2023 Sixth International Conference of Women in Data Science at Prince Sultan University (WiDS PSU) | 978-1-6654-7723-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/WiDS-PSU57071.2023.00049

Prediction of epidemic disease cases using ARIMA and SARIMAX models

Narendra Kumar
School of Computing
DIT University Dehradun
Uttarakhand, INDIA
narendra298@gmail.com

Vardhan Jain
School of Computing
DIT University Dehradun
Uttarakhand, INDIA
vardhanjaindit@gmail.com

Kritika Joshi
School of Computing
DIT University Dehradun
Uttarakhand, INDIA
jkritika531@gmail.com

Ishaan Dawar
School of Computing
DIT University Dehradun
Uttarakhand, INDIA
ishaan.dawar@dituniversity.edu.in

***Abstract*—Human movement has a significant influence on disease propagation. To suppress an outbreak, movement restrictions are imposed in locations classified as disease hotspots. It is essential to monitor the sickness on a local level in order to identify locations that need concentrated attention in order for several nations to succeed in preventing the further spread of COVID-19 without reimposing national bans. The goal of this research is to identify epidemic disease hotspots within local communities by utilizing publicly available health- related data and unstructured data analysis methods. We used data that was self-reported from the whole population and obtained via a mobile application, in combination with data from targeted PCR(polymerase chain reaction) testing, to develop estimates of the frequency and incidence of sickness across the area. We also demonstrate how these estimations may be utilized to identify infection hotspots in real time. A regression analysis is done based on the numerical data to determine the best regression model for these important parameters. The proposed regression models are useful for estimating hotspots under various ventilation settings. We used the ARIMA(Autoregressive Integrated Moving Average) and SARIMAX(Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors) models to identify hotspots for future COVID waves using a data set of COVID-19 cases.**

Keywords— COVID-19, ARIMA, SARIMAX, epidemic disease, Hotspot

I. INTRODUCTION

One example of an infectious disease that may spread from person to person is shown by the ongoing COVID-19 epidemic. Studies have shown, on the whole, that limiting people's movement may help reduce the spread of epidemics, but this alone is not enough to end the outbreaks. The World Health Organization (WHO)[1] proposes that, in the case of a pandemic, activities at educational institutions, government agencies, and private businesses should be suspended as a practical approach to reducing the risk of disease transmission. People find themselves in unpleasant situations when they have mobility restrictions, and these restrictions are not always obeyed. Law enforcement personnel may trace the whereabouts of those who violate the law in order to identify them and compel them to comply with the law. By linking the mobile phones of network users to base transceiver stations, also known as base stations (BSs), it is possible to monitor movement using data obtained from cellular networks. Because cellular networks reach the majority of inhabited areas, it is possible to make use of them for location tracking in a straightforward manner without the need for the construction of any additional infrastructure. At any one time, a BS will provide service to each mobile device user, and the purpose of each BS is to serve a certain geographical area. In order to provide their users with a better level of coverage, telecom service providers position base stations in key locations. These base stations have a high degree of accuracy. It is possible to ascertain the location of a user by locating the base station (BS) that serves that user. We are just aware of the place at which the BS is delivered, hence, it is crucial to bear in mind that we do not know the actual location. In areas with a high population density, there is also a high density of BSs, and the location information that is gathered may be fairly precise (within a few meters). Because users' affiliation with the BSs changes when they move about, it is possible to track where they go using this information. As a consequence of this, CDR might potentially be used to track the whereabouts of users. In fact, a number of countries have been using data from telecom companies to monitor the transmission of COVID-19. In this category, telecom service providers are also mentioned. A similar sign that a unaware user has entered the region is when a user who was not previously linked to any of the BSs in the area becomes connected with one of the BSs in the area. BS triangulation that enables a user to receive signals from a given location However, if one applies this line of reasoning, it becomes clear that it is impossible to monitor a hotspot whose boundaries are known in advance. This is because users located beyond the boundary may also be connected with the BSs, and their connections may be mistaken for those of users located within the region.

The development of Machine learning (ML) algorithms for early pandemic identification indicates that this may soon be a viable path to aiding improved preparation[27]. As these technologies continue to improve in accuracy, they are likely to play an important role in encouraging the development of ground-breaking health policy. This research presents preliminary evidence demonstrating that improved data sharing procedures will contribute to future urban health policy worldwide by surveying how data and Ml procedures helped in the early phases of identification of the COVID-19 epidemic [29].We have taken the data from website of Ministry of Health and Family Welfare government of india and API provided by https://www.covid19india.org/.

978-1-6654-7723-9/23/$31.00 ©2023 IEEE
DOI 10.1109/WiDS-PSU57071.2023.00049

This paper is organized as sections, Section II provides an overview of the related works background. Section III describes the suggested technique, which is based on machine learning; the block diagram of this proposed work is shown in this section. Section IV discusses the experimental outcomes. Finally, Sections V present a discussion and conclusion remarks, respectively.

II. LITERATURE REVIEW

With the help of methods for analyzing unstructured data, local hotspots for the COVID-19 virus were found.As a consequence of this, exhaustive research that focuses mostly on ethnic profiling based on zip code is being conducted in order to discover which local communities in the Central

Florida region are most significantly influenced by COVID-19 [1]. The individual AQI, total AQI, but also corresponding AQI classifications for six pollutants in Asia's four most badly affected locations were analyzed in order to determine the pollution impact and key contaminants both after and during COVID-19 lockdowns. This was done in order to determine how the lockdowns affected air pollution and key pollutants. The cities of Wuhan, Tokyo, and Mumbai, India, were identified as hotspot cities [2]. Telemedicine, disease monitoring, data intelligence, sickness modeling, and medication modeling are the five research topics that have been identified as a result of the thematic analysis of the solutions that have been investigated. Each topic is supported by a tool (or set of tools) that offers automation and helps users make decisions. In addition, researchers provide four reference designs that might potentially address recurrent challenges in the process of developing the systems. these systems of ontology-based COVID-19 analytics platforms [3]. According to the statistics, there are significantly higher levels of air quality in eastern India around clusters of coal-fired power stations. Based on the research results, India (Eastern) would have higher air pollution levels, which would establish it as a new hot spot for air quality with the biggest magnitudes [4]. Information from mobile phone use enables the evaluation of mobility behavior for an entire country on a moment-by-moment basis. In order to effectively deal with COVID-19 and any future pandemics, it is essential that more of these types of data be made available in an anonymous format, which is a point that we want to stress [5]. Assessing potential risk factors for infections as well as clinical symptoms among employees who tested positive for antibodies was a secondary objective [6]. The occurrence of COVID-19 was calculated by using the decided-to-invite swab (RT-PCR) tests that were provided in the app. The researchers have used spatially granular estimates to locate hotspots, which they defined as regions with rapidly expanding case numbers [7]. The authors have offered a software solution that exhibits the effectiveness of their approaches as well as implementing their methods themselves [8]. Researchers have explored the association between COVID-19 fatalities and a variety of air pollutants (using two independent temporal datasets), and they discovered excellent positive correlations with PM10 and AQI indicators [9]. The researchers conducted spatial analyses throughout Kazakhstan using geographic information systems such as QGIS and GeoDa. This allowed them to uncover COVID-19 risk clusters. After extensive installation and testing, the authors found that S-Nav provides reliable route suggestions in a manner that is almost identical to real-time. S-Nav, on the other hand, limits travel via red and orange zones to less than 2% of all trips while increasing the percentage of trips taken through green zones to over 100%. On the other hand, researchers see an 18% increase in travel lengths in comparison to the potentially hazardous paths that are the shortest [11]. The authors built a computer architecture that is Internet of Things (IoT) based and also integrates fog and cloud computing technologies. The authors have presented IoT-based strategies as a means of illness detection. These techniques make use of analytical tools and suggest an architecture, both of which make use of the big data idea for data collection and analysis [12]–[18]. The authors have used a variety of quality estimation methods in order to evaluate the effectiveness of the recorded data and images [19]–[21].

In this specific instance, the LSTM model was used to analyze the non-linear component of the data, while the ARIMA model was utilized to investigate the linear component of the data. Both models were used in conjunction with each other.The hybrid model combines the benefits of the ARIMA model with the LSTM model[22]. Intraday forecast updates are used to determine how accurate the SARIMA and SARIMAX models are when it comes to making predictions [23]. The authors provide a variety of different methods for assessing data [24]–[25]. The authors investigated and compared four different models (exponential smoothing, integrated autoregressive average, and seasonally adjusted moving average) to see which was the most effective at predicting. The SARIMAX model is also utilized to account for the influence of India's vacation seasons on the assessment of variations in the number of new cases. This is done by taking into account the number of people who are away from work during these times. Analysis of time series may be carried out in either a univariate or multivariate fashion by employing the ARIMAX model [26].

III. PROPOSED WORK

Many libraries have been used for the further processing of data, like NumPy, which is used for array manipulation in thedomain of linear algebra and the Fourier transform. Pandas isused for data processing and importing CSV files in the Python script.

Steps of proposed work

1. Importing dataset for API.
2. Since we are predicting total cases ('TT') column we don't need state data, hence dropping out state data.
3. Copying data into train_df Data frame so that our original data is untouched
4. Plotting current data.
5. Decomposing data to observe if there exists a seasonal trend.
6. Calling the function gives below result, where we can observe the huge gap between original data and mean, standard deviation
7. Also, the p value is 0.9778 which is not so good and hence, the output says, "The series is likely non-

stationary."

8. Here are various methods for making series stationary like log, differencing and so on.
9. Here we are using differencing, shift operator shifts the 'TT' column of df by 4 places and difference is taken.
10. Plotting the data after differencing we see the P value is reduced to 0.3427 which is quite good as compared to our previous value 0.9778
11. You can try different values in shift to reduce the p value (if possible, #try to choose one where number of observations used is MAX and p value is MIN)
12. Plotting autocorrelation and partial autocorrelation for both data (data before differencing and data after differencing)
13. We can see a recurring correlation exists in both ACF and PACF hence we should choose SARIMAX model which also deals with seasonality.

The matplotlib and seaborn libraries have been used for data visualization and statsmodel.api for ARIMA and SERIMAX.

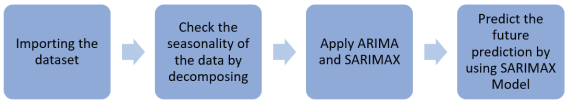
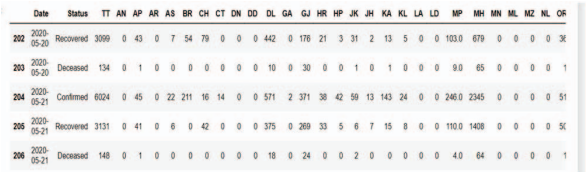


Fig. 1. Processes of proposed work.

Then we import data from covid website API. The data that we have imported has data for every state and union territory, and since we have to predict total cases, we have to drop out state data.



	Date	Status	TT	AN	AP	AR	AS	BR	CH	CT	DN	DD	DL	GA	GJ	HR	HP	JK	JH	KA	KL	LA	LD	MP	MH	MN	ML	MZ	NL	OR
202	2020-05-20	Recovered	3099	0	43	0	7	54	79	0	0	0	442	0	176	21	3	31	2	13	5	0	0	103.0	679	0	0	0	0	3(
203	2020-05-20	Deceased	134	0	1	0	0	0	0	0	0	0	10	0	30	0	0	1	0	1	0	0	0	9.0	65	0	0	0	0	1
204	2020-05-21	Confirmed	6024	0	45	0	22	211	16	14	0	0	571	2	371	38	42	59	13	143	24	0	0	246.0	2345	0	0	0	0	51
205	2020-05-21	Recovered	3131	0	41	0	6	0	42	0	0	0	375	0	269	33	5	6	7	15	8	0	0	110.0	1408	0	0	0	0	5(
206	2020-05-21	Deceased	148	0	1	0	0	0	0	0	0	0	18	0	24	0	0	2	0	0	0	0	0	4.0	64	0	0	0	0	1

Fig. 2. Sample data set

After coping the data to some other variable, we decompose the data to observe if there is a seasonal trend.

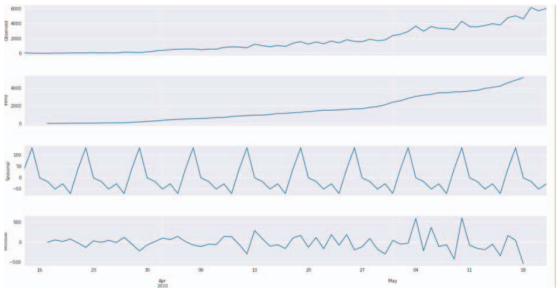


Fig. 3. Data set graph

IV. RESULT ANALYSIS

The "trend" graph has an upward trend, which shows that the data is not stationary. Time series forecasting can be done on stationary data, so we need to make it stationary.

To further check the data, we perform an ad hoc test. AD-Fuller stands for Augmented Dickey-Fuller Unit Root Test, which checks the mean and standard deviation and returns the p-value. The smaller the p-value, the more stationary the series.

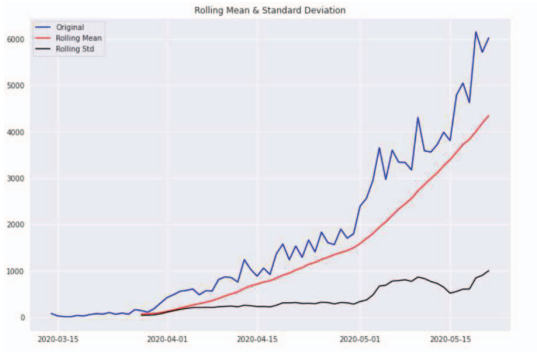


Fig. 4. Graph among original , mean and std.

After calling the function, we can observe the huge gap between the original data and the rolling mean and standard deviation, and in this data set, the p-value is 0.9778, which is not so good, so we can say that "The series is likely not stationary."

So, to make the series stationary, we can employ various techniques such as logging, differencing, and so on.

Here we are using differencing; the shift operation shifts the "TT" column by 4 places and the difference is taken.

Plotting the data after differencing, we see the p value is reduced to 0.3427, which is quite good as compared to our previous value of 0.9778.

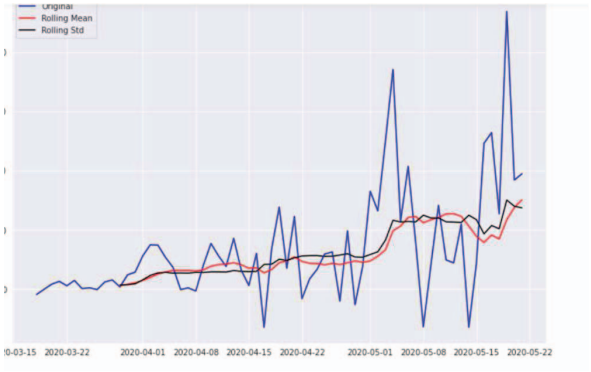


Fig. 5. plotting the data after differencing

Now we put the graphs of autocorrelation and partial autocorrelation for both the data sets (data before differencing and data after differencing).

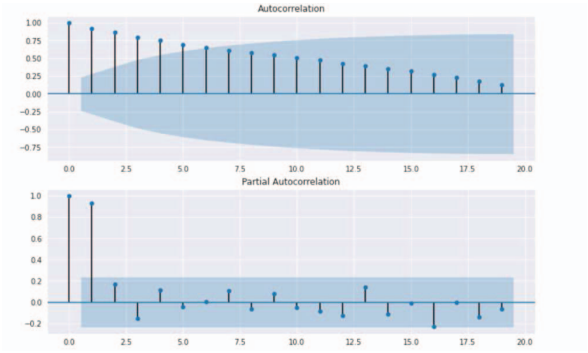


Fig. 6. graph of auto correlation

We can see a recurring correlation exists in both ACF andPACF; hence, we should choose the SARIMAX model, which also deals with seasonality.

So after using the SARIMAX model, we again plot the ACFand PACF graphs.

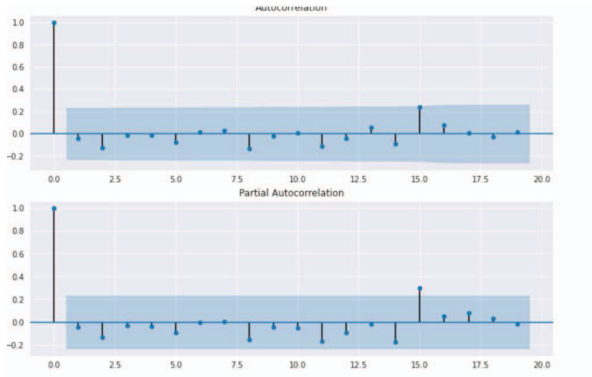


Fig. 7. Graph after using SARIMAX model

So here we simply see that there is an upward trend in the graph, which is written, and after that we start forecasting the data, and we plot a graph of original data and forecasted data.

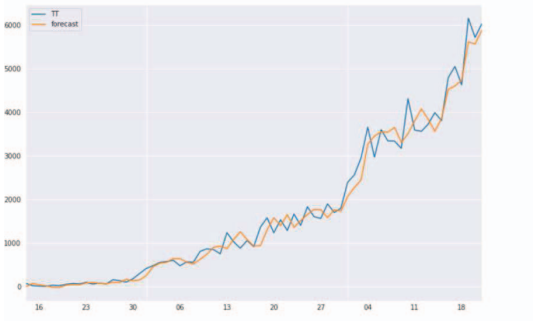
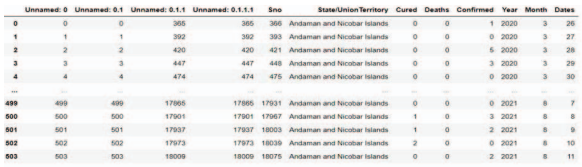


Fig. 8. Graph of original data and forecasted data

Here we can see that the forecasted data goes hand in handwith original data.



	Unnamed: 0	Unnamed: 0.1	Unnamed: 0.1.1	Unnamed: 0.1.1.1	Sno	State/UnionTerritory	Cured	Deaths	Confirmed	Year	Month	Dates
0	0	0	365	365	366	Andaman and Nicobar Islands	0	0	1	2020	3	26
1	1	1	392	392	393	Andaman and Nicobar Islands	0	0	0	2020	3	27
2	2	2	420	420	421	Andaman and Nicobar Islands	0	0	5	2020	3	28
3	3	3	447	447	448	Andaman and Nicobar Islands	0	0	3	2020	3	29
4	4	4	474	474	475	Andaman and Nicobar Islands	0	0	0	2020	3	30
...	...	...	...	...	...	...	...	...	...	...	...	...
499	499	499	17865	17865	17931	Andaman and Nicobar Islands	0	0	0	2021	8	7
500	500	500	17901	17901	17967	Andaman and Nicobar Islands	1	0	3	2021	8	8
501	501	501	17937	17937	18003	Andaman and Nicobar Islands	1	0	2	2021	8	9
502	502	502	17973	17973	18039	Andaman and Nicobar Islands	2	0	0	2021	8	10
503	503	503	18009	18009	18075	Andaman and Nicobar Islands	0	0	2	2021	8	11

Fig. 9. Confirm cases of disease

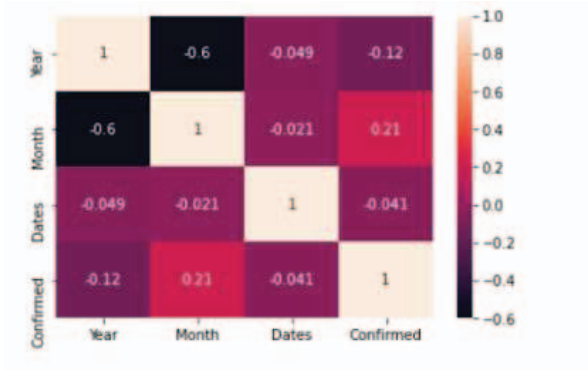


Fig. 10. Confusion matrix

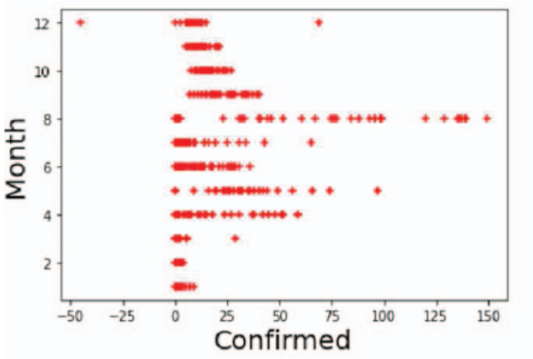


Fig. 11. Chart for confirm cases

	Date	TT	UN
64	2020-05-17	5049	0
65	2020-05-18	4628	0
66	2020-05-19	6154	814
67	2020-05-20	5716	282
68	2020-05-21	6024	307

Fig. 12. Table of covid cases

V. CONCLUSION

Hot-spot analysis is a sophisticated statistical method for epidemic analytics and visualization. It can provide public health policymakers real-time epidemic patterns and projections. hot-spot analysis has helped quickly identify high-risk clusters and adapt public health policies. In today's time we are very depressed with the Covid cases and the deaths due to that virus, more than 3.4 Cr peoplealone India is affected due to this deathly virus and 4.51L people died due to this virus. So our model which is based on ARIMA and SARIMAX models will predict the number of cases due to corona in the identified hotspots with high accuracy, so that people can take preventive measures and government and local authorities canmake future strategy to deal with it.

REFERENCES

[1] N. Joseph, S. Bernadin, D. Hodges, and P. Sekhar, "A Case Study on using Unstructured Data Analysis Methods to identify local Covid-19 Hotspots," in SoutheastCon 2021, 2021.

[2] M. Hu, Z. Chen, H. Cui, T. Wang, C. Zhang, and K. Yun, "Air pollution and critical air pollutant assessment during and after COVID-19 lockdowns: Evidence from pandemic hotspots in China, the Republic of Korea, Japan, and India," Atmos. Pollut. Res., vol. 12, no. 2, pp. 316–329, 2021.

[3] A. Ahmad, M. Bandara, M. Fahmideh, H. A. Proper, G. Guizzardi, and J. Soar, "An Overview of Ontologies and Tool Support for COVID-19 Analytics," in 2021 IEEE 25th International Enterprise Distributed Object Computing Workshop (EDOCW), 2021.

[4] B. Tyagi, G. Choudhury, N. K. Vissa, J. Singh, and M. Tesche, "Changing air pollution scenario during COVID-19: Redefining the hotspot regions over India," Environ. Pollut., vol. 271, no. 116354, p. 116354, 2021.

[5] G. Heiler et al., "Country-wide mobility changes observed using mobile phone data during COVID-19 pandemic," in 2020 IEEE International Conference on Big Data (Big Data), 2020..

[6] A. Hildebrandt, O. Hökelekli, L. Uflacker, H. Rudolf, and S. G. Gatermann, "COVID-19: Hotspot hospital?- seroprevalence of SARS-CoV-2 antibodies in hospital employees in a secondary care hospital network in Germany: Intermediate results of a prospective surveillance study," Int. J. Hyg. Environ. Health, vol. 235, no. 113771, p. 113771, 2021.

[7] T. Varsavsky et al., "Detecting COVID-19 infection hotspots in England using large-scale self-reported data from a mobile application: a prospective, observational study," Lancet Public Health, vol. 6, no. 1, pp. e21–e29, 2021.

[8] A. Singh and M. K. Hanawal, "Monitoring COVID hotspots using telecom data: Voronoi tessellations for marking buffer zones," in 2021 International Conference on COMmunication Systems & NETworkS (COMSNETS), 2021.

[9] H. R. Naqvi, G. Mutreja, A. Shakeel, and M. A. Siddiqui, "Spatio-temporal analysis of air quality and its relationship with major COVID-19 hotspot places in India," Remote Sens. Appl. Soc. Environ., vol. 22, no. 100473, p. 100473, 2021.

[10] S.-Y.-D. Zhou et al., "Discarded masks as hotspots of antibiotic resistance genes during COVID-19 pandemic," J. Hazard. Mater., vol. 425, no. 127774, p. 127774, 2022.

[11] T. Yabe, K. Tsubouchi, Y. Sekimoto, and S. V. Ukkusuri, "Early warning of COVID-19 hotspots using human mobility and web search query data," Comput. Environ. Urban Syst., vol. 92, no. 101747, p. 101747, 2022.

[12] K. Kumar, N. Kumar, and R. Shah, "Role of IoT to avoid spreading of COVID-19," International Journal of Intelligent Networks, vol. 1, pp. 32–35, 2020.

[13] N. Kumar, A. K. Dahiya, K. Kumar, and S. Tanwar, "Application of IoT in Agriculture," in 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2021.

[14] K. Kumar, G. Saini, N. Kumar, M. S. Kaiser, R. Kannan, and R. Shah, "Prediction of energy generation target of hydropower plants using artificial neural networks," Sustainable Developments by Artificial Intelligence and Machine Learning for Renewable Energies, pp. 309–320, 2022.

[15] B. Suneja, A. Negi, N. Kumar, and R. Bhardwaj, "Cloud-based tomato plant growth and health monitoring system using IOT," 2022 3rd International Conference on Intelligent Engineering and Management (ICIEM), 2022.

[16] K. Kumar, N. Kumar, A. Kumar, M. A. Mohammed, A. S. Al-Waisy, M. M. Jaber, N. K. Pandey, R. Shah, G. Saini, F. Eid, and M. N. Al-Andoli, "Identification of cardiac patients based on the medical conditions using machine learning models," Computational Intelligence and Neuroscience, vol. 2022, pp. 1–15, 2022.

[17] K. Kumar, A. Kumar, N. Kumar, M. A. Mohammed, A. S. Al-Waisy, M. M. Jaber, R. Shah, and M. N. Al-Andoli, "Dimensions of internet of things: Technological taxonomy architecture applications and open challenges—a systematic review," Wireless Communications and Mobile Computing, vol. 2022, pp. 1–23, 2022.

[18] N. Kumar, K. Kumar, and A. Kumar, "Application of internet of things in image processing," 2022 IEEE Delhi Section Conference (DELCON), 2022.

[19] N. Kumar, H. Shukla, and R. Tripathi, "Image restoration in noisy free images using fuzzy based median filtering and adaptive particle swarm optimization - richardson-lucy algorithm," International Journal of Intelligent Engineering and Systems, vol. 10, no. 4, pp. 50–59, 2017.

[20] K. K. Narendra Kumar and Anil Kumar Dahiya, "Image Restoration Using a Fuzzy-Based Median Filter and Modified Firefly Optimization Algorithm", Int. J. Adv. Sci. Technol., vol. 29, no. 4s, pp. 1471-1477, 2020.

[21] N. Kumar, A. K. Dahiya and K. Kumar, "Modified Median Filter for Image Denoising", Int. J. Adv. Sci. Technol., vol. 29, no. 4s, pp. 1495-1502, 2020.

[22] E. Dave, A. Leonardo, M. Jeanice, and N. Hanafiah, "Forecasting Indonesia exports using a hybrid model Arima-LSTM," Procedia Computer Science, vol. 179, pp. 480–487, 2021.

[23] M. Xie, C. Sandels, K. Zhu and L. Nordström, "A seasonal ARIMA model with exogenous variables for elspot electricity prices in Sweden," 2013 10th International Conference on the European Energy Market (EEM), 2013, pp. 1-4, doi: 10.1109/EEM.2013.6607293.

[24] R. Gothwal, S. Gupta, D. Gupta and A. K. Dahiya, "Color image segmentation algorithm based on RGB channels," Proceedings of 3rd International Conference on Reliability, Infocom Technologies and Optimization, 2014, pp. 1-5, doi: 10.1109/ICRITO.2014.7014669.

[25] M. Jain and A. Kumar, "Secure medical communication using sixteen rectangle substitution cipher and information-location dependent steganography," 2017 International Conference on Intelligent Communication and Computational Techniques (ICCT), 2017, pp. 189-193, doi: 10.1109/INTELCCT.2017.8324043.

[26] A. Jain, T. Sukhdeve, H. Gadia, S. P. Sahu and S. Verma, "COVID19 Prediction using Time Series Analysis," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021, pp. 1599-1606, doi: 10.1109/ICAIS50930.2021.9395877.

[27] A. Sharma and N. Kumar, "Encryption of text using fingerprints as input to various algorithms," Int. J. Sci. Res., 2014.

[28] R. P. T. H.S Shukla Narendra Kumar, "Image Restoration using modified binary particle Swarm Optimization Richardson-Lucy (MBSO-RL) algorithm," Int. J. Appl. Eng. Res., vol. 10, no. 22, pp. 43077–43081, 2015.

[29] Z. Allam, G. Dey, and D. Jones, "Artificial Intelligence (AI) provided early detection of the coronavirus (covid-19) in China and will influence future urban health policy internationally," AI, vol. 1, no. 2, pp. 156–165, 2020.