Epidemo: A Machine Learning Regression-Based Model

Joel Lee George
Department of Computer Science and Engineering
Amal Jyothi College of Engineering
Kanjirappally,Kerala
joelleegeorge2024@cs.ajce.in

Karthik S Kumar
Department of Computer Science and Engineering
Amal Jyothi College of Engineering
Kanjirappally, Kerala
karthikskumar2024@cs.ajce.in

Riya Merce Thomas
Department of Computer Science and Engineering
Amal Jyothi College of Engineering
Kanjirappally, Kerala
riyamercethomas2024@cs.ajce.in

Roshan Roy Varghese
Department of Computer Science and Engineering
Amal Jyothi College of Engineering
Kanjirappally, Kerala
roshanroyvarghese2024@cs.ajce.in

Simy Mary Kurian
Department of Computer Science and Engineering
Amal Jyothi College of Engineering
Kanjirappally, Kerala
simymarykurian@amaljyothi.ac.in

***Abstract*— In recent years, the world has witnessed the devastating consequences of disease outbreaks, highlighting the urgent need for effective epidemic management. An epidemic signifies the rapid transmission of illness to a substantial portion of a population within a short timeframe. The proposed system offers a proactive approach to this challenge by leveraging advanced Machine Learning (ML) regression tools. By analyzing diverse data sources such as historical disease trends, environmental conditions, and human behaviors, the system predicts the onset and spread of diseases, providing crucial early warnings for public health authorities and communities. Through timely implementation of preventive measures informed by these forecasts, authorities can mitigate the impact of epidemics, safeguard public health, and alleviate strain on healthcare systems. This proactive strategy underscores the importance of early intervention and data-driven approaches in combating and controlling disease outbreaks.**

I. INRODUCTION

In an increasingly interconnected and globalized world, the emergence of epidemics and infectious diseases poses a constant threat to public health and global stability. The need for a proactive, data-driven solution to predict and preempt these outbreaks has never been more urgent. This is where our proposed system steps in.

Our groundbreaking system leverages the power of advanced Machine Learning (ML) regression tools to revolutionize the way we approach epidemic prevention. By harnessing the vast volumes of data available, we are able to forecast when and how rapidly a disease is likely to spread, providing vital insights that are instrumental in preventing and mitigating the impact of epidemics.

Epidemics, like the recent pandemics, have demonstrated their devastating potential, not only in terms of public health but also in the profound economic, social, and political consequences

they bring. With our system, we aim to shift the paradigm from reactive crisis management to proactive epidemic prevention.

By accurately predicting the trajectory of disease outbreaks, our system empowers governments, healthcare agencies, and communities to implement timely and targeted preventive measures. These interventions can include vaccination campaigns, quarantine strategies, travel restrictions, and public health awareness initiatives. Ultimately, the objective is to curtail the spread of diseases, save lives, and safeguard the global community against the devastating impact of epidemics.

Our system represents a vital step towards a safer, more resilient world, where science, technology, and data converge to protect humanity from the unpredictable and often catastrophic nature of epidemics. With this powerful tool at our disposal, we are better equipped to safeguard public health and the well-being of future generations.

II. RELATED WORK

Over the past decade, there has been a notable rise in the occurrence of severe epidemics. The emergence of COVID-19 has underscored the critical importance of epidemic control measures. Enhancing our capacity to forecast the likelihood of these outbreaks can significantly enhance our efforts in preventing and managing them. It has become evident that the primary objective in this battle is the mitigation of disease transmission. Saumya Yashmohini Sahai et al. proposed a methodology used in nowcasting with the help of random forest algorithm to predict Covid-19. This model was easy to understand and make use due to it's simplicity [3] but it was less effective for long-term forecasting and trend analysis. Using an enriched model of random forest called RF EP (EP for Epidemiological Prediction) Patrick Loola Bokonda et al. proposed an epidemic prediction system. The authors found that RF EP gave a better performance in recall metric than the default RF and other traditional ML models [1]. But this model still had issues when compared to non-traditional methods.

Soudeh Ghafouri-Fard et al. devised a methodology employing three machine learning (ML) algorithms: Artificial Neural Network (ANN), Bidirectional Long Short-Term Memory (LSTM), and Auto-Regressive Integrated Moving Average (ARIMA). This approach utilized machine learning and deep learning techniques for trend estimation, effectively addressing the nonlinear and intricate nature of the disease. [7]However, the predictions were susceptible to uncertainty and necessitated access to high-quality data and computational resources.Sangwon Chae et al. employed deep learning algorithms, optimizing their parameters while considering vast datasets, including social media data, to predict infectious diseases[12]. They compared the performance of Deep Neural Network (DNN) and Long Short-Term Memory (LSTM) models with that of the Autoregressive Integrated Moving Average (ARIMA) model when forecasting three infectious diseases one week ahead. Their findings indicated that the DNN and LSTM models outperformed ARIMA. Although this investigation uncovered distinctive features of the DNN and LSTM models, the search query used was relatively outdated.S. Grampurohit et al. used three data mining algorithms- Decision tree, Random Forest and Naïve Bayes to analyze medical records and predict the most likely disease based on the symptoms provided by the patient. This model had very efficient data analysis[5] but had a limited symptom selection and was not flexible. J. Tetteroo et al. customized the AutoML framework auto-sklearn for time series forecasting purposes, presenting two variations for multi-step COVID-19 forecasting ahead. These were termed as (a) multi-output and (b) repeated single output forecasting[15]. It required significant computational resources and could sometimes lead to over-fitting.

Tianyu Zhang et al. proposed using AutoML and BrewAI to predict covid-19 which enhanced the accessibility of ML models. It was found that this model could be used by even non-ML experts whereas the traditional models needed expertise[2].This method requires a suitable AutoML platform which is a bit hard to achieve.

Martuza Ahamad et al. developed a model that employed supervised machine learning algorithms to find the affected individuals and to isolate them. Among the five algorithms used XGBoost and Gradient Boosting were found to be the best[4].Though the model was quiet successful in identifying patients with Covid-19 methodology used here showed a performance variability in different age groups. Yue Gao et al. proposed an ensemble model by making use of four machine learning methods that are Logistic Regression, Support Vector Machine, Gradient Boosted Decision Tree, and Neural Network[6]. Though this showed a better accuracy when compared to some of the other methods it is highly-complex and reduces the ease of use.

It sought to investigate automated COVID-19 detection through the utilization of machine learning techniques. Their objective was to develop an intelligent web application by training and assessing various classifiers. These included logistic regression, random forest, decision tree, k-nearest neighbor, support vector machine (SVM), ensemble models (adaptive boosting and extreme gradient boosting), as well as deep learning techniques such as artificial neural network, convolutional neural network, and long short-term memory[8]. Yet due to it's complex nature, it required diverse skill sets and collaboration of different experts.

Goodman-Meza D et al. aimed to build and analyze a machine learning algorithm to diagnose COVID-19 in the inpatient

setting[9].The algorithm relied on fundamental demographic and laboratory characteristics to function as a screening tool in hospitals lacking adequate testing resources. While the algorithm improves COVID-19 diagnosis in settings with limited resources, its data may not encompass all pertinent patient factors.

Zoabi Y. et al. [10] devised a machine-learning method that utilized data from 51,831 tested individuals, of whom 4769 were confirmed COVID-19 cases. They constructed a model capable of identifying COVID-19 cases by leveraging straightforward features obtained through basic inquiries. Remarkably, the model accurately predicted COVID-19 test outcomes using merely eight binary features: gender, age under 60 years, documented contact with a confirmed case, and the manifestation of five initial clinical symptoms. But since the model was based on data-set which included self reporting, there was a high chance of the symptoms being misinterpreted which possibly could lower the accuracy rate of the model.

S.Shinde et al. aimed to investigate the spread of outbreaks in villages and suburban areas where medical resources may be scarce. They sought to develop a machine learning model to predict epidemic dynamics and identify areas most likely to experience future outbreaks[11]. Recognizing the significance of factors that subtly influence disease epidemic patterns, the method incorporated considerations of climate, geography, and population distribution in the affected regions. The methodology discussed in the paper involved various approaches and models for outbreak prediction and data analysis related to infectious diseases. These methodologies include Epidemiological Models(SEIR) and Specific Disease Outbreak Prediction Models(Bayesian). Yet due to it's performance variability and limited availability of the data, it might have provided inaccurate predictions.

Sanzida Solayman et al. [13] conducted research focused on automatic COVID-19 detection using machine learning techniques to construct an intelligent web application. The dataset underwent preprocessing steps including dropping null values, feature engineering, and synthetic oversampling (SMOTE) techniques. Various classifiers were trained and evaluated, including logistic regression, random forest, decision tree, k-nearest neighbor, support vector machine (SVM), ensemble models (adaptive boosting and extreme gradient boosting), as well as deep learning techniques such as artificial neural network, convolutional neural network, and long short-term memory. Additionally, Explainable AI utilizing the LIME framework was employed to interpret prediction results. But it needed highly accurate data and required diverse skill set and collaboration due to the system's high complexity.

Seyed Ali Rakhshan et al. devised a recurrent SEIRS compartmental model to forecast the recurrence of disease outbreaks. Subsequently, they integrated machine learning models to enhance the paper's methodology in data analysis, particularly focusing on prediction scenarios. MLP, RBF, LSTM, ANFIS, and GRNN were employed for the training and evaluation of COVID-19. The outcomes were then juxtaposed with those obtained from the recurrent dynamical system in both the fitting process and prediction scenarios.[14]. Yet due to the high complexity of the model,limited data quality and availability can compromise prediction. It was also ineffective in providing a long-term forecast.

M. S. Kaiser et al. [16] introduced a mobile app-based intelligent portable healthcare (pHealth) tool named i WorkSafe. This tool aims to aid industries in identifying potential COVID-19 infection suspects among their employees who may require primary care. The app incorporates a fuzzy neural network model that amalgamates data on employees' health statuses from the industry's database, proximity and contact tracing information from mobile devices, and self-reported COVID-19 test data by users. By utilizing Bluetooth low energy sensing technology and employing K Nearest Neighbor and K-means techniques, the app can track users' proximity and trace their contact with other employees. The main limitation of this study was to utilize an open-source dataset comprising patient data from a specific region

III. PROPOSED METHODOLOGY

The proposed methodology presents a systematic and comprehensive approach to data pre-processing, analysis, and machine learning model development. Beginning with meticulous data cleansing and transformation, it ensures the dataset's accuracy and alignment with modeling assumptions. The subsequent exploration and correlation studies unveil crucial insights into data characteristics and relationships. The division of the dataset into training and testing sets enables robust model training, with a focus on parameter adjustment and fine-tuning for optimal performance. Rigorous testing and evaluation on unseen data, accompanied by performance metrics and comparative analysis, guide decision-making, ultimately determining the suitability of the proposed model for addressing research objectives. This methodology forms a critical framework for data-driven decision-making and model deployment.

Choice of epidemic:

Choosing an epidemic for experimentation involves a systematic methodology to ensure that the selected case is representative, relevant, and conducive to meaningful analysis. Selecting an epidemic for experimentation necessitates a systematic approach to ensure the chosen case is both representative and relevant. This comprehensive process ensures the chosen epidemic aligns with research goals and

provides a suitable foundation for meaningful and insightful analysis.

Choice of Dataset:

After selecting the epidemic of interest, the next crucial step is choosing a dataset from the available literature. The dataset serves as the empirical foundation for testing the proposed model. This selection involves careful consideration of factors such as data completeness, quality, temporal and geographic scope, demographic information, and relevance to research questions. The chosen dataset should align with the study's objectives and provide a comprehensive and reliable basis for evaluating the proposed model's effectiveness in addressing the dynamics of the selected epidemic.

Selection of ML models:

To validate the effectiveness of our model, it is imperative to engage in a comparative analysis by selecting and evaluating it alongside other machine learning models. This approach allows for a robust assessment of its performance in relation to existing benchmarks. By comparing against established models, we gain insights into the relative strengths and weaknesses of our proposed solution, aiding in the determination of its efficacy and suitability for addressing the specific challenges posed by the chosen epidemic. This process allows us to identify strengths and weaknesses relative to existing solutions, providing valuable insights into its potential application in addressing the challenges posed by the chosen epidemic.

ML Enriched Model:

The models used here include Gradient Boosting and XGBoost algorithms as they have a history of producing accurate predictions. Gradient Boosting is an ensemble learning technique used for both regression and classification tasks. It builds a strong predictive model by combining the predictions of multiple weak models, typically decision trees. It works iteratively , with each tree addressing the errors of the previous ones. In each iteration, the model focuses on the residual errors, adjusting its parameters to minimize them. This process continues until a predefined number of trees are built. Gradient Boosting is known for its high predictive accuracy and ability to handle complex relationships in data, making it a powerful and widely used machine learning algorithm.

XGBoost, or Extreme Gradient Boosting, is an advanced implementation of the gradient boosting algorithm designed for speed and performance. Widely utilized in machine learning competitions and diverse applications, XGBoost excels in predictive accuracy and efficiency. It incorporates regularization techniques to prevent overfitting, utilizes a unique "tree pruning" algorithm for enhanced computational speed, and supports parallel computing. XGBoost is versatile, handling regression, classification, and ranking tasks. Its ability to capture complex relationships in data, handle missing values, and provide feature importance rankings makes it a popular choice for data scientists aiming for high-performance predictive models in various domains.

Data Pre-processing and Analysis:

Data Cleansing:

This step involves identifying and handling missing data, outliers, and errors in the dataset. Cleaning the data ensures that the subsequent analysis and modeling are based on accurate and reliable information.

Data Transformation: Data may need to be transformed to meet the assumptions of the chosen modeling techniques. This can include scaling, normalization, or feature engineering to create new variables.

Data Exploration: This is an essential step for understanding the characteristics of the data. It includes data visualization, summary statistics, and exploratory data analysis (EDA) to uncover patterns, trends, and potential relationships within the data.

Correlation Studies: Correlation analysis helps identify relationships between variables. It's used to determine which variables are strongly related to the target variable or to each other. This information can guide feature selection and model building.

Dataset Splitting:

Training and Testing Sets: The dataset is split into two subsets: a training set and a testing set. The training set is used to build and optimize the models, while the testing set is reserved for evaluating their performance. This separation helps assess how well the models generalize to unseen data.

Model Training:

Utilizing Training Data: In this phase, machine learning models, including the proposed model, are trained using the training dataset. Training involves adjusting model parameters to minimize the difference between predicted and actual outcomes, a process known as model fitting.

Fine-tuning: Models may be fine-tuned by optimizing hyperparameters to improve their performance. This can involve techniques like cross-validation to find the best hyperparameter values.

Model Training:

Model training is a critical phase where machine learning models, including the proposed one, are fed with the training dataset to learn patterns and relationships. During this process, the model adjusts its parameters to minimize the difference between predicted and actual outcomes, a concept known as model fitting. Successful training is essential for the model's ability to make accurate predictions on new, unseen data, and it

forms the foundation for subsequent fine-tuning and evaluation stages in the modeling process.

Model Testing:

Testing on the Test Dataset: The trained models, including the proposed one, are applied to the test dataset to make predictions. This phase is crucial for assessing how well the models generalize to new, unseen data. It helps detect issues like overfitting (when a model fits the training data too closely) and ensures that the models perform as expected. In the testing phase, models, including the proposed one, are utilized to predict outcomes on a separate test dataset. This critical step evaluates the generalizability of the models to unseen data, detecting potential overfitting issues. By assessing performance on new data, the testing phase ensures that models are not merely memorizing the training set but demonstrating a robust ability to make accurate predictions in real-world scenarios, thus affirming their reliability and effectiveness.

Model Evaluation:

Performance Metrics: Models are evaluated using predefined performance metrics such as accuracy, precision, recall, F1-score, or mean squared error, depending on the nature of the problem (classification, regression, etc.).

Comparative Analysis: The results of all models, including the proposed solution, are compared based on the chosen performance metrics. This analysis helps determine which model best addresses the research objectives and is the most suitable for deployment.

Decision-Making: The evaluation results guide decision-making on whether to adopt the proposed model or opt for an alternative. It also informs potential improvements or further iterations of the modeling process.

These methodologies are integral to the process of data analysis and machine learning model development, ensuring that the chosen model is well-prepared, thoroughly tested, and evaluated for its effectiveness in addressing the research problem or objectives.

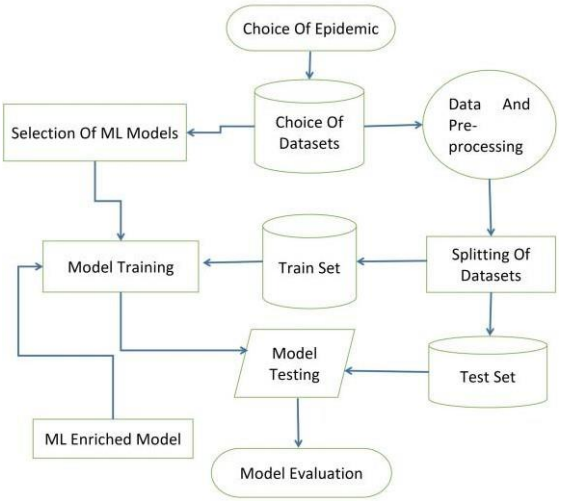

Fig1: Architecture of model

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

All datasets utilized in this project originate from the Git repository. This source serves as the primary reservoir for data acquisition, ensuring consistency and reliability in the datasets employed for analysis and model development. Overall, relying on the Git repository for dataset acquisition ensures a robust foundation for the project's objectives and contributes to the establishment of a reliable data-driven framework.

In this process, we examine the gathered data to extract any relevant insights that can provide meaningful interpretations. We analyze the primary factors that could significantly influence the occurrence frequency of various diseases.

For COVID-19, these factors include difficulty in breathing and a decrease in SpO2 levels, while for Dengue, they involve decreased platelet counts and blood pressure. As for the "others" category, it encompasses three diseases sharing common symptoms, such as fever, joint pains, yellowing of skin and eyes and much more.

For prediction purposes, we constructed a regression model to estimate the likely number of cases. Employing a gradient boosting classifier, we identified the best-performing model for classification and regression, aiming for more robust conclusions and results.We utilized the Flask API to deploy the model, while the front-end was developed entirely using HTML, CSS, React, and JavaScript.

The website's functionality is as follows: Users need to log in using their Google credentials to access prediction services. Subsequently, they can navigate to the "Take Action" page where they choose from three options:

i) Predicting whether a person has contracted COVID-19.

ii) Predicting whether a person has contracted Dengue.

iii) Predicting the likelihood of a person with certain symptoms being diagnosed with a disease.

The model forecasts the likelihood of an outbreak based on the number of cases.

The prediction models were constructed using a gradient boosting classifier and evaluated based on a dataset split of 75:25 for training and testing. The accuracy scores, measured using the 'accuracy-score' function from scikit-learn's metrics module, were expressed as percentages. The parameters for the COVID-19 and Dengue prediction models were selected to encompass the major symptoms specific to each disease. Additionally, along with symptoms the infection model incorporated Location as a primary parameter.

The COVID-19 prediction model showcased a notable accuracy rate of 98.9 percent, underscoring its effectiveness in accurately identifying cases within the dataset. This high level of accuracy suggests that the model is proficient in distinguishing between COVID-19 positive and negative cases, making it a valuable tool for diagnosis and prognosis in healthcare settings.

In contrast, despite utilizing a dataset comprising 1500 data values, the Dengue prediction model yielded a comparatively lower accuracy rate of 56.52 percent. This suggests that while the model may have some capability in predicting Dengue cases, its performance is not as reliable or consistent as desired. Further refinement or exploration of additional features may be necessary to improve the Dengue prediction model's accuracy and efficacy.

Similarly, the infection model, also trained on 1500 data values, exhibited an impressive accuracy rate of approximately 97.3 percent. This suggests that the model is highly effective in predicting infections based on the provided data. The notable accuracy rate indicates that the model is capable of accurately classifying instances of infection, highlighting its potential utility in disease surveillance and management efforts.

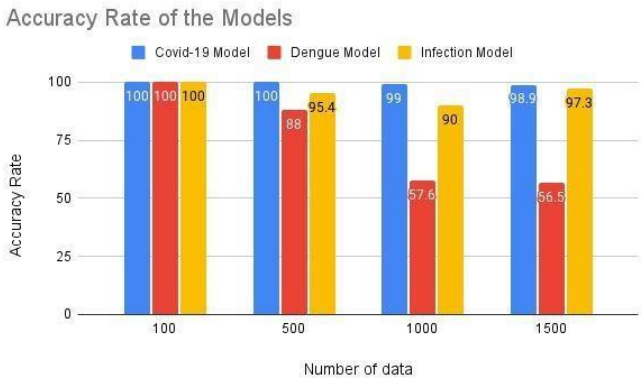

Fig2: Accuracy Rate of the Model

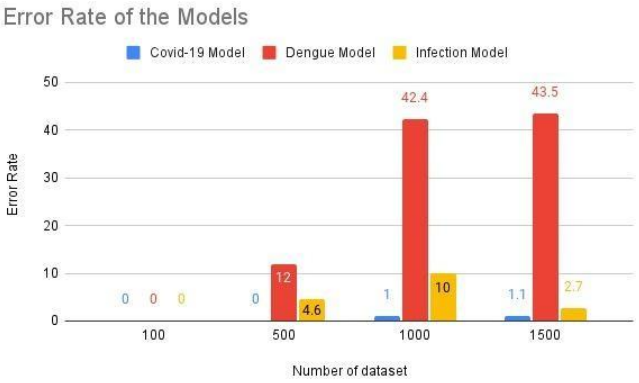

Fig3: Error Rate of the Model

V. CONCLUSION

In conclusion, the proposed system harnesses advanced Machine Learning regression techniques to forecast disease onset and transmission, offering a proactive approach to epidemic management. Through comprehensive analysis of varied datasets including historical disease trends, environmental factors, and human behaviors, it delivers precise predictions crucial for early intervention by public health authorities and communities alike. This system represents a significant advancement in epidemic control, providing an essential tool to bolster readiness and response measures. By leveraging ML capabilities, it presents a forward-thinking solution to the worldwide challenge of disease control, emphasizing the vital contribution of technology in safeguarding public health.

REFERENCES

[1] Loola Bokonda, P.; Sidibe,M.; Souissi, N.; Ouazzani-Touhami,K. Machine Learning Model For Predicting Epidemics. Computers2023, 12, 54.

[2] Tianyu Zhang, Fethi Rabhi, Ali Behnaz, Xin Chen, Hye-young Paik, Lina Yao, Chandini Raina MacIntyre,
Use of automated machine learning for an outbreak risk prediction tool, Informatics in Medicine Unlocked, Volume 34,2022.

[3] Saumya Yashmohini Sahai, Saket Gurukar, Wasiur R. KhudaBukhsh, Srinivasan Parthasarathy, Grzegorz A. Rempała, A machine learning model for nowcasting epidemic incidence, Mathematical Biosciences,Volume 343,2022.

[4] Md. Martuza Ahamad, Sakifa Aktar, Md. Rashed-Al-Mahfuz, Shahadat Uddin, Pietro Liò, Haoming Xu, Matthew A. Summers, Julian M.W. Quinn, Mohammad Ali Moni, A machine learning model to identify early stage symptoms of SARS-Cov-2 infected patients,Expert Systems with Applications,Volume 160,2020.

[5]S. Grampurohit and C. Sagarnal, "Disease Prediction using Machine Learning Algorithms," 2020 International Conference for Emerging Technology (INCET), Belgaum, India, 2020, pp. 1-7.

[6] Gao, Y., Cai, GY., Fang, W. et al. Machine learning based early warning system enables accurate mortality risk prediction for COVID-19, Nat Commun 11, 5033 (2020).

[7] Soudeh Ghafouri-Fard, Hossein Mohammad-Rahimi, Parisa Motie, Mohammad A.S. Minabi, Mohammad Taheri, Saeedeh Nateghinia, Application of machine learning in the prediction of COVID-19 daily new cases: A scoping review,Heliyon, Volume 7, Issue 10,2021.

[8] Sanzida Solayman, Sk. Azmiara Aumi, Chand Sultana Mery, Muktadir Mubassir, Riasat Khan,Automatic COVID-19 prediction using explainable machine learning techniques,International Journal of Cognitive Computing in Engineering, Volume 4,2023,Pages 36-46.

[9] Goodman-Meza D, Rudas A, Chiang JN, Adamson PC, Ebinger J, Sun N, Botting P, Fulcher JA, Saab FG, Brook R, Eskin E, An U, Kordi M, Jew B, Balliu B, Chen Z, Hill BL, Rahmani E, Halperin E, Manuel V. A machine learning algorithm to increase COVID-19 inpatient diagnostic capacity. PLoS One, 2020 Sep 22.

[10] Zoabi, Y., Deri-Rozov, S. Shomron, N. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. npj Digit. Med. 4, 3 (2021).

[11]S. Shinde, S. Yadav and A. Somvanshi, "Epidemic Outbreak Prediction Using Machine Learning Model," 2022 5th International Conference on Advances in Science and Technology (ICAST), Mumbai, India, 2022.

[12] Chae S, Kwon S, Lee D. Predicting Infectious Disease Using Deep Learning and Big Data. Int J Environ Res Public Health. 2018 Jul 27;15(8):1596. doi: 10.3390/ijerph15081596. PMID: 30060525; PMCID: PMC6121625.

[13] Sanzida Solayman, Sk. Azmiara Aumi, Chand Sultana Mery, Muktadir Mubassir, Riasat Khan,
Automatic COVID-19 prediction using explainable machine learning techniques,International Journal of Cognitive Computing in Engineering,Volume 4,2023.

[14] Seyed Ali Rakhshan, Mahdi Soltani Nejad, Marzie Zaj, Fatemeh Helen Ghane, Global analysis and prediction scenario of infectious outbreaks by recurrent dynamic model and machine learning models: A case study on COVID-19, Computers in Biology and Medicine, Volume 158, 2023.

[15] J. Tetteroo, M. Baratchi and H. H. Hoos, "Automated Machine Learning for COVID-19 Forecasting," in IEEE Access, vol. 10, pp. 94718-94737, 2022.

[16]M. S. Kaiser et al., "iWorksafe: Towards Healthy Workplaces During COVID-19 With an Intelligent Phealth App for Industrial Settings," in IEEE Access, vol. 9, pp. 13814-13828, 2021.

[17] Chatterjee R, Bajwa S, Dwivedi D, Kanji R, Ahammed M, Shaw R. COVID-19 Risk Assessment Tool: Dual application of risk communication and risk governance. Prog Disaster Sci. 2020.

[18] https://github.com/Simranpandey16/COVID-19-prediction/blob/master/Madedata1.csv