COVID-19 Prediction Model With K-Nearest Neighbour Algorithm in India

Vitthal M. Patil[1], Dr. Manoj S. Sonawane[2]
[1]*Asst. Professor, RCPET's IMRD, Shirpur*
[2]*Asst. Professor, RCPET's IMRD, Shirpur*

***Abstract*—This study proposes a COVID-19 prediction model based on symptoms using the k-nearest neighbor (KNN) algorithm. The model's effectiveness was evaluated through an experiment, and the results were analyzed to assess its predictive accuracy. The study utilized COVID-19 prediction data from the GitHub machine learning data repository, which includes 2575 patients from India, who were either positive or negative for COVID-19. These patients exhibited symptoms such as fever, body aches, runny nose, and dyspnea, with infection probabilities labeled as 1 (positive) or 0 (negative). The dataset encompasses patients ranging in age from 1 to 100, with varying fever levels. Breath issues were categorized into three types: mild (0), severe (1), and none (-1), while body soreness and runny nose symptoms were classified into two distinct categories. The data was primarily gathered through self-collection methods. After analyzing the experimental results, the KNN model demonstrated a predictive accuracy of 98.36%.**

***Index Terms*—COVID-19, KNN Model, Dataset, Supervised Machine Learning, Disease prediction.**

I. INTRODUCTION

The SARS-CoV-2 virus causes COVID-19, which has affected many people in India. While most infected individuals recover without needing special treatment, some may become seriously ill and require medical care. Understanding the disease and how the virus spreads is crucial for preventing infection. Proper hygiene, wearing masks, maintaining physical distance, and getting vaccinated are essential measures to protect yourself and others.

With limited testing capacity, quickly identifying COVID-19 symptoms is vital for controlling the spread of the virus. Machine learning algorithms are increasingly used in disease diagnosis, helping healthcare professionals improve accuracy and decision-making. This study utilizes a dataset of 2575 COVID-19 patients, both positive and negative, with symptoms like fever, body aches, runny nose, and difficulty breathing. The dataset includes patients aged 1 to 100, providing insights into the predictive power of machine learning models in diagnosing COVID-19.

II. OBJECTIVE

The objective of this study is to develop a COVID-19 prediction model using the K-Nearest Neighbor (KNN) algorithm in the context of India. Specifically, the goals are:

1.To design a machine learning model that predicts COVID-19 infection based on key symptoms such as fever, body aches, runny nose, and difficulty breathing.

2.To assess the performance of the KNN algorithm in accurately predicting COVID-19 infection in a dataset of patients from India.

3.To analyze the predictive accuracy of the KNN model using various performance metrics such as accuracy, confusion matrix, and learning curves.

4.To assist healthcare professionals by providing a tool for rapid and efficient COVID-19 diagnosis, especially in areas with limited testing resources.

5.To explore the potential of using KNN for enhancing decision-making and resource allocation in the healthcare system during the pandemic in India.

III. RELATED WORK

Using machine-learning methods to identify COVID-19 has been the subject of numerous research projects. The research projects used various machine learning methods to create a prediction model for the

COVID-19 categorization. This section discusses a few of the earlier studies on COVID-19 prediction.

1. The survey paper by Sharma et al. (2021) discusses an approach that uses machine learning and deep learning applications to predict the progression of COVID-19 cases at different stages of the epidemic. This method helps assess the impact of non-pharmaceutical measures implemented to control the spread of the virus, based on the dynamics of the epidemic.

2. The review study by Boddu et al. (2022) examines the use of application-based methods for diagnosing lung cancer in the context of COVID-19. However, according to some experts, the ARIMA approach used in the study needs further improvement, as highlighted in the contrasting analysis.

3. The overview by Algani et al. (2022) discusses various classification models applied to COVID-19. It considers factors such as the total number of COVID-19 cases, population density, the percentage of people over 65, the number of lockdown days, the duration of the COVID-19 outbreak, the availability of medical professionals, and hospital beds per 1000 people in the country. Additionally, it includes two categorical parameters: average income and the climate regions of the country.

4. According to Vidya M. Mukri (2023), Convolutional Neural Networks (CNNs) are highly efficient deep learning techniques. They have been used to achieve significant results in computer vision tasks, such as object detection and face recognition. CNNs consist of neurons with adjustable weights and biases.

IV. KNN ALGORITHM

In Machine Learning, one of the key types of learning is Supervised Learning. This involves having the correct output already available, along with a set of features related to that output. We use algorithms to train on existing data and then predict the output of new data, which only contains the features. This is similar to a teacher who teaches students, tells them what is correct, and then expects them to apply what they've learned to give the correct answers in exams. KNN is used for both classification and regression tasks in machine learning.

How the KNN Algorithm Works:

1. Select a value for K: First, choose the value of K, which determines how many neighbors to consider.

2.Choose a distance metric: For this example, we use the Euclidean distance. Calculate the Euclidean distance between the new point and its neighbors.

3.Identify the nearest neighbors: Compare all neighbors of the new point and find the K-nearest ones.

4.Classify the new point: Count the number of instances from each class among the K-nearest neighbors. Assign the new point to the class that has the highest frequency.

5.Assign the class: Finally, the new point is assigned to the class that most of its K-nearest neighbors belong to.

V. RESEARCH METHODOLGY

In this research, the researcher collected Covid-19 data from github data repository for training and testing the proposed KNN model. For implementation and experimental testing, the researcher employed Python 3.7 programming language. A statistical method is Pearson's correlation analysis and data visualization as well as feature relationship measures are employed for the identification and interpretation of heart disease data repository to find out the relationship between the class and the features in observations. To develop COVID-19 prediction, model the researcher employed KNN algorithm. Figure 1 demonstrates COVID-19 distribution in the dataset.

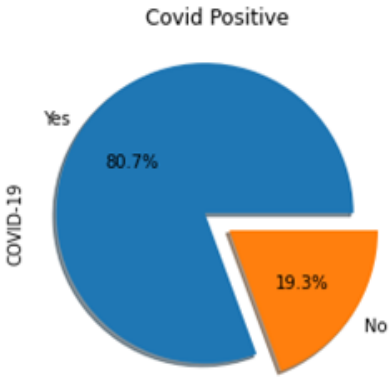

Figure 1. demonstrates COVID-19 distribution in the dataset

Dataset Description

Breathing Problem	Breathing Problem (Yes=1 No=0)
Fever	Fever (Yes=1 , No=0)
Dry Cough	Dry Cough (Yes=1 , No=0)
Sore throat	Sore throat (Yes=1 , No=0)
Running Nose	Running Nose (Yes=1 , No=0)
Asthma	Asthma (Yes=1 , No=0)
Chronic Lung Disease	Chronic Lung Disease (Yes=1 , No=0)
Headache	Headache (Yes=1 , No=0)
Heart Disease	Heart Disease (Yes=1 , No=0)
Diabetes	Diabetes (Yes=1 , No=0)
Hyper Tension	Hyper Tension (Yes=1 , No=0)
Gastrointestinal	Gastrointestinal (Yes=1 , No=0)
Abroad travel	Abroad travel (Yes=1 , No=0)
Contact with COVID Patient	Contact with COVID Patient (Yes=1 , No=0)
Attended Large Gathering	Attended Large Gathering (Yes=1 , No=0)
Visited Public Exposed Places	Visited Public Exposed Places (Yes=1 , No=0)
Family working in Public Exposed Places	Yes=1 , No=0
Wearing Masks	Wearing Masks (Yes=1 ,No=0)
Sanitization from Market	Sanitization from Market (Yes=1 , No=0)
COVID-19	Presence or absence of COVID-19 infection. (Yes=1 , No=0)

Table 1 Feature of COVID-19 Dataset

VI. FEATURE CORRELATION MODEL

The author has employed Pearson's correlation analysis for visualization of the relationship between each feature. This helps to identify the feature that is strongly related to the class feature in the data repository. The Pearson's correlation matrix for each feature of the COVID-19 dataset is shown in Figure 2. As illustrated in Figure 2, some of the features are highly correlated. For instance, Dry Cough and abroad travel has correlation value 0.33 Similarity fever is highly correlated to Sore throat with correlation coefficient 0.32. In addition, number of Contact with COVID Patient has high correlation with Attended Large Gathering with correlation coefficient of 0.23. In contrast, features such as Gastrointestinal, Hyper Tension and Fatigue has negative correlation value with some feature.

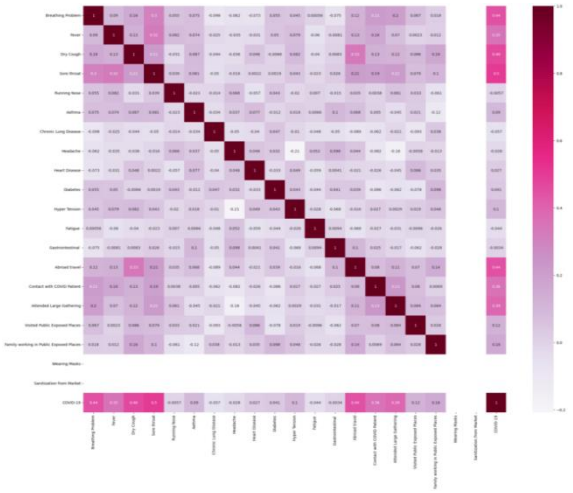

Figure 2: Heatmap COVID-19 Feature

VII. RESULTS

This section explains the experimental test results of the proposed model. The predictive performance of the Decision Tree and Adaptive Boosting algorithms is analysed using performance metrics such as accuracy, confusion matrix, and the learning curve of the algorithms. Figure 3 illustrates the accuracy of the proposed KNN model. The highest accuracy score achieved on a random test is 98.36%. The predictive performance of the proposed model was also tested on the training set, and the results are shown in Figure 3.

```
[47] K-NEAREST NEIGHBOURS
ROC_AUC value : 97.47213154797593 %

Mean Squared Error : 2.5758969641214353 %

R2 score is : 83.10350187640175 %

Accuracy Score : 98.3666896710375 %

Classification Report :
              precision    recall  f1-score   support

           0       0.90      0.98      0.93       204
           1       0.99      0.97      0.98       883

    accuracy                           0.97      1087
   macro avg       0.95      0.97      0.96      1087
weighted avg       0.98      0.97      0.97      1087
```

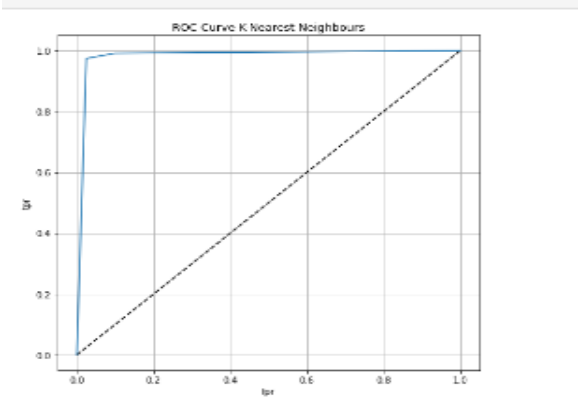

Figure 3 Performance of KNN Model

VIII. CONCLUSION

In conclusion, this study used a COVID-19 dataset from the GitHub machine learning repository and proposed a KNN-based model for prediction. Performance metrics like accuracy and the confusion matrix were used to evaluate the model, showing that the KNN algorithm outperforms logistic regression. Despite challenges like rising hospital demands and shortages of medical equipment, the KNN model plays a crucial role in supporting quick clinical decisions and optimizing healthcare resources. The RT-PCR test, the gold standard for COVID-19 diagnosis, remains scarce in developing countries, exacerbating infection rates and delays in preventive actions. Effective screening through models like KNN can help diagnose COVID-19 efficiently, easing the burden on healthcare systems and assisting medical staff in triaging patients, particularly with limited resources.

REFERENCES

[1] World Health Organization, "Coronavirus disease (COVID-19)," WHO, 2020. [Online]. Available: https://www.who.int

[2] S. Sharma, R. Sharma, and A. Singh, "A survey on machine learning and deep learning applications for COVID-19 prediction," Journal of Medical Systems, vol. 45, no. 4, pp. 1–15, 2021.

[3] R. Boddu, S. R. Kumar, and P. Reddy, "Machine learning approaches for COVID-19 diagnosis: A review," Journal of Healthcare Engineering, vol. 2022, pp. 1–12, 2022.

[4] Y. Algani, M. H. Alshahrani, and A. Alharbi, "COVID-19 prediction using machine learning classification models," International Journal of Advanced Computer Science and Applications, vol. 13, no. 2, pp. 456–464, 2022.

[5] V. M. Mukri, "Applications of convolutional neural networks in medical diagnosis," International Journal of Computer Applications, vol. 185, no. 8, pp. 22–27, 2023.

[6] T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21–27, Jan. 1967.

[7] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann, 2011.

[8] I. H. Witten, E. Frank, and M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, 4th ed. Burlington, MA, USA: Morgan Kaufmann, 2017.

[9] GitHub Machine Learning Repository, "COVID-19 Dataset," 2020. [Online]. Available: https://github.com

[10] S. Aggarwal, "Machine learning approaches for disease prediction in healthcare," International Journal of Engineering Research & Technology, vol. 9, no. 6, pp. 112–118, 2020.

[11] A. Esteva et al., "A guide to deep learning in healthcare," Nature Medicine, vol. 25, pp. 24–29, 2019.

[12] Centers for Disease Control and Prevention, "Symptoms of COVID-19," CDC, 2020. [Online]. Available: https://www.cdc.gov