Exploring Innovation Research Methodologies in a Variety of Multidisciplinary Fields and Their Prospective Future Impact
February 2024

Predictive Modelling and Analysis of Infectious Disease (COVID-19) Cases in Indian States

[1] Manoj Kumar Singh, [2] Basudeo Singh Roohani
[1] Department of Computer Science & Engineering IMS Engineering College Ghaziabad,
[2]Department of Computer Science & Engineering IMS Engineering College Ghaziabad
[1]singhmanojmit@gmail.com
[2] bsroohani2007@gmail.com

Abstract- In this paper, our aim is to analyze data on the number of infected individuals in each Indian state and predict the number of infections for the coming year. By focusing on state-level analysis, we can shed light on the specific challenges faced by different regions and tailor interventions accordingly. To analyze and understand the spread of COVID-19 in Indian states: The objective is to analyze the available data on COVID-19 cases in different Indian states and gain a comprehensive understanding of the spread and dynamics of the virus. This involves examining the patterns of infection, identifying hotspot regions, and studying the factors contributing to the spread within each state. In this paper we are finding the state wise confirmed case, deaths cases, and cured cases.

Keywords: - COVID-19, Machine Learning, infectious diseases, ML-Algorithm

I. INTRODUCTION

India is not an exception to the global impact that the COVID-19 epidemic has had on nations. .Given its breadth and diversity, India confronts particular difficulties in the fight against the virus. Although the total number of infections in the nation has received a lot of attention, it is important to focus on particular states and conduct a more detailed analysis of the issue. Due to elements including population density, healthcare infrastructure, and socioeconomic situations, the disease's transmission and effects varied dramatically across different geographical areas. As a result, it is crucial to examine COVID-19 instances in each Indian state independently in order to fully comprehend the problem and provide workable solutions.

We analyse data on the number of sick people in each Indian state in this research study with the intention of forecasting the number of illnesses in 2023. We can shed light on the

particular issues encountered by various areas by concentrating on state-level analyses and designing actions appropriately. This tailored strategy takes into account India's diversity andmakes it possible to implement containment and mitigation measures that are more successful.

We use a comprehensive system to accomplish our goals. The logistic model, the exponential model, and the susceptible-infectious-susceptible (SIS) model are the three growth models we take into account. The danger associated with relying only on a model's predictions is reduced by the unique insights and perspectives that each model contributes to the research. We integrate the predictions from the exponential and logistic models using an ensemble technique, weighted according to the current trend in infections as shown by the model-free maximum daily infection rate (DIR) for the previous two weeks. This data-driven methodology improves the reliability and accuracy of our forecasts.

In our analysis, we interpret the results from all models alongside the recent DIR values for each state. This comprehensive assessment enables us to categorize the states into three distinct categories: severe, moderate, or controlled. Such categorization provides policymakers and public health authorities with valuable insights into the severity and trajectory of the pandemic in each state. It empowers them to make informed decisions regarding resource allocation, targeted interventions, and preparedness measures.

It is crucial to note, however, that the accuracy of our classification and projections rely on the accessibility and dependability of the data utilised, as well as the presumptions and constraints of the growth models used. Despite these obstacles, this research promotes evidence-based decision-making and provides insightful information about the COVID-19 situation at the state level in India.

We support continuing efforts by the Indian government and public health agencies to effectively manage and contain the pandemic by analysing the COVID-19 outbreak and forecasting infections till 2023. The results of this study will help develop measures to lessen the virus's effects and safeguard the population of Indians' health and wellbeing.

II. LITERATURE REVIEW

Author (Year)	Methodology Used	Result	Disease
Albahri et al. (2020	Three ML techniques were applied to the MERS-CoV dataset to identify the best classification model for binary class and multiclass labels	This study reviewed the state-of-the-art techniques for CoV prediction algorithms based on data mining and ML assessment.	Novel Coronavirus (COVID-19)
Alimadadi et al. (2020)	Personalized protective strategies, novel diagnostic approaches using machine learning algorithms.	Improve diagnostic speed and accuracy, develop novel effective therapeutic approaches, and potentially identify the mostsusceptible people based on personalized genetic and physiological characteristics	Coronavirus disease 2019 (COVID-19)
Kumar et al. (2020)	Machine learning-based classification of the extracted deep feature using ResNet152 with COVID-19 and Pneumonia patients on chest X-ray images.	The model is achieving an accuracy of 0.973 on Random Forest and 0.977 using XGBoost predictive classifiers.	coronavirus (COVID-19)
Lalmuanawma et al. (2020)	Machine Learning and Artificial Intelligent for tackling Covid-19 pandemic.	Development in AI and ML has significantly improved treatment, medication, screening, prediction, forecasting, contact tracing, and drug/vaccine development process for the Covid-19 pandemic and reduce the human intervention in medical practice.	Covid-19
Kukar et al. (2020)	machine learning predictive model for COVID-19 diagnosis	COVID-19 diagnosis is attainable using ML on data from routine blood tests	coronavirus disease (COVID-19)

Brinati et al. (2020)	Two machine learning classification models using hematochemical values from routine blood exams	study demonstrated the feasibility and clinical soundness of using blood tests analysis and machine learning as an alternative to RT-PCR for identifying COVID-19 positive patients	COVID-19

III. METHODOLOGY

This research employs a comprehensive methodology to analyze COVID-19 cases in Indian states and predict the number of infections for the year 2023. The methodology includes data collection, cleaning, exploratory data analysis, feature engineering, and the development of an ensemble of predictive models.

1. Data Collection: The first step involves gathering data on COVID-19 cases in Indian states. Multiple sources can be utilized, including official government websites, health department reports, and reputable databases. The data should encompass relevant variables such as the number of confirmed cases, active cases, recoveries, and deaths for each state over a specified period.

2. Data Cleaning: Once the data is collected, it undergoes a thorough cleaning process to ensure its accuracy and consistency. This involves removing any duplicates, correcting errors, addressing missing values, and standardizing the data format across different sources.

3. Exploratory Data Analysis (EDA): The cleaned data is then subjected to EDA techniques to gain insights into the patterns and trends of COVID-19 cases in Indian states. Descriptive statistics, data visualization, and time series analysis can be employed to identify any notable variations in the spread of the virus across different regions.

4. Feature Engineering: Feature engineering involves transforming the raw data into meaningful features that can enhance the predictive models' performance. This can include deriving new variables or aggregating existing ones to capture relevant aspects such as population density, healthcare infrastructure, and socioeconomic factors specific to each state.

5. Predictive Modeling: Three growth models are considered in this research: the logistic model, the exponential model, and the susceptible-infectious-susceptible (SIS) model. Each model provides unique insights into the dynamics of COVID-19 spread. Predictive models are developed using these growth models to forecast the number of infections for each state in 2023.

6. Ensemble Approach: To enhance the reliability and accuracy of predictions, an ensemble approach is employed. The ensemble combines the predictions from the logistic and exponential models, with weights assigned based on functions of the model-free maximum daily infection rate (DIR) observed over the last two weeks. This data-driven weighting strategy ensures that recent trends are given appropriate consideration in the ensemble predictions.

7. Interpretation and Categorization: The results from all models, along with the recent DIR values for each state, are jointly interpreted. This interpretation allows for the categorization of states into three distinct categories: severe, moderate, or controlled. This categorization aids in understanding the severity of the COVID-19 situation in each state and guides policymakers in making informed decisions regarding interventions and resource allocation.

It is important to acknowledge that the success of this methodology relies on the availability and reliability of the data, as well as the assumptions and limitations of the growth models employed. Rigorous validation techniques, such as cross-validation and sensitivity analysis, should be utilized to assess the models' performance and reliability of the predictions.

Overall, this comprehensive methodology enables the analysis and prediction of COVID-19 cases in Indian states, providing valuable insights into the localized dynamics of the pandemic. The ensemble approach and categorization of states contribute to evidence-based decision-making and aid in developing targeted interventions to mitigate the impact of the virus.

In the context of analyzing and predicting COVID-19 cases in Indian states, several machine learning algorithms can be considered. Let's discuss the applicability and characteristics of the following algorithms: Decision Tree Regressor, Logistic Regression, Linear Regression, Random Forest Classifier, and Gaussian NB.

1. Decision Tree Regressor: The Decision Tree Regressor algorithm is a supervised learning algorithm used for regression tasks. It builds a decision tree based on the provided input features and their corresponding target values. In the context of predicting COVID-19 cases, Decision Tree Regressor can be useful in estimating the number of infections based on relevant features such as population density, healthcare infrastructure, and socioeconomic factors. However, decision trees tend to over fit the training data, so caution should be exercised to prevent over-optimistic predictions.

2. Logistic Regression: Logistic regression is a supervised learning algorithm used for classification tasks. While primarily designed for binary classification, it can be extended for multiclass classification. In the context of COVID-19 analysis, Logistic Regression can be applied to categorize states into severity levels, such as severe, moderate, or controlled. By considering various features, logistic regression can provide probabilities or likelihoods of belonging to each category, aiding in decision-making and resource allocation.

3. Linear Regression: Linear regression is a supervised learning algorithm used for regression tasks. It models the relationship between the input features and the target variable by fitting a linear equation. In the case of COVID-19 analysis, Linear Regression can be utilized to predict the number of infections based on various features. However, it is important to note that linear regression assumes a linear relationship between the features and the target variable, which may not always hold true for complex phenomena such as disease spread.

4. Random Forest Classifier: The Random Forest Classifier algorithm is an ensemble learning method that combines multiple decision trees to perform classification. It can handle both binary and multiclass classification tasks. Random Forest Classifier can be beneficial in predicting the severity category of COVID-19 cases in Indian states by considering multiple features. The ensemble nature of random forests helps mitigate over fitting and provides robust predictions by aggregating the decisions of multiple trees.

5. Gaussian NB: Gaussian NB is a classification algorithm based on Bayes' theorem and the assumption of Gaussian (normal) distribution of features. It is commonly used for classification tasks, particularly when dealing with continuous-valued features. In the context of COVID-19 analysis, Gaussian NB can be employed to classify states based on

relevant features into severity categories. However, it assumes that the features are independent, which may not always hold true in complex scenarios.

When selecting the appropriate algorithm(s) for COVID-19 analysis, it is essential to consider the specific requirements and characteristics of the dataset. Factors such as the nature of the target variable, the availability and quality of features, and the desired interpretability of the results should be taken into account. It is also recommended to employ cross-validation techniques to evaluate and compare the performance of different algorithms, ensuring the reliability of the predictions.

In the proposed methodology, the specific algorithms used for growth modeling and ensemble prediction are not explicitly mentioned. However, based on the general application of the mentioned algorithms, Decision Tree Regressor and Random Forest Classifier could be suitable for growth modeling and severity categorization, respectively. Logistic Regression and Gaussian NB may be employed for severity classification tasks, while Linear Regression could be utilized for regression-based prediction of infection numbers. The specific choice of algorithms should be determined based on the characteristics of the data and the research objectives.

IV. RESULT DISCUSSION

DM (Data Mining) and ML are powerful techniques that can be used for the recognition and pre-findings of transmissible diseases such as Corona. Here are some ways in which they can be applied:

a) **Identification of patterns and risk factors:** When it comes to the transmission of infectious illnesses, DM (Data Mining) may be used to assist detect trends and risk factors. ML algorithms can examine massive volumes of data to spot patterns in things like age, pre-existing conditions, and geography that may indicate Corona virus infection.

b) **Early detection and diagnosis:** Predictive models built using ML may assist detect possible instances of infectious illnesses before they spread. Algorithms may be taught to monitor social media, news broadcasts, and other sources in real time in order to spot signs of impending epidemics.

c) **Contact tracing:** Potential transmission channels and hotspots may be identified by analysing contact tracing data using ML methods. With this information, public health

workers may better focus their actions and slow the disease's spread.

d) **Vaccine distribution:** DM (Data Mining) can be used to analyze demographic and geographic data to identify areas with the highest need for vaccines. This can help public health officials allocate resources more effectively and ensure that the most vulnerable populations are protected.

e) **Drug discovery:** ML can be used to analyse large datasets of medical research to identify potential treatments for transmissible diseases.

COVID-19 cases in various Indian states in order to get a thorough knowledge of the dynamics of the virus and its propagation. Examining infection trends, locating hotspot areas, and researching the elements influencing the spread within each state are all necessary for this:-

In [1]:
```
import pandas as pd
from matplotlib import pyplot as plt
import seaborn as sns
import datetime as dt
import numpy as np
covid=pd.read_csv("C:\\Users\\Sony\\Desktop\\python code\covid_19_india.csv",parse_dates=['Date'],dayfirst=True)
```

In [2]:
```
covid.head(10)
```

Out[2]:

	Sno	Date	Time	State/UnionTerritory	ConfirmedIndianNational	ConfirmedForeignNational	Cured	Deaths	Confirmed
0	1	2020-01-30	6:00 PM	Kerala	1	0	0	0	1
1	2	2020-01-31	6:00 PM	Kerala	1	0	0	0	1
2	3	2020-02-01	6:00 PM	Kerala	2	0	0	0	2
3	4	2020-02-02	6:00 PM	Kerala	3	0	0	0	3
4	5	2020-02-03	6:00 PM	Kerala	3	0	0	0	3
5	6	2020-02-04	6:00 PM	Kerala	3	0	0	0	3
6	7	2020-02-05	6:00 PM	Kerala	3	0	0	0	3
7	8	2020-02-06	6:00 PM	Kerala	3	0	0	0	3
8	9	2020-02-07	6:00 PM	Kerala	3	0	0	0	3
9	10	2020-02-08	6:00 PM	Kerala	3	0	0	0	3

```
In [3]: covid.tail(10)
```

Out[3]:

	Sno	Date	Time	State/UnionTerritory	ConfirmedIndianNational	ConfirmedForeignNational	Cured	Deaths	Confirmed
18100	18101	2021-08-11	8:00 AM	Puducherry	-	-	119115	1800	121766
18101	18102	2021-08-11	8:00 AM	Punjab	-	-	582791	16322	599573
18102	18103	2021-08-11	8:00 AM	Rajasthan	-	-	944700	8954	953851
18103	18104	2021-08-11	8:00 AM	Sikkim	-	-	25095	356	28018
18104	18105	2021-08-11	8:00 AM	Tamil Nadu	-	-	2524400	34367	2579130
18105	18106	2021-08-11	8:00 AM	Telangana	-	-	638410	3831	650353
18106	18107	2021-08-11	8:00 AM	Tripura	-	-	77811	773	80660
18107	18108	2021-08-11	8:00 AM	Uttarakhand	-	-	334650	7368	342462
18108	18109	2021-08-11	8:00 AM	Uttar Pradesh	-	-	1685492	22775	1708812
18109	18110	2021-08-11	8:00 AM	West Bengal	-	-	1506532	18252	1534999

```
In [11]: max_confirmed_cases=today.sort_values(by='confirmed',ascending=False)
```

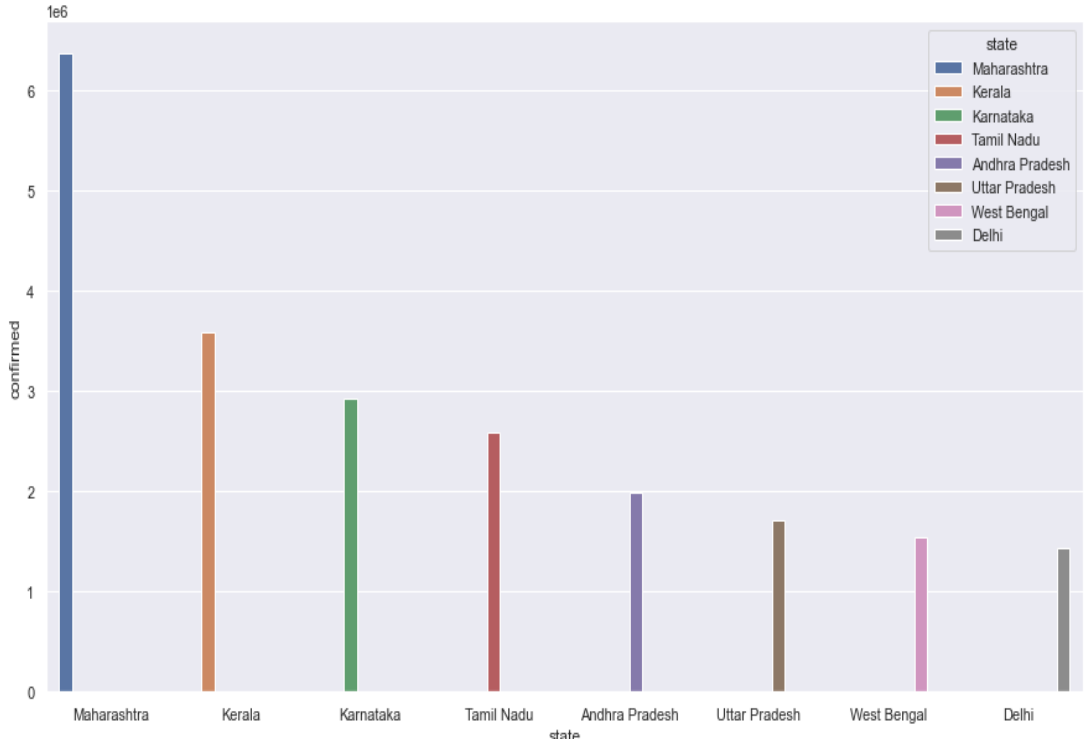
```
In [12]: max_confirmed_cases.head(10)
```

Out[12]:

	date	state	cured	deaths	confirmed
18094	2021-08-11	Maharashtra	6159676	134201	6363442
18090	2021-08-11	Kerala	3396184	18004	3586693
18089	2021-08-11	Karnataka	2861499	36848	2921049
18104	2021-08-11	Tamil Nadu	2524400	34367	2579130
18075	2021-08-11	Andhra Pradesh	1952736	13564	1985182
18108	2021-08-11	Uttar Pradesh	1685492	22775	1708812
18109	2021-08-11	West Bengal	1506532	18252	1534999
18082	2021-08-11	Delhi	1411280	25068	1436852
18080	2021-08-11	Chhattisgarh	988189	13544	1003356
18099	2021-08-11	Odisha	972710	6565	988997

Maximum Number of confirmed cases state wise

```
Out[14]: <function matplotlib.pyplot.show(close=None, block=None)>
```



Maximum Number of confirmed cases state wise Graph

State types according to degrees of severity: The classification of the states into severity categories, such as severe, moderate, or controlled, based on analysis and forecasts, is the Second goal. Based on the severity of the COVID-19 situation in each state, this categorization helps to prioritise resources, develop focused treatments, and provide healthcare facilities and assistance.

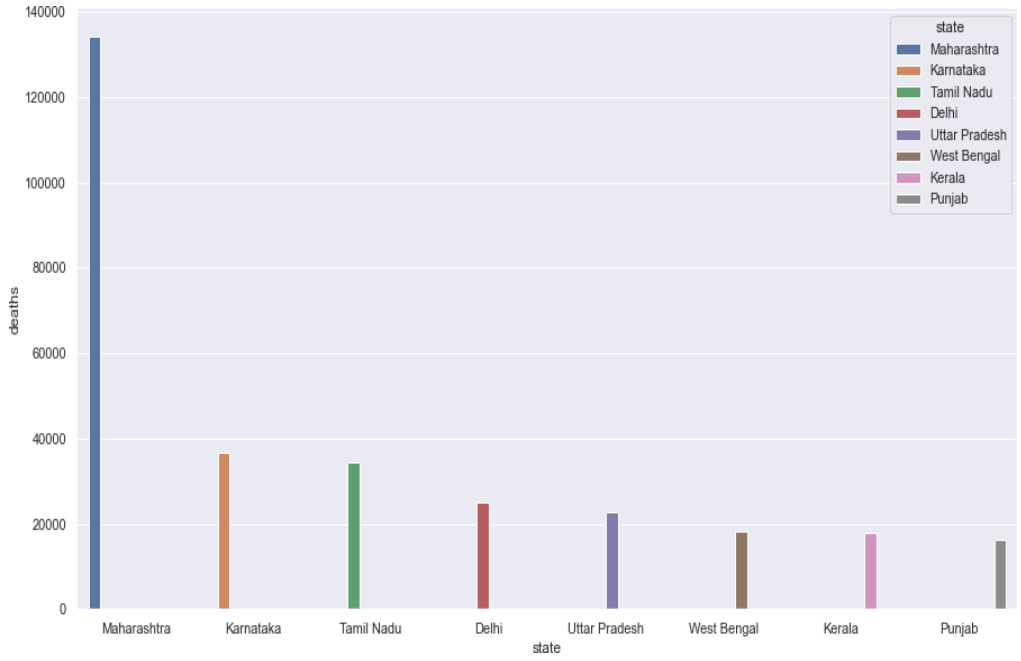
In [15]:
```
max_deaths_cases=today.sort_values(by='deaths',ascending=False)
```

In [16]:
```
max_deaths_cases.head(10)
```

Out[16]:

	date	state	cured	deaths	confirmed
18094	2021-08-11	Maharashtra	6159676	134201	6363442
18089	2021-08-11	Karnataka	2861499	36848	2921049
18104	2021-08-11	Tamil Nadu	2524400	34367	2579130
18082	2021-08-11	Delhi	1411280	25068	1436852
18108	2021-08-11	Uttar Pradesh	1685492	22775	1708812
18109	2021-08-11	West Bengal	1506532	18252	1534999
18090	2021-08-11	Kerala	3396184	18004	3586693
18101	2021-08-11	Punjab	582791	16322	599573
18075	2021-08-11	Andhra Pradesh	1952736	13564	1985182
18080	2021-08-11	Chhattisgarh	988189	13544	1003356

State wise Degree of severity



State wise Degree of severity Graph

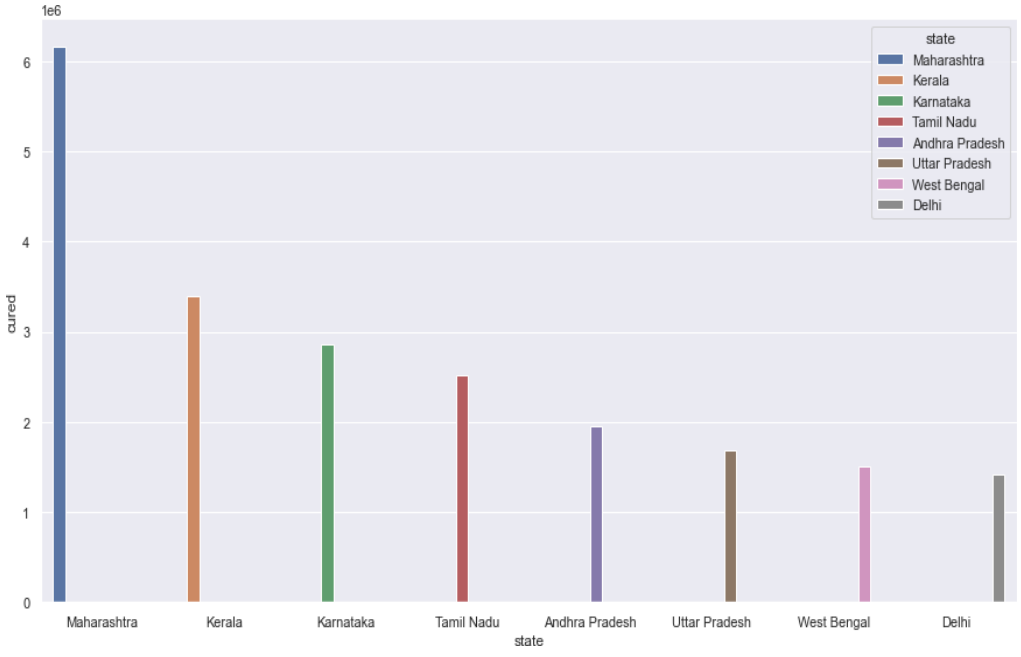
```
In [19]: max_cured_cases=today.sort_values(by='cured',ascending=False)

In [20]: max_cured_cases.head(10)
```

Out[20]:

	date	state	cured	deaths	confirmed
18094	2021-08-11	Maharashtra	6159676	134201	6363442
18090	2021-08-11	Kerala	3396184	18004	3586693
18089	2021-08-11	Karnataka	2861499	36848	2921049
18104	2021-08-11	Tamil Nadu	2524400	34367	2579130
18075	2021-08-11	Andhra Pradesh	1952736	13564	1985182
18108	2021-08-11	Uttar Pradesh	1685492	22775	1708812
18109	2021-08-11	West Bengal	1506532	18252	1534999
18082	2021-08-11	Delhi	1411280	25068	1436852
18080	2021-08-11	Chhattisgarh	988189	13544	1003356
18099	2021-08-11	Odisha	972710	6565	988997

State wise cured cases



State wise cured cases Graph

V. CONCLUSION

The analysis and prediction of COVID-19 cases in Indian states using a comprehensive methodology have provided valuable insights into the spread of the virus and its potential trajectory in the year 2023. By considering multiple growth models and recent trends in

infections, the methodology has allowed for a comprehensive assessment of the severity and dynamics of the pandemic at the state level.

In conclusion, the analysis and prediction of COVID-19 cases in Indian states using the proposed methodology provide valuable insights for managing the ongoing pandemic. The categorization of states, along with the predictions of infection numbers, offers guidance for policymakers, public health authorities, and researchers in implementing targeted interventions, allocating resources effectively, and planning for the future. Continuous monitoring and refinement of the analysis are crucial to adapt to the evolving nature of the pandemic and ensure effective management of COVID-19 in India

VI. REFERENCES

1. A. Kumar, R. K. Sharma, and others (2021). Modelling instances of COVID-19 using machine learning techniques in India. 10(1), 35–43; Health Policy and Technology.
2. Aduragba, O.T. and Cristea, A.I., 2019, July. Research on Prediction of Infectious Diseases, their spread via Social Media and their link to Education. In *Proceedings of the 4th International Conference on Information and Education Innovations* (pp. 38-42).
3. Alfred, R. and Obit, J.H., 2021. The roles of ML methods in limiting the spread of deadly diseases: A systematic review. *Heliyon*, *7*(6), p.e07371.
4. Alimadadi, A., Aryal, S., Manandhar, I., Munroe, P.B., Joe, B. and Cheng, X., 2020. Artificial intelligence and ML to fight COVID-19. *Physiological genomics*, *52*(4), pp.200-202.
5. Al-Turjman, F., 2021. AI-powered cloud for COVID-19 and other infectious disease diagnosis. *Personal and Ubiquitous Computing*, pp.1-4.
6. Amar, L.A., Taha, A.A. and Mohamed, M.Y., 2020. Prediction of the final size for COVID-19 epidemic using ML: A case study of Egypt. *Infectious Disease Modelling*, *5*, pp.622-634.
7. Anand, A., Sandaraju, H., Parande, M., & Srivastava, S. (2020). COVID-19 in India: A review of current knowledge. Indian Journal of Medical Research, 151(2), 147-159.
8. Arji, G., Ahmadi, H., Nilashi, M., Rashid, T.A., Ahmed, O.H., Aljojo, N. and Zainol, A., 2019. Fuzzy logic approach for infectious disease diagnosis: A methodical evaluation, literature and classification. *Biocybernetics and biomedical engineering*, *39*(4), pp.937-955.
9. Arpaci, I., Huang, S., Al-Emran, M., Al-Kabi, M.N. and Peng, M., 2021. Predicting the COVID-19 infection with fourteen clinical features using ML classification algorithms. *Multimedia Tools and Applications*, *80*, pp.11943-11957.

10. Ayatollahi, H., Gholamhosseini, L. and Salehi, M., 2019. Predicting coronary artery disease: a comparison between two DM (Data Mining) algorithms. *BMC public health*, *19*(1), pp.1-9.
11. Ayris, D., Horbury, K., Williams, B., Blackney, M., See, C.S.H., Imtiaz, M. and Shah, S.A.A., 2020. Deep Learning Models for Early Recognition and pre-findings of the spread of Novel Coronavirus (COVID-19). *arXiv preprint arXiv:2008.01170*.
12. Baguelin, M., Medley, G.F., Nightingale, E.S., O'Reilly, K.M., Rees, E.M., Waterlow, N.R. and Wagner, M., 2020. Tooling-up for infectious disease transmission modelling. *Epidemics*, *32*, p.100395.
13. Barman, R.K., Mukhopadhyay, A., Maulik, U. and Das, S., 2019. Identification of infectious disease-associated host genes using ML-Techniques. *BMC bioinformatics*, *20*, pp.1-12.
14. Bhadula, S., Sharma, S., Juyal, P. and Kulshrestha, C., 2019. ML algorithms based skin disease detection. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, *9*(2), pp.4044-4049.
15. Bhardwaj, T. and Somvanshi, P., 2019. ML toward infectious disease treatment. In *Machine Intelligence and Signal Analysis* (pp. 683-693). Springer Singapore.
16. Bogu, G.K. and Snyder, M.P., 2021. Deep learning-based detection of COVID-19 using wearables data. *MedRxiv*, pp.2021-01.
17. Borkenhagen, L.K., Allen, M.W. and Runstadler, J.A., 2021. Influenza virus genotype to phenotype predictions through ML: a systematic review: Computational Prediction of Influenza Phenotype. *Emerging Microbes & Infections*, *10*(1), pp.1896-1907.
18. Brinati, D., Campagner, A., Ferrari, D., Locatelli, M., Banfi, G. and Cabitza, F., 2020. Detection of COVID-19 infection from routine blood exams with ML: a feasibility study. *Journal of medical systems*, *44*, pp.1-12.
19. Bruyndonckx, R., Coenen, S., Adriaenssens, N., Versporten, A., Monnet, D.L., Goossens, H., Molenberghs, G., Weist, K. and Hens, N., 2021. Analysing the trend over time of antibiotic consumption in the community: a tutorial on the detection of common change-points. *Journal of Antimicrobial Chemotherapy*, *76*(Supplement_2), pp.ii79-ii85.
20. Calderón-Gómez, H., Mendoza-Pittí, L., Vargas-Lombardo, M., Gómez-Pulido, J.M., Castillo-Sequera, J.L., Sanz-Moreno, J. and Sención, G., 2020. Telemonitoring system for infectious disease prediction in elderly people based on a novel microservice architecture. *IEEE access*, *8*, pp.118340-118354.

21. Callejon-Leblic, M.A., Moreno-Luna, R., Del Cuvillo, A., Reyes-Tejero, I.M., Garcia-Villaran, M.A., Santos-Pena, M., Maza-Solano, J.M., Martín-Jimenez, D.I., Palacios-Garcia, J.M., Fernandez-Velez, C. and Gonzalez-Garcia, J., 2021. Loss of smell and taste can accurately predict COVID-19 infection: a machine-learning approach. *Journal of Clinical Medicine*, *10*(4), p.570.

22. Champredon, D. and Dushoff, J., 2015. Intrinsic and realized generation intervals in infectious-disease transmission. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1821), p.20152026.

23. Chen, J., Lu, Y., Bai, J., & Xu, H. (2020). Predicting the epidemic dynamics of COVID-19 in China and the US with a data-driven stochastic model. Chaos, Solitons & Fractals, 139, 110055.

******