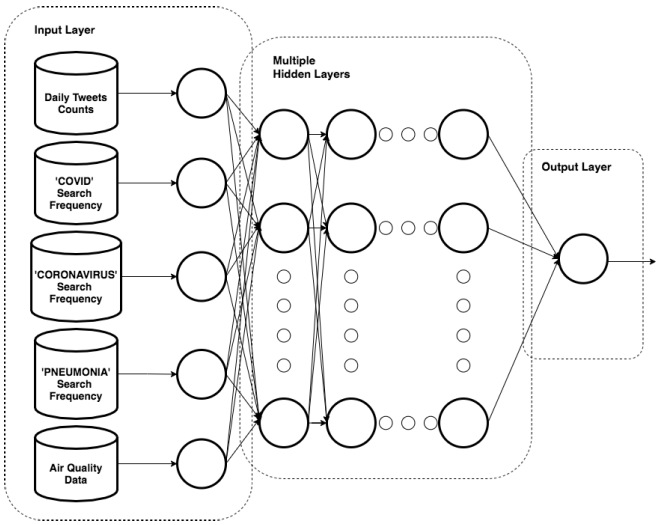

Fig. 3. DNN architecture.

$$k = \beta_l + \sum_{t=1} q_t x_t \tag{2}$$

where x_t is an n-dimensional input vector, q_t is a weight vector, and β_l is the bias [53]. The SVR used for the experiment uses an RBF kernel with a regularization parameter of 1. The results are then compared by applying the ElasticNet regression model, which is introduced in [54]. The model is well-explored in multiple studies and has been effective in reliable prediction compared to other explored techniques. This is a modification over the traditional linear regression model by achieving a lasso and ridge regularization-based technique to minimize the β parameter [55], [56]. The ElasticNet used for the experiment is trained for 1000 epochs with a mixing parameter of 0.5 and an alpha value of 1. The MLPR for the experiment is constructed with a single hidden layer of 100 units and trained for 200 epochs with Adam optimization and a constant learning rate of 10^{-3}. To undertake a concrete investigation, the authors further explored shallow and deep neural models in order to justify the experiment [57]. Fig. 3 provides a pictorial representation of a typical DNN. The weights assigned in each of the synapses get updated with each epoch with a goal to minimize the residual cost. For the experiment, the DNN is constructed by adding three layers of 1-D convolutional layers with a constant learning rate of 10^{-3}. The network is trained for 500 iterations with Adam optimization.

III. Experimental Analysis and Results

The primary intention of the experiment is to analyze the propagation of web-based data during the peak of COVID-19 outbreak tenure in India and understand their effectiveness in predicting new cases. To achieve the optimal understanding of the idea, the authors segregated the investigation into three separate studies:

1) extensive analysis of COVID-19-related web-based data and newly identified case frequency propagation in a range covering the four phases of nationwide COVID-19 lockdown; its effectiveness in machine learning and deep learning-based prediction of COVID-19 cases by fusing air quality data;
2) a special justification of the proposed study by assessing its impact during the period covering lockdown and unlock phases in India;
3) a special study on the impact of the second COVID-19 wave in India.

The collected data are divided into three separate datasets with dissimilar densities. Machine learning and deep learning models are applied at the final stage of the experiment. Models are evaluated by using a tenfold cross-validation technique. The performances of the models are calculated by using the root mean square error (RMSE) and mean absolute error (MAE) metrics defined as

$$\text{RMSE} = \sqrt{\sum_{i=0}^{n} \frac{(\sigma_{p,i} - \sigma_{t,i})^2}{n}} \tag{3}$$

$$\text{MAE} = \frac{\sum_{i=0}^{n} |\sigma_{p,i} - \sigma_{t,i}|}{n}. \tag{4}$$

A. Analysis and New Case Prediction During the Nationwide Lockdown

The study focusing on the impact of national lockdown in India is conducted over a span of 132 days starting from January 22, 2020. The quotidian steady increment in the number of COVID-19-related tweets is distinctly shown in Fig. 4. It is found that the highest count is documented soon after the end of the fourth phase of lockdown in India. It can be seen that the frequency of daily tweets undergoes a sharp fall after the month of August. An average count of 1000 tweets is observed in the month of December, which is 4000 less than the highest count documented within a range of June to August. The figure illustrates that the propagation of daily Twitter posts is analogous to the trajectory of the newly identified COVID-19 cases. This supports our assumption that the drift of tweets that are addressing COVID-19-related information is influenced by the daily increment in cases. It is visible in the figure that the search frequency of keywords, such as "COVID" and "pneumonia," attains peak during the month of April and May. It can, thus, be inferred that home quarantine during the period of lockdown in the country has resulted in a higher frequency of searches. This special study is intended to analyze the constructiveness of the proposed approach during a nationwide-lockdown situation in India. The entire population of the country is 1.3 billion as of December 2020 [58]. By the end of the second phase of the nationwide lockdown, the count of daily confirmed cases documented a number more than 35 000 in the country. Further analysis has revealed that 2/7th of this count is documented in the region of Maharashtra, which is situated in the southwestern part of the country. By the end of the fourth phase, the total number of documented cases crossed 180 000. From Fig. 6, it is found

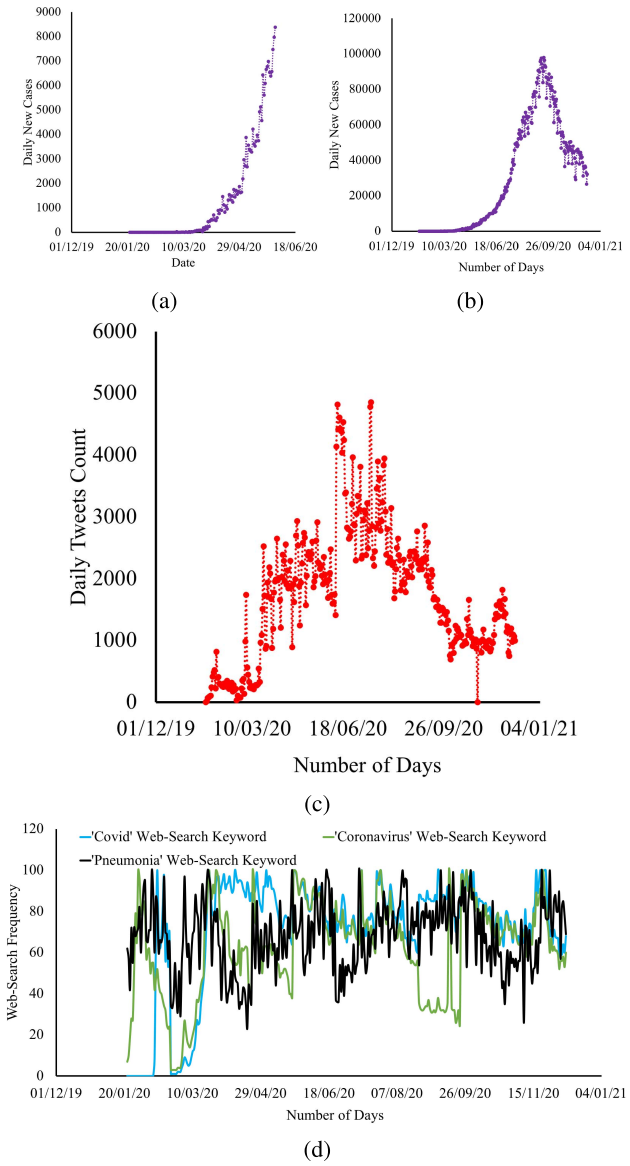

Fig. 4. Data frequency. (a) COVID-19 new cases during nationwide lockdown. (b) COVID-19 new cases during lockdown and unlock phases. (c) Frequency of daily Tweets. (d) Frequency of web search.

that, by the end of the second phase, the maximum number of COVID-19-related tweets is from the region of Maharashtra. The cumulative count is documented above 12 000 followed by that from the Uttar Pradesh region. The region of Maharashtra is found to be the most prominently affected region by Phase 2 of the nationwide lockdown. The region has recorded the highest number of COVID-19 cases and deaths related to the virus as of May 1, 2020. The region is found to be home to over 8000 cases and over 400 deaths. This is followed by regions such as Gujarat and Madhya Pradesh belonging to the western and central parts of the country. The average air quality level during the lockdown period in the country is measured based on the levels of particulate matter, nitrogen oxides (NO_x), ammonia gas (NH_3), carbon monoxide (CO), and sulfur dioxide (SO_2). The presence of NO_x is measured by the existence of nitric oxide (NO), nitric dioxide (NO_2),

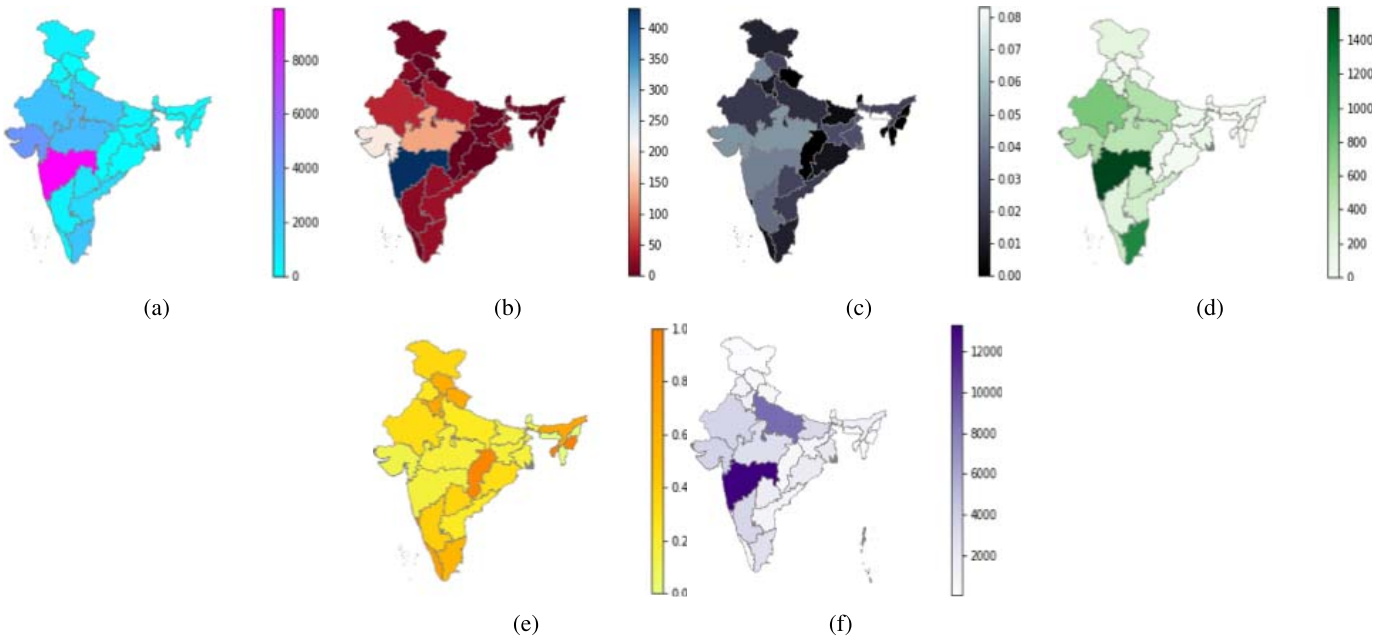

Fig. 5. Statewise distribution during the end of Phase 2 of COVID-19 lockdown. (a) COVID-19 cases. (b) Number of deaths related to COVID-19 outbreak. (c) Deaths per COVID-19 cases. (d) Number of recovered COVID-19 cases. (e) Recovery count per COVID-19 cases. (f) COVID-19-related tweets' density.

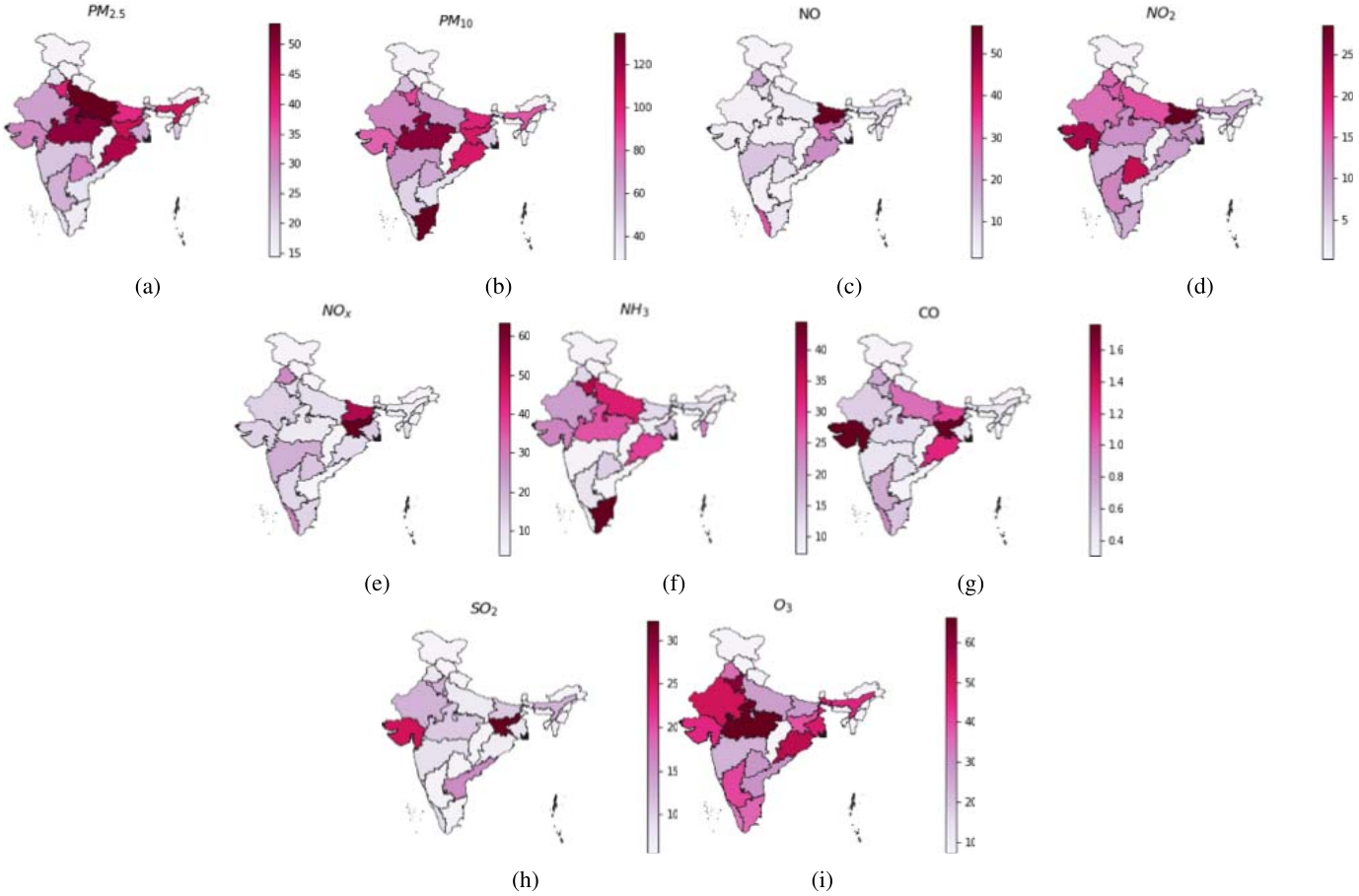

Fig. 6. Regional distribution of air quality levels during the four phases of nationwide lockdown. (a) Fine particulate matters with a diameter of 2.5 μm or less. (b) Fine particulate matters with diameter between 2.5 and 10 μm. (c) NO. (d) NO_2. (e) NO_x. (f) NH_3. (g) CO. (h) SO_2. (i) Ozone gas (O_3).

and other oxides of nitrogen represented as NO_x. The regional distribution of the particles constructed in Fig. 7 is based on the average levels documented during the first two lockdowns in India. The variation of the levels is pictorially illustrated for the country in Fig. 7. It is evident from the figure that, in the active phases of lockdown, the air becomes clearer

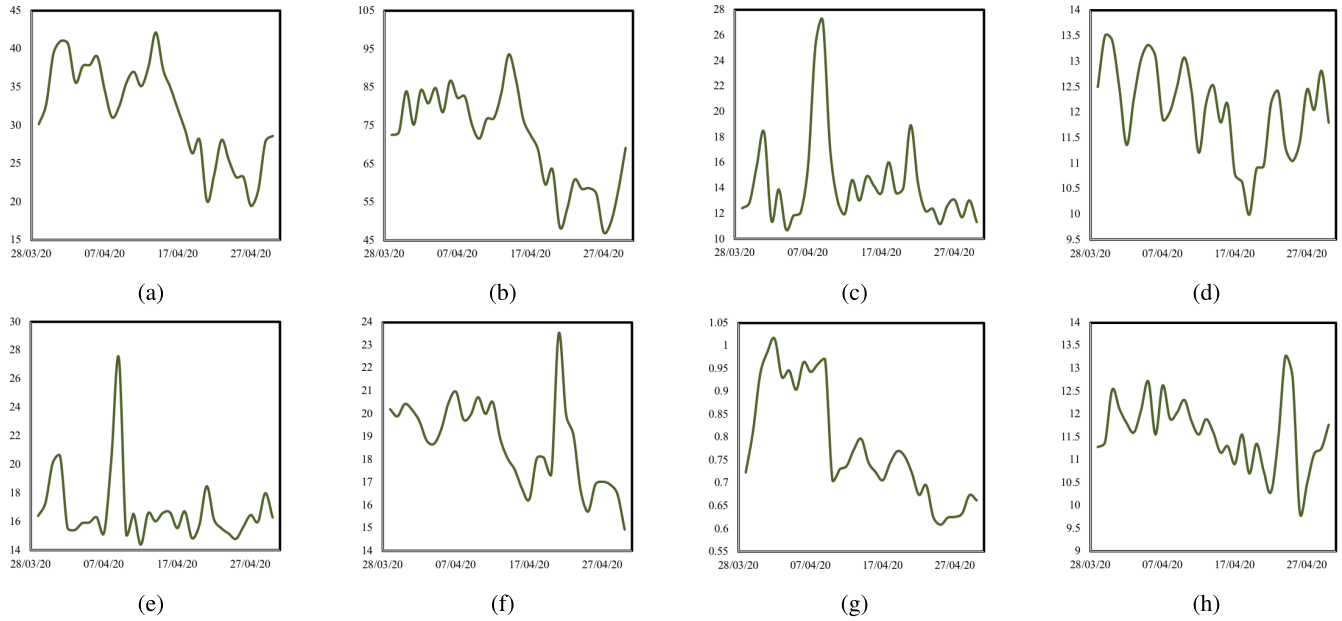

Fig. 7. Plots of air quality levels during first two lockdown phases. (a) Fine particulate matters with a diameter of 2.5 μm or less. (b) Fine particulate matters with a diameter between 2.5 and 10 μm. (c) NO. (d) Nitrogen dioxide (NO_2). (e) NO_x. (f) NH_3. (g) CO. (h) SO_2.

every day by recording a diurnal decrease in the quantity of fine particulate matters, NH_3, CO, and SO_2. This observation distinctly illustrates the effect of nationwide lockdown on air quality.

The plots in Fig. 8 represent the lag correlation values obtained between the web-related social data and daily new cases data collected from late January 2020 till the end of phase 4 of the lockdown period in India. The frequency of social media posts is found to be highly correlated with the daily increment of COVID-19 cases. The trajectory depicts the highest Pearson r-value between the daily tweets count and daily laboratory documented COVID-19 cases ($r > 0.85$) between a lag range of 40 and 60 days. This range has also resulted in the highest correlation for the search frequencies of "COVID" and "coronavirus" web-search keywords. The highest Spearman correlation value between the "COVID" search keywords and the daily new cases is recorded to be greater than 0.8, whereas the highest r-value for the "coronavirus" search keyword is documented as greater than 0.5. However, the r-value obtained between the "pneumonia" web-search and daily documented cases is found to portray a negative correlation within a lag of the first 60 days and then again between a lag of 80 and 100 days. The optimal Pearson correlation value ($r > 0.59$) is recorded between a lag period of 100 and 110 days. Each of the web-based data is found to be negatively correlated with the daily frequency of laboratory-confirmed cases within a lag period of 90–100 days (three months). As illustrated in Fig. 4, the count of COVID-19 cases has been steadily increasing during the months of May and June. It is observed that the daily frequency of web-search data and tweets presented a descending trajectory in the month of February, which is three months prior to the month of May. This, in turn, resulted in a negative correlation. This can also be justified by the obtained P-values in the corresponding lag period. Fig. 9 illustrates that the hypotheses fail for tweets and web-data frequency in the lag period greater than 90 days.

In Table I, observations have been recorded for predicting new COVID-19 cases from tweets and search keywords on the dataset containing data points from January to May. MLP regression has achieved the lowest RMSE value of 0.19 and the lowest MAE value of 0.13; thus, it becomes a preferable choice. On the other hand, Elastic Net is the least recommended choice for an intelligent model. It yields a score as high as 0.25 when search keywords are employed to predict new COVID-19 cases. Predicting these cases from tweets and search keywords together yields a better outcome when PolynomialR is used. By observing the results, it is found that tweets were more considerate in predicting new COVID-19 cases than search keywords. In Table II, initially, the air quality data are used to predict the new COVID-19 cases, and later web data are cumulated to investigate the changes in observations. The lowest RMSE and MAE values are observed for the SVR regression when considering only air quality data. Elastic Net is comparatively unsuccessful in providing good RMSE and MAE values. As the RMSE value is greater or equal to the MAE value, attention is preferably given to the RMSE values for the comparison of models. The RMSE score for SVR regression is 0.13 and is ideal for predicting the new COVID-19 cases. A nominal deterioration of score can be witnessed for DNN when web data are cumulated to the air quality data. However, the highest score of 0.63 can be observed for PolynomialR when the aforementioned inclusion takes place.

B. Special Analysis Combining the Lockdown and Unlock Phases

On analyzing the correlation values obtained through data collected within a range of January 22 and December 9,

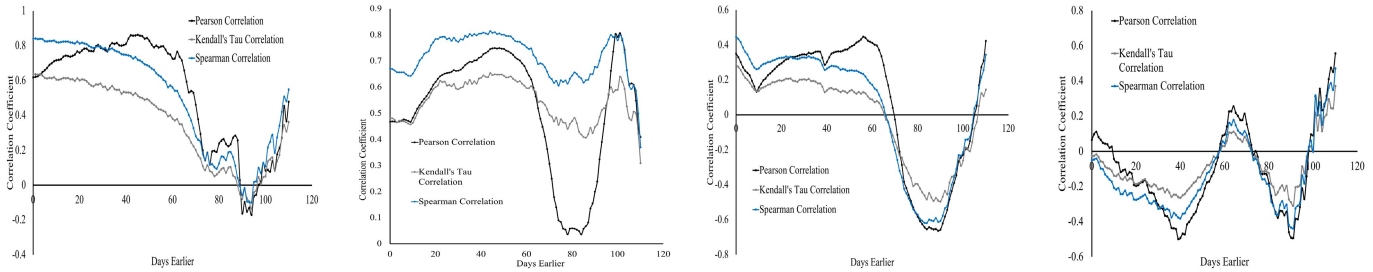

Fig. 8. Correlation coefficient between January 22, 2020, and May 31 (end of fourth lockdown period). (a) Tweets. (b) COVID search keyword. (c) Coronavirus search keyword. (d) Pneumonia search keyword.

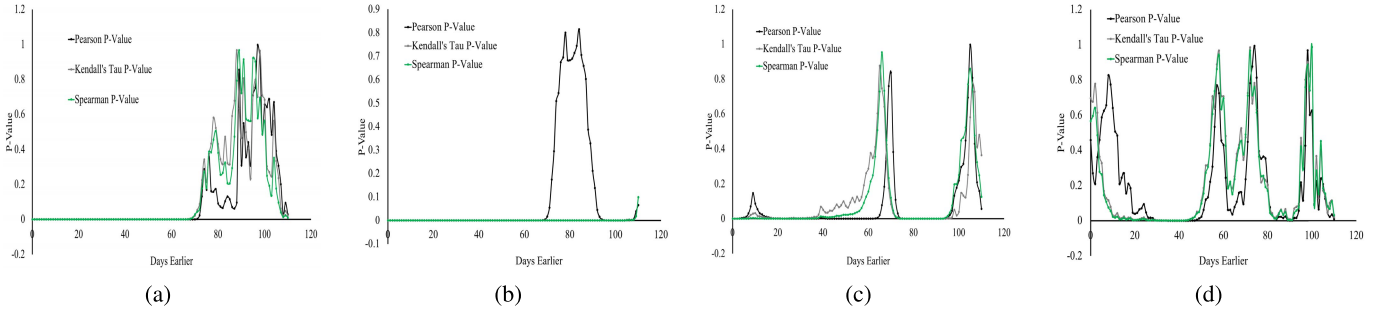

Fig. 9. P-Value obtained between January 22, 2020, and May 31 (end of fourth lockdown period). (a) Tweets. (b) COVID search keyword. (c) Coronavirus search keyword. (d) Pneumonia search keyword.

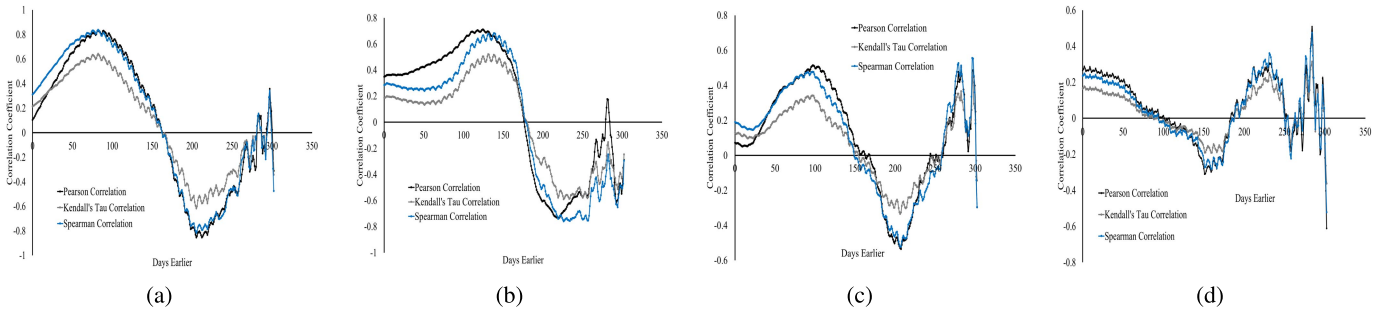

Fig. 10. Correlation coefficient between January 22, 2020, and December 9. (a) Tweets. (b) COVID search keyword. (c) Coronavirus search keyword. (d) Pneumonia search keyword.

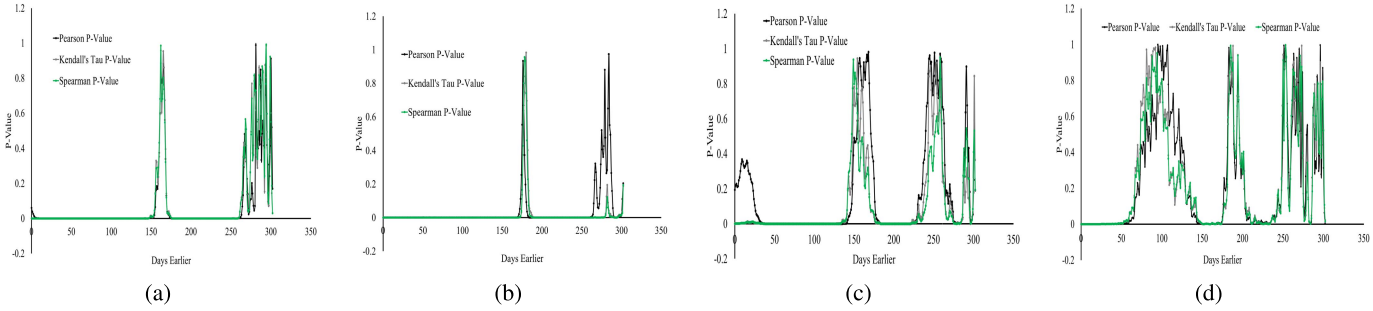

Fig. 11. P-Value obtained between January 22, 2020, and December 9, 2020. (a) Tweets. (b) COVID search keyword. (c) Coronavirus search keyword. (d) Pneumonia search keyword.

2020, it is found that a high correlation exists between the daily frequency of COVID-19-related tweets and daily new cases. As shown in Fig. 10, the optimal r-value ($r > 0.82$) is obtained within a lag range of 70 and 100 days. The correlation follows a steady increase within a lag range of 100 days and then degrades up to the lowest r-value of -0.8 around a lag of 200 days. The correlation between the "COVID" web-search keyword and the daily laboratory-confirmed cases

TABLE I

WEB-BASED PREDICTION RESULTS DURING LOCKDOWN

X Values	Measure	PolynomialR	SVR	MLPR	ElasticNet	DNN
Tweets	RMSE±SD	0.19±0.07	0.21±0.05	0.19±0.047	0.24±0.09	0.24±0.07
	MAE±SD	0.14±0.05	0.16±0.03	0.13±0.041	0.2±0.066	0.126±0.05
SK	RMSE±SD	0.23±0.05	0.22±0.06	0.22±0.064	0.25±0.043	0.23±0.14
	MAE±SD	0.18±0.04	0.16±0.043	0.16±0.047	0.19±0.024	0.13±0.099
Tweets+SK	RMSE±SD	0.21±0.026	0.19±0.067	0.21±0.056	0.25±0.063	0.22±0.11
	MAE±SD	0.15±0.018	0.13±0.042	0.15±0.04	0.19±0.046	0.11±0.06

TABLE II

CUMULATED PREDICTION USING AIR QUALITY AND WEB DATA

X Values	Measure	PolynomialR	SVR	MLPR	ElasticNet	DNN
Air Quality	RMSE±SD	0.19±0.11	0.13±0.051	0.17±0.062	0.25±0.03	0.22±0.067
	MAE±SD	0.14±0.09	0.09±0.022	0.12±0.032	0.19±0.04	0.1±0.05
All	RMSE±SD	0.63±0.23	0.13±0.04	0.17±0.05	0.25±0.063	0.21±0.08
	MAE±SD	0.44±0.18	0.1±0.023	0.13±0.04	0.2±0.04	0.11±0.05

TABLE III

WEB-BASED PREDICTION RESULTS FROM JANUARY TO DECEMBER

X Values	Measure	PolynomialR	SVR	MLPR	ElasticNet	DNN
Tweets	RMSE±SD	0.29±0.017	0.29±0.028	0.31±0.024	0.31±0.02	0.36±0.04
	MAE±SD	0.25±0.016	0.21±0.023	0.27±0.02	0.27±0.018	0.25±0.04
SK	RMSE±SD	0.26±0.035	0.24±0.033	0.27±0.032	0.31±0.014	0.25±0.038
	MAE±SD	0.22±0.036	0.19±0.025	0.23±0.033	0.27±0.018	0.16±0.027
Tweets+SK	RMSE±SD	0.25±0.015	0.22±0.052	0.3±0.025	0.31±0.021	0.24±0.065
	MAE±SD	0.19±0.015	0.16±0.036	0.27±0.024	0.28±0.02	0.14±0.045

was found to be optimal ($r > 0.75$) 150 days earlier. The degrading correlation trajectory of the web search is found to demonstrate a path similar to the one obtained for daily tweets count. This is also valid for the correlation trajectory obtained for the "coronavirus" web search. The figure shows that the highest correlation between newly documented laboratory cases and entities, such as daily tweets count, "COVID," and "coronavirus" web-search frequencies, is obtained within a lag range of 50–150 days. However, the optimal correlation value ($r > 0.57$) between the daily documented COVID-19 cases and the "pneumonia" web-search frequency is found 250–300 days earlier. Each of the collected web-based data is found to be negatively correlated with the newly documented COVID-19 cases for a lag period between 150 and 200 days (five to seven months). This result can be justified with the pictorial representation of data propagation in our projected range of study presented in Fig. 4. A decrease in the count of newly detected cases is found to experience a sharp drop in a range between late September to early December. It is shown in the figure that the daily frequency of Twitter posts and web searches is increasing five to seven months prior to the range. The consequence of this is the negative correlation observed in Fig. 9. This is also valid for the correlation values obtained at the maximum range of lag (>300 days) in our analysis. The result can be further justified by the P-values represented in Fig. 11.

In Table III, tweets and search keywords are separately and cooperatively tested for predicting new COVID-19 cases. When PolynomialR is employed as an intelligent model, the RMSE value is observed to be the lowest. As lower values of RMSE define the practicality of the model, PolynomialR is found to be the best model in this experiment. Elastic Net yields a score as high as 0.31, and DNN exhibits the highest score of 0.36 when tweets are applied to predict new COVID-19 cases. However, DNN performs adroitly when search keywords are used to predict new cases. It can be observed that the intelligent models perform in a superior manner when collaborated with search keywords.

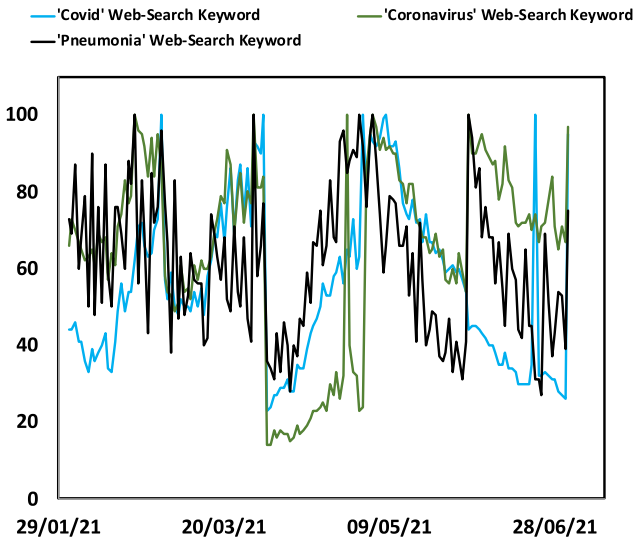

Fig. 12. Frequency of web search during the second wave.

C. Analysis of the Second Wave of COVID-19 Outbreak in India

To analyze the data during the second wave in India, a similar study is performed by utilizing the daily tweets and web-search frequency obtained between February 1, 2021, and July 2, 2021 [42]. From Figs. 12, 13, and 14, it is visible that, in late March 2021, the search frequencies of

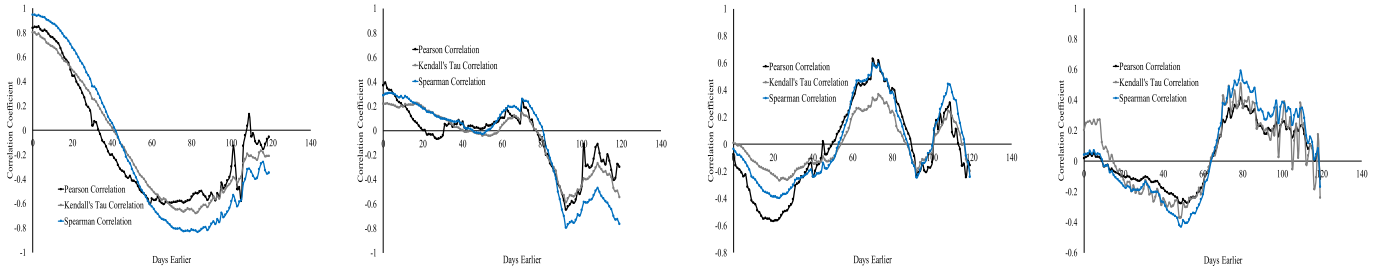

Fig. 13. Correlation coefficient during the second wave. (a) Tweets. (b) COVID search keyword. (c) Coronavirus search keyword. (d) Pneumonia search keyword.

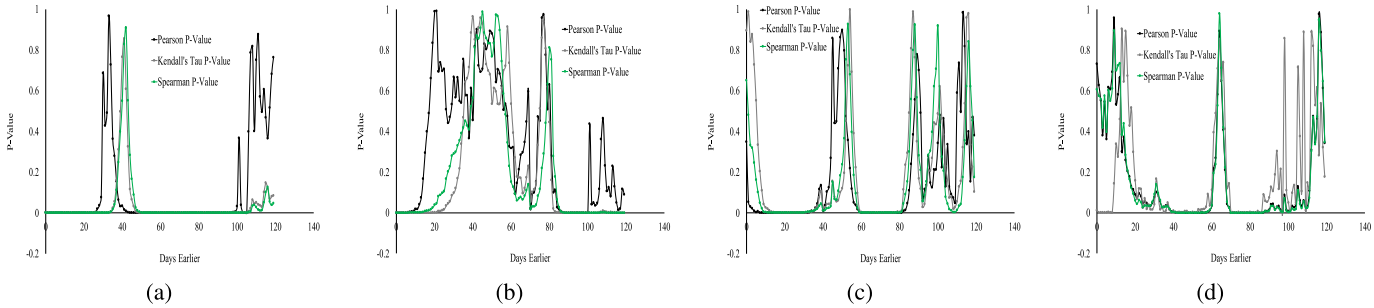

Fig. 14. P-value obtained during the second wave. (a) Tweets. (b) COVID search keyword. (c) Coronavirus search keyword. (d) Pneumonia search keyword.

each keyword depicted a sharp decline and are found to be even low throughout the month of April. However, this was followed by a gradual rise in the first week of May. From this, it is understood that, at the beginning of the second wave, people lost interest in searching essential keywords on Google. Similar to the results obtained for the first wave, it is observed that, even for the second wave, the highest correlation is obtained within a lag range of 70–100 days. As illustrated in Fig. 13, the lowest correlation is found for the daily frequency of COVID-19-related tweets and new cases at a lag period of 40 days. However, a high positive correlation is visible at a zero-day-lag range ($r > 0.97$) and a high negative correlation at a 75-day-lag range ($r < -0.8$). This is similar to what is exhibited in Fig. 8. Fig. 13 further reveals that the correlation with the number of new COVID-19 cases and frequency of tweets, COVID search keywords, coronavirus search keywords, and pneumonia search keywords is changing as the lag period is increased from zero to 120 days. In the case of tweet frequency, it can be observed that, till 40-day lag, the correlation remains positive, and after that, it remains negative till 110 days. This indicates that the frequency of the COVID-19-related tweets has started decreasing with an increase in the number of COVID-19 cases after 40 days. Furthermore, it can be noted that the peak of the second wave is reported to be on May 8, 2021, which is, approximately, the exact lag period ahead of the starting day of the study when the correlation reaches the minimum value, whereas an exactly opposite nature is reflected by the coronavirus search keyword frequency. This leads to the conclusion that an increase in COVID-19-related tweets indicates an increase in the number of COVID-19 cases only at the beginning (approximately 20 days) of a wave, whereas coronavirus-related keywords frequency may be useful to infer the number of cases after a 40-day-lag period. The highest positive correlations for the web-search keywords are found at a lag of 70 days.

IV. Conclusion

An extensive analysis of COVID-19 social media data collected from Twitter and search engine trends is used to predict daily COVID-19 cases in India. Data are collected and used for the pandemic period of the entire 2020. A separate study has been carried out to better understand the interrelation between social media information and daily new cases during the lockdown and the rest of the period. Lag correlation analysis has revealed that daily new cases can be predicted 70–100 days ahead with reasonable confidence. In order to find the effect of air quality in daily new cases of COVID-19 during the lockdown, an extensive analysis is done, and it has been revealed that the presence of air quality indicators in prediction models has improved the performance. MLPR and SVR regression models resulted in RMSE scores of 0.17 and 0.13, respectively, by predicting on the basis of air quality data during the lockdown. MLPR was also found to be successful in predicting the new cases from the frequency and resulted in an RMSE score of 0.19. The DNN-based model is found to be effective in predicting daily new cases in terms of social media data extracted from Twitter, Internet search engine trends, and air quality data in India. Future studies can be directed toward understanding the effects of social network users and content interactions in the prediction of the COVID-19 pandemic.

REFERENCES

[1] World Health Organization (WHO). (2020). *Coronavirus Disease (COVID-19) Pandemic*. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019

[2] Q. Lin *et al.*, "A conceptual model for the coronavirus disease 2019 (COVID-19) outbreak in wuhan, China with individual reaction and governmental action," *Int. J. Infectious Diseases*, vol. 93, pp. 211–216, Apr. 2020.

[3] R. Vaman, M. Valamparampil, A. Ramdas, A. Manoj, B. Varghese, and F. Joseph, "A confirmed case of COVID-19 among the first three from Kerala, India," *Indian J. Med. Res.*, vol. 151, no. 5, p. 493, 2020.

[4] K. Sarkar, S. Khajanchi, and J. J. Nieto, "Modeling and forecasting the COVID-19 pandemic in India," *Chaos, Solitons Fractals*, vol. 139, Oct. 2020, Art. no. 110049.

[5] P. Pulla, "The epidemic is growing very rapidly': Indian government adviser fears coronavirus crisis will worsen," *Nature*, vol. 583, no. 7815, p. 180, 2020.

[6] (2020). *India COVID-19 Tracker*. [Online]. Available: https://www.covid19india.org/

[7] T. Lancet, "India under COVID-19 lockdown," *Lancet*, vol. 395, no. 10233, p. 1315, Apr. 2020.

[8] A. Maji, T. Choudhari, and M. B. Sushma, "Implication of repatriating migrant workers on COVID-19 spread and transportation requirements," 2020, *arXiv:2005.04424*.

[9] S. M. Dev and R. Sengupta, *COVID-19: Impact on the Indian economy*. Mumbai, India: Indira Gandhi Institute of Development Research, Apr. 2020.

[10] V. K. R. Chimmula and L. Zhang, "Time series forecasting of COVID-19 transmission in Canada using LSTM networks," *Chaos, Solitons Fractals*, vol. 135, Jun. 2020, Art. no. 109864.

[11] F. Piccialli, V. S. di Cola, F. Giampaolo, and S. Cuomo, "The role of artificial intelligence in fighting the COVID-19 pandemic," *Inf. Syst. Frontiers*, vol. 23, pp. 1467–1497, Apr. 2021.

[12] K. N. Nabi, M. T. Tahmid, A. Rafi, M. E. Kader, and M. A. Haider, "Forecasting COVID-19 cases: A comparative analysis between recurrent and convolutional neural networks," *Results Phys.*, vol. 24, May 2021, Art. no. 104137.

[13] I. Rahimi, F. Chen, and A. H. Gandomi, "A review on COVID-19 forecasting models," *Neural Comput. Appl.*, pp. 1–11, Feb. 2021.

[14] S. Shastri, K. Singh, S. Kumar, P. Kour, and V. Mansotra, "Time series forecasting of COVID-19 using deep learning models: India-USA comparative case study," *Chaos, Solitons Fractals*, vol. 140, Nov. 2020, Art. no. 110227.

[15] D. Liu *et al.*, "A machine learning methodology for real-time forecasting of the 2019–2020 COVID-19 outbreak using internet searches, news alerts, and estimates from mechanistic models," 2020, *arXiv:2004.04019*.

[16] A. Goel and L. Gupta, "Social media in the times of COVID-19," *JCR, J. Clin. Rheumatol.*, vol. 26, no. 6, pp. 220–223, 2020.

[17] C. Li, L. J. Chen, X. Chen, M. Zhang, C. P. Pang, and H. Chen, "Retrospective analysis of the possibility of predicting the COVID-19 outbreak from internet searches and social media data, China, 2020," *Eurosurveillance*, vol. 25, no. 10, Mar. 2020, Art. no. 2000199.

[18] Y.-H. Lin, C.-H. Liu, and Y.-C. Chiu, "Google searches for the keywords of 'wash hands' predict the speed of national spread of COVID-19 outbreak among 21 countries," *Brain, Behav., Immunity*, vol. 87, pp. 30–32, Jul. 2020.

[19] X. Yuan, J. Xu, S. Hussain, H. Wang, N. Gao, and L. Zhang, "Trends and prediction in daily new cases and deaths of COVID-19 in the United States: An internet search-interest based model," *Explor. Res. Hypothesis Med.*, vol. 5, no. 2, pp. 1–6, Apr. 2020.

[20] A. Mavragani, "Tracking COVID-19 in Europe: Infodemiology approach," *JMIR Public Health Surveill.*, vol. 6, no. 2, Apr. 2020, Art. no. e18941.

[21] S. M. Ayyoubzadeh, S. M. Ayyoubzadeh, H. Zahedi, M. Ahmadi, and S. R N. Kalhori, "Predicting COVID-19 incidence through analysis of Google trends data in Iran: Data mining and deep learning pilot study," *JMIR Public Health Surveill.*, vol. 6, no. 2, Apr. 2020, Art. no. e18828.

[22] A. Husnayain, A. Fuad, and E. C.-Y. Su, "Applications of Google search trends for risk communication in infectious disease management: A case study of the COVID-19 outbreak in Taiwan," *Int. J. Infectious Diseases*, vol. 95, pp. 221–223, Jun. 2020.

[23] Y. Ortiz-Martínez, J. E. Garcia-Robledo, D. L. Vásquez-Castañeda, D. K. Bonilla-Aldana, and A. J. Rodriguez-Morales, "Can Google trends predict COVID-19 incidence and help preparedness? The situation in Colombia," *Travel Med. Infectious Disease*, vol. 37, Sep. 2020, Art. no. 101703.

[24] C. Shen, A. Chen, C. Luo, J. Zhang, B. Feng, and W. Liao, "Using reports of symptoms and diagnoses on social media to predict COVID-19 case counts in mainland China: Observational infoveillance study," *J. Med. Internet Res.*, vol. 22, no. 5, May 2020, Art. no. e19421.

[25] P. Majumdar, A. Biswas, and S. Sahu, "COVID-19 pandemic and lockdown: Cause of sleep disruption, depression, somatic pain, and increased screen exposure of office workers and students of India," *Chronobiol. Int.*, vol. 37, no. 8, pp. 1191–1200, Aug. 2020.

[26] X. Yuan, J. Xu, S. Hussain, H. Wang, N. Gao, and L. Zhang, "Trends and prediction in daily new cases and deaths of COVID-19 in the United States: An internet search-interest based model," *Explor. Res. Hypothesis Med.*, vol. 5, no. 2, pp. 1–6, Apr. 2020.

[27] S. Zangari, D. T. Hill, A. T. Charette, and J. E. Mirowsky, "Air quality changes in New York city during the COVID-19 pandemic," *Sci. Total Environ.*, vol. 742, Nov. 2020, Art. no. 140496.

[28] M. Masum and S. Pal, "Statistical evaluation of selected air quality parameters influenced by COVID-19 lockdown," *Global J. Environ. Sci. Manage.*, vol. 6, pp. 85–94, Aug. 2020.

[29] M. A. Zambrano-Monserrate and M. A. Ruano, "Has air quality improved in Ecuador during the COVID-19 pandemic? A parametric analysis," *Air Qual., Atmos. Health*, vol. 13, no. 8, pp. 929–938, Aug. 2020.

[30] D. Fattorini and F. Regoli, "Role of the chronic air pollution levels in the COVID-19 outbreak risk in Italy," *Environ. Pollut.*, vol. 264, Sep. 2020, Art. no. 114732.

[31] R. P. Singh and A. Chauhan, "Impact of lockdown on air quality in India during COVID-19 pandemic," *Air Qual., Atmos. Health*, vol. 13, no. 8, pp. 921–928, Aug. 2020.

[32] *Manual Monitoring Data: Central Pollution Control Board, India*. Accessed: Jan. 15, 2021. [Online]. Available: https://cpcb.nic.in/manual-monitoring/

[33] L. Li *et al.*, "Characterizing the propagation of situational information in social media during COVID-19 epidemic: A case study on Weibo," *IEEE Trans. Comput. Social Syst.*, vol. 7, no. 2, pp. 556–562, Mar. 2020.

[34] K. Xu, F. Wang, H. Wang, Y. Wang, and Y. Zhang, "Mitigating the impact of data sampling on social media analysis and mining," *IEEE Trans. Comput. Social Syst.*, vol. 7, no. 2, pp. 546–555, Apr. 2020.

[35] S. Priya, M. Bhanu, S. K. Dandapat, K. Ghosh, and J. Chandra, "TAQE: Tweet retrieval-based infrastructure damage assessment during disasters," *IEEE Trans. Comput. Social Syst.*, vol. 7, no. 2, pp. 389–403, Apr. 2020.

[36] P. K. Pandey and M. Singh, "Quantifying nonrandomness in evolving networks," *IEEE Trans. Comput. Social Syst.*, vol. 7, no. 6, pp. 1447–1459, Dec. 2020.

[37] X. Dong, U. Victor, and L. Qian, "Two-path deep semisupervised learning for timely fake news detection," *IEEE Trans. Comput. Social Syst.*, vol. 7, no. 6, pp. 1386–1398, Dec. 2020.

[38] A. Banerjee, M. Bhattacharjee, K. Ghosh, and S. Chatterjee, "Synthetic minority oversampling in addressing imbalanced sarcasm detection in social media," *Multimedia Tools Appl.*, vol. 79, nos. 47–48, pp. 35995–36031, Dec. 2020.

[39] G. Shrivastava, P. Kumar, R. P. Ojha, P. K. Srivastava, S. Mohan, and G. Srivastava, "Defensive modeling of fake news through online social networks," *IEEE Trans. Comput. Social Syst.*, vol. 7, no. 5, pp. 1159–1167, Oct. 2020.

[40] P. Gupta, S. Kumar, R. R. Suman, and V. Kumar, "Sentiment analysis of lockdown in India during COVID-19: A case study on Twitter," *IEEE Trans. Comput. Social Syst.*, vol. 8, no. 4, pp. 992–1002, Aug. 2021.

[41] J. Zhou, H. Zogan, S. Yang, S. Jameel, G. Xu, and F. Chen, "Detecting community depression dynamics due to COVID-19 pandemic in Australia," *IEEE Trans. Comput. Social Syst.*, vol. 8, no. 4, pp. 982–991, Aug. 2021.

[42] M. Safi. (2021). *India's Shocking Surge in COVID Cases Follows Baffling Decline*. [Online]. Available: https://www.worldometers.info/world-population/india-population/

[43] R. Lamsal, "Design and analysis of a large-scale COVID-19 tweets dataset," *Applied Intelligence*, vol. 51, no. 5, pp. 2790–2804, 2020.

[44] U. Venkatesh and P. A. Gandhi, "Prediction of COVID-19 outbreaks using Google trends in India: A retrospective analysis," *Healthcare Informat. Res.*, vol. 26, no. 3, pp. 175–184, Jul. 2020.

[45] A. Mavragani, "Tracking COVID-19 in Europe: Infodemiology approach," *JMIR Public Health Surveill.*, vol. 6, no. 2, Apr. 2020, Art. no. e18941.

[46] R. Ferrucci *et al.*, "Psychological impact during the first outbreak of COVID-19 in Italy," *Frontiers Psychiatry*, vol. 11, pp. 1–9, Nov. 2020.
[47] T. Fahrudin, D. R. Wijaya, and A. A. G. Agung, "COVID-19 confirmed case correlation analysis based on Spearman and Kendall correlation," in *Proc. Int. Conf. Data Sci. Its Appl. (ICoDSA)*, Aug. 2020, pp. 1–4.
[48] R. Gupta, G. Pandey, P. Chaudhary, and S. K. Pal, "Machine learning models for government to predict COVID-19 outbreak," *Digit. Government, Res. Pract.*, vol. 1, no. 4, pp. 1–6, Oct. 2020.
[49] M. Yadav, M. Perumal, and M. Srinivas, "Analysis on novel coronavirus (COVID-19) using machine learning methods," *Chaos, Solitons Fractals*, vol. 139, Oct. 2020, Art. no. 110050.
[50] D. Parbat and M. Chakraborty, "A Python based support vector regression model for prediction of COVID19 cases in India," *Chaos, Solitons Fractals*, vol. 138, Sep. 2020, Art. no. 109942.
[51] M. R. H. Mondal, S. Bharati, P. Podder, and P. Podder, "Data analytics for novel coronavirus disease," *Informat. Med. Unlocked*, vol. 20, Jan. 2020, Art. no. 100374.
[52] A. Mahmoodzadeh, M. Mohammadi, A. Daraei, H. F. H. Ali, A. I. Abdullah, and N. K. Al-Salihi, "Forecasting tunnel geology, construction time and costs using machine learning methods," *Neural Comput. Appl.*, vol. 33, pp. 321–348, Jan. 2020.
[53] Z. Malki, E.-S. Atlam, A. E. Hassanien, G. Dagnew, M. A. Elhosseini, and I. Gad, "Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches," *Chaos, Solitons Fractals*, vol. 138, Sep. 2020, Art. no. 110137.
[54] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Statist. Soc., B, Stat. Methodol.*, vol. 67, no. 2, pp. 301–320, 2005.
[55] T. K. Johnsen and J. Z. Gao, "Elastic net to forecast COVID-19 cases," in *Proc. Int. Conf. Innov. Intell. Informat., Comput. Technol. (3ICT)*, Dec. 2020, pp. 1–6.
[56] E. M. Giusti *et al.*, "The psychological impact of the COVID-19 outbreak on health professionals: A cross-sectional study," *Frontiers Psychol.*, vol. 11, p. 1684, Jul. 2020.
[57] F. Shahid, A. Zameer, and M. Muneeb, "Predictions for COVID-19 with deep learning models of LSTM, GRU and bi-LSTM," *Chaos, Solitons Fractals*, vol. 140, Nov. 2020, Art. no. 110212.
[58] Worldometer. (2020). *Coronavirus Disease (COVID-19) Pandemic*. [Online]. Available: https://www.worldometers.info/world-population/india-population/

Sankhadeep Chatterjee received the B.Tech. degree in computer science and engineering from the Maulana Abul Kalam Azad University of Technology, Kolkata, India, in 2015, and the M.Tech. degree in computer science and engineering from the University of Calcutta, Kolkata, in 2017. He is currently pursuing the Ph.D. degree with the Indian Institute of Engineering Science and Technology, Shibpur, Howrah, India.

He is currently an Assistant professor with the University of Engineering & Management, Kolkata. He has published and presented more than 60 research papers in reputed international journals/conferences. His current research interests include machine learning, deep learning, metaheuristics, and text data analysis.

Mr. Chatterjee obtained the prestigious Council of Scientific & Industrial Research (CSIR) Senior Research Fellowship from the Government of India in 2019.

Kushankur Ghosh finished his high school education in 2017. He is currently pursuing the bachelor's degree in computer science and engineering with the University of Engineering & Management, Kolkata, India.

He has been doing research on topics such as social computing and data-driven challenges. He has over 12 peer-reviewed publications in his name as an undergraduate researcher. Some of his major contributions are in domains such as sarcasm detection and sentiment analysis. His primary research interests are machine learning, data mining, and social network analysis.

Arghasree Banerjee is currently pursuing the bachelor's degree in computer science and engineering with the University of Engineering & Management, Kolkata, India.

She is an active researcher, currently focusing on mitigating the class imbalance problem prevalent in numerous genuine datasets in domains such as image and text classification. Having over 13 publications, she investigates research areas, such as data mining and computer vision.

Soumen Banerjee (Senior Member, IEEE) received the B.Sc. degree (Hons.) in physics and the B.Tech. and M.Tech. degrees in radiophysics and electronics from the University of Calcutta, Kolkata, India, in 1998, 2001, and 2003, respectively, and the Ph.D. degree in engineering from the Indian Institute of Engineering Science and Technology (IIEST), Shibpur, Howrah, India, in 2019.

He was the former Guest Faculty with the Department of Applied Physics, University of Calcutta. He has a teaching/research experience of 20+ years. He is currently the Head of the Department of Electronics and Communication Engineering, University of Engineering & Management, Kolkata. He has published and presented more than 100 contributory papers in international journals/conferences. He is the author of ten books and an editor of two books from Springer Nature. His current research interests include design, fabrication, and characterization of wide bandgap semiconductor-based IMPATT diodes at D-band, W-band, and THz frequencies, SIW technology-based antennas and arrays, printed antennas, frequency selective surface (FSS), dielectric resonator antennas, body wearable antennas, machine learning, computer vision, and optimization techniques.

He has chaired many technical sessions in IEEE international conferences. He is also a reviewer of many international journals of IEEE, Springer, Wiley, and so on. He has organized Optronix-2019 (IEEE) and Optronix-2020 (Springer).