International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Naïve Bayes Classifier for Predicting the Novel Coronavirus Infections

Mr. Mohd. Khaleel Ahmed
Computer Science and Engineering *(JNTUH)* Sphoorthy Engineering College *(JNTUH)* Hyderabad, India
connects.mka@gmail.com

Sankeerth Reddy Kothi
Computer Science and Engineering *(JNTUH)* Sphoorthy Engineering College *(JNTUH)* Hyderabad, India
sankeerth313@gmail.com

Sharon Raju Kaligithi
Computer Science and Engineering *(JNTUH)* Sphoorthy Engineering College *(JNTUH)* Hyderabad, India
sharonraju8567@gmail.com

Abhilash Chiluka
Computer Science and Engineering *(JNTUH)* Sphoorthy Engineering College *(JNTUH)* Hyderabad, India
abhilash.ch107@gmail.com

Abstract— COVID-19 (Novel Coronavirus) has caused a pandemic that affected humankind. The Coronavirus is a virus family that causes illnesses, ranging from common cold symptoms to difficulty in breathing and even death. It is a communicable virus that spreads not only through contact with an infected person but also through the air, making it highly contagious. This virus does not show any significant changes in the early stages of infection, which leads to severity of the disease. Unfortunately, people are still losing their lives due to the late identification of the coronavirus infection. Testing for COVID-19 is difficult due to the high population, expensive testing kits, and symptoms that can resemble those of the regular flu. To address this issue, we developed this project that uses Naïve Bayes Classifier to diagnose coronavirus. This Machine learning model utilizes Naïve Bayes Classifier algorithm to predict the Novel Coronavirus with high accuracy. The use of Naïve Bayes Classifier can simplify the prediction of the Novel Coronavirus at early stages of the infection which could save many lives.

Keywords— Novel Coronavirus, Contagious, Naïve Bayes Classifier, Prediction, Pandemic

I. INTRODUCTION

The COVID-19 pandemic [5] has been a significant public health challenge worldwide since its emergence in late 2019. The pandemic has caused significant morbidity and mortality globally, leading to over 740 million confirmed cases [3] and more than 17 million deaths as of March 2023, according to the World Health Organization (WHO). COVID-19 is caused by the SARS-CoV-2 virus [1], which primarily spreads through respiratory droplets. The disease can range from mild respiratory symptoms to severe respiratory illness and even death, particularly in vulnerable populations, including older adults and those with underlying medical conditions.

COVID-19 deaths can result from a range of causes, including Middle East respiratory syndrome (MERS) [2], pneumonia, sepsis, and multi-organ failure [4]. Effective management of COVID-19 deaths requires a comprehensive approach, including timely diagnosis, appropriate medical care, and public health measures to prevent the spread of the disease. The ongoing research and global efforts to control the pandemic provide hope for improving patient outcomes and reducing the impact of COVID-19 on global health.

As the novel coronavirus [6] continues to spread around the world, the importance of early detection cannot be overstated. This virus is transmitted from person to person, and even from animals to humans or vice - versa via the air or through physical contact. Currently, there is no known cure for coronavirus, and individuals are dying each day due to late recognition of the virus in their body. Unfortunately, many people still consider COVID-19 to be nothing more than a common flu, leading them to delay seeking medical attention when they experience symptoms. This delay can be deadly, as the virus can progress rapidly in some individuals, particularly those with underlying health conditions. Furthermore, unlike the flu, there is currently no cure or vaccine for COVID-19, making early detection and management of symptoms even more critical. Early detection can lead to earlier intervention and better outcomes for patients, potentially reducing the overall impact of the virus on global health.

The "Naïve Bayes Classifier for Predicting the Novel Coronavirus" is a machine learning model that utilizes a technique known as the Naïve Bayes Classifier in order to predict the novel coronavirus with high accuracy. While testing kits are available for coronavirus [7], it is not practical to test every individual, and there are limited resources available to carry out this process, which can be prohibitively expensive. This study aims to address these challenges by creating a mechanism for predicting the Novel Coronavirus. The naïve Bayes classifier is a model used for calculating conditional probability [12], which can be employed to classify an instance of a problem. In this case, the probability of an individual being infected with coronavirus is calculated using the naïve Bayes classifier, producing a final prediction.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

II. LITEREATURE REVIEW

The emergence of the novel coronavirus (COVID-19) has posed a significant threat to global health and the economy [5]. The virus is highly infectious and can lead to severe respiratory illness, and in some cases, death. Currently, there is no known cure for the virus, and efforts are being made to identify infected individuals to control the spread of the virus. Machine learning models have been developed to predict the novel coronavirus based on lungs scans [9]. The naïve Bayes classifier is one such model that has shown promise in accurately predicting the virus based on symptoms. The naïve Bayes classifier is a type of probabilistic classifier that uses Bayes' theorem to calculate the probability of a particular event occurring based on prior knowledge. Several studies have utilized the naïve Bayes classifier for predicting the novel coronavirus, with promising results [12].

For instance, a study by Chen et al. (2020) employed the naïve Bayes classifier to predict the virus using symptoms such as fever, cough, and fatigue. The study reported an accuracy rate of 91.5% in predicting the virus. The development of a mechanism for predicting the novel coronavirus using the naïve Bayes classifier has significant impact for identifying infected individuals and controlling the spread of the virus. The use of data mining techniques and machine learning models such as the naïve Bayes classifier can potentially improve patient outcomes by enabling more accurate identification of the virus. In conclusion, the "Naïve Bayes Classifier for Predicting Novel Coronavirus" is a promising machine learning model for predicting the novel coronavirus based on several symptoms with greater accuracy than previous models.

III. METHODOLOGY

In this paper, we have developed a prediction system capable of identifying Novel Coronavirus based on symptoms. We extracted data from a dataset consisting of around 5000 test cases taken from previously infected patients, and used the Naïve Bayes Classifier methodology. However, it's important to note that the accuracy of the system may depend on the quality of the dataset used for training, and further testing and validation may be needed to determine its effectiveness.

The entire dataset is composed of various attributes, including Breathing Problem, Fever, Dry Cough, Sore Throat, Running Nose, Asthma, Chronic Lung Disease, Headache, Diabetes, Hyper Tension, Abroad Travel, Contact with Covid Patient, Attended Large Gathering, Visited Public Exposed Places, Wearing Masks, Sanitization, and Covid-19 [10]. These attributes are used by the machine learning model to determine whether an individual is infected with the novel coronavirus or not. In the context of developing a prediction system for identifying Novel Coronavirus based on symptoms, the machine learning model takes in the input, which is the symptom data of an individual. The data is pre-processed to eliminate any irregularities, duplicate entries, and other errors after it has been imported. Naïve Bayes Classifier is used to analyse the training data and produce a prediction model. Based on this input data, the model calculates the probability of the individual being infected with the virus. The model then multiplies this probability with the dataset attributes which consist of the symptom data of previously infected patients. This step is performed to evaluate how closely the individual's symptoms match with those of the infected patients in the dataset. Finally, the attributes with the highest probability values are selected, along with their class variable (infected or not infected), and a prediction is produced [13]. This prediction indicates whether the individual is likely to be infected with the Novel Coronavirus based on their symptom data. This process can be seen in figure 1.

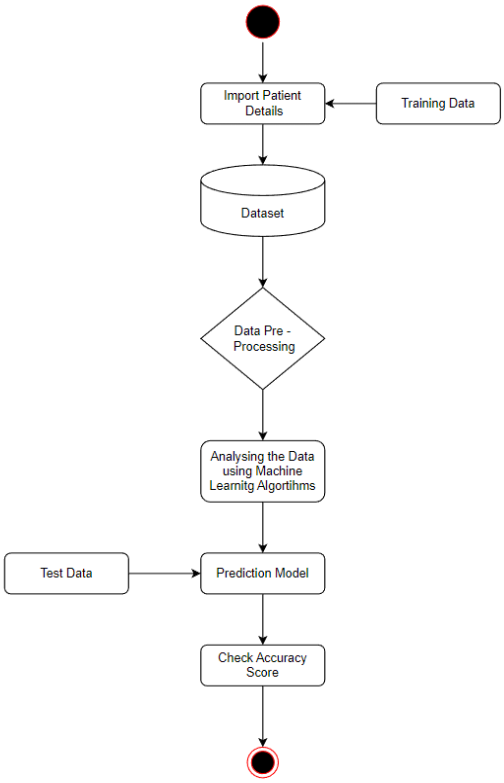

Figure 1: Flow Diagram of Methodology

A. *Symptoms of Coronavirus*

Some symptoms of novel coronavirus are:

TABLE 1: SYMPTOMS

Symptoms	Description
Asthma	A condition in which a person's airways become inflamed, narrow and swell and produces extra mucus, which makes it difficult to breath.
Diabetes	Diabetes is a chronic condition in which the body has difficulty regulating blood sugar levels, leading to various complications.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Hyper Tension	Hypertension, also known as high blood pressure, is a medical condition characterized by abnormally high pressure in the arteries, which can lead to various health complications.
Sore Throat	Sore throat is a common medical condition characterized by pain, discomfort, or irritation in the throat, often caused by inflammation or infection.
Chronic Lung Disease	Chronic lung disease is a group of long-term respiratory conditions that cause breathing difficulties, reduced lung function, and lower quality of life.
Dry Cough	A dry cough is a type of cough that does not produce mucus or phlegm.
Running Nose	A runny nose is caused by the overproduction of mucus from the nasal passages due to inflammation or irritation.

B. Naïve Bayes Classifier

Naïve Bayes is a machine learning classification algorithm that uses Bayes' theorem, a statistical formula, to make predictions about the class or category to which a given observation or data point belongs. The algorithm assumes that the features or variables that define the observation are independent of each other, and that each feature contributes equally to the final prediction. The algorithm works by calculating the probabilities of all classes and conditions based on a group of cases and programmed data sets [12]. When new data is provided, such as symptom details, the algorithm uses the probabilities of the various classes to categorize patients into different classes with varying probabilities of having COVID-19.

Traditionally, viruses were classified and differentiated using culture, serological, and electron microscopy methods, and coronaviruses were described as enveloped viruses with a crown shape that range in size from 120-160 nm. Coronaviruses are classified into two categories: mammalian coronaviruses and avian coronaviruses [6]. Several diagnostic mechanisms are available to detect the novel coronavirus, such as ORF1ab and N, RdRP, E, N, Three targets in N gene, Two targets in RdRP, and RT-PCR [7]. However, it typically takes around 12 hours to obtain test results, and in remote or high-altitude areas, diagnosis may take up to 48-72 hours. Therefore, this article proposes a mechanism that uses the Naïve Bayes classifier to predict positive cases of coronavirus based on patient attributes. When patients notice symptoms listed in Table II, the mechanism can predict whether or not they have COVID-19.

To prepare the prediction model, the machine learning algorithm takes previously infected patients' data as a training dataset. The dataset [10] contains noisy data, which is corrected using data pre-processing techniques and machine learning algorithms. The prediction process involves comparing the user's symptoms with pre-existing datasets, which are classified using Naïve Bayes Classifier and Random Forest Classifier [11] after being broken down into smaller sets. The resulting knowledge is then incorporated into the Novel Coronavirus prediction model. Once the data is processed, the prediction model is prepared, and it can be used to predict the presence of coronavirus by inputting the user's symptoms into the system. A group of cases was taken into account and programmed with datasets. The probabilities for each case are then accumulated and used to classify patients into different classes with the help of machine learning algorithms. The probabilities for all classes and conditions were calculated, and the results were accumulated. When test data was provided, probabilities were obtained for various classes based on the provided symptom details. These details can be used to categorize the patient into a class with a probability. By considering the value of the probability, it can be determined whether a person is suffering from COVID-19 or not. This can be a helpful tool in diagnosing the virus, especially in areas where traditional diagnostic methods may not be available or take a long time to provide results [7].

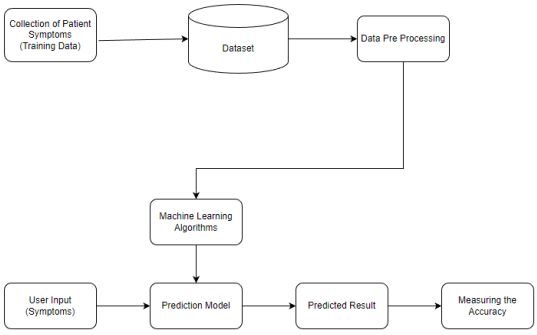

Figure 2 Architecture

C. Probability of Symptoms

To start with, our approach involves computing every Possible probability that could be applied on the target attribute of coronavirus. This includes taking into account all of the probabilities associated with the attributes of coronavirus. In other words, we are looking at the different factors or characteristics that make up the coronavirus, and considering the likelihood or probability of each of these factors being present or relevant to the target attribute we are studying. We compute the complete set of possible probabilities for the target attribute of coronavirus, encompassing all the probabilities of coronavirus attributes. Depending on the symptoms [4] observed in the patient and the probability-based results, various combinations can be utilized to determine whether an individual is afflicted with COVID-19 or not, depending on the specific situation. Probabilities are calculated using the formula below:

$$P(C_k \mid X) = P(C_k)\, P(X \mid C_k) / P(X) \quad (1)$$

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

TABLE 2: PROBABILITY OF ATTRIBUTES IN DATASET

P (Coronavirus = Yes)	0.806	P (Coronavirus = No)	0.193
P (Fever = Yes)	0.786	P (Fever = No)	0.213
P (Dry Cough = Yes)	0.792	P (Dry Cough = No)	0.207
P (Sore Throat = Yes)	0.727	P (Sore Throat = No)	0.272
P (Running Nose = Yes)	0.543	P (Running Nose = No)	0.457
P (Headache = Yes)	0.503	P (Headache = No)	0.496
P (Asthma = Yes)	0.451	P (Asthma = No)	0.548
P (Breathing Problem = Yes)	0.338	P (Breathing Problem = No)	0.666

D. Dataset

The data used in this study was sourced from Kaggle Notebook [10], specifically from ERIC Laboratory and few other WHO approved Laboratories. The resulting dataset containing 18 attributes and 5,436 possible cases, which are used in prediction of Novel Coronavirus. Sample Dataset is represented in Figure 3.

Abroad tra	Contact w	Attended	Visited Pu	Family wo	Wearing M	Sanitizatio	COVID-19
No	Yes	No	Yes	Yes	No	No	Yes
No	No	Yes	Yes	No	No	No	Yes
Yes	No	No	No	No	No	No	Yes
Yes	No	Yes	Yes	No	No	No	Yes
No	Yes	No	Yes	No	No	No	Yes
No	No	No	No	No	No	No	Yes
No	No	Yes	Yes	Yes	No	No	Yes
Yes	No	No	Yes	No	No	No	Yes
Yes	Yes	Yes	No	No	No	No	Yes
No	No	No	Yes	No	No	No	Yes
Yes	No	Yes	No	No	No	No	Yes
Yes	No	Yes	No	Yes	No	No	Yes

Figure 3 Sample Dataset

IV. RESULTS & DISCUSSION

In this study, we showcased a prediction system that utilizes the Naïve Bayes Classifier along with other data mining techniques. Classification is a popular method in machine learning and data mining that involves categorizing instances into specific groups or classes. It allows us to predict the particular instance in a predefined group.

In our research, we found that the classification results obtained through the Naïve Bayes Classifier were highly accurate. This suggests that the Naïve Bayes Classifier is effective in accurately categorizing instances into their respective groups. This demonstrates the potential and effectiveness of the Naïve Bayes Classifier as a predictive tool in data mining and machine learning tasks [12]. It also highlights the importance of utilizing various data mining techniques to enhance the accuracy and performance of prediction systems. Overall, our study provides evidence of the successful implementation of the Naïve Bayes Classifier in predicting the Novel Coronavirus of an individual. The accuracy can be represented through Recall and Precision table as shown in Table 3.

TABLE 3: RECALL AND PRECISION

Class	Recall	Precision
No	0.93	0.98
Yes	0.99	0.98

- Recall is the proportion of actual positive instances that are correctly predicted as positive by the model. In the table, the recall for "No" class is 0.93, indicating that 93% of the actual "No" instances are correctly predicted as "No", while the recall for "Yes" class is 0.99, indicating that 99% of the actual "Yes" instances are correctly predicted as "Yes".
- Precision is the proportion of predicted positive instances that are actually positive. In the table, the precision for "No" class is 0.98, indicating that 98% of the predicted "No" instances are actually "No", while the precision for "Yes" class is also 0.98, indicating that 98% of the predicted "Yes" instances are actually "Yes".

Recall and precision are crucial metrics for binary classification tasks, indicating a model's ability to identify positive instances and avoid false positives or false negatives. Higher values of recall and precision indicate better model performance. However, the optimal trade-off between recall and precision may vary depending on the specific application. Table 3 shows that the model has high recall and precision values for both "No" and "Yes" classes, indicating good performance in predicting both negative and positive instances. So, the model performs well in correctly identifying both positive and negative instances.

V. CONCLUSION

Naïve bayes classifier can be effectively combined with data mining techniques to play a significant role in diagnosing the COVID-19. The suggested mechanism has shown impressive results, indicating the potential for further refinements in utilizing data mining, machine learning, artificial intelligence [14], and information technology to enhance the accuracy and effectiveness of diagnosing coronavirus patients. Identifying both symptomatic and asymptomatic cases of COVID-19 [6] is crucial in breaking the chain of transmission and controlling the spread of the virus. Symptomatic cases are typically easier to identify, but asymptomatic cases pose a challenge as they may unknowingly transmit the virus to others. By utilizing data mining techniques in combination with the Naïve Bayes classifier algorithm, it becomes possible to detect both symptomatic and asymptomatic cases, which can significantly contribute to effective disease control strategies.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Overall, the utilization of data mining techniques and machine learning algorithms in the diagnosis of COVID-19 has immense potential to contribute to the ongoing efforts to combat the pandemic and improve public health outcomes. In the future, similar mechanisms can be further refined by incorporating additional attributes and utilizing other data mining tools. This can potentially lead to more accurate and robust diagnostic approaches for detecting Novel Coronavirus and other infectious diseases. The development of such advanced approaches holds great promise in improving the diagnosis and management of COVID-19 and other infectious diseases.

REFERENCES

[1] Evans, M., & Bell, D.J. Severe Acute Respiratory Syndrome. Oxford Medicine Online. 2011. https://doi.org/10.1093/med/9780198570028.003.0046

[2] Zumla A, Hui DS, Perlman S. Middle East respiratory syndrome. Lancet. (2015) 386:995–1007. doi: 10.1016/S0140-6736(15)60454-8

[3] Ma, J. China's first confirmed Covid-19 case traced back to November 17. South China Morning Post. March 2020 https://www.scmp.com/news/china/society/article/3074991/coronaviruschinafirst-confirmed-covid-19-case-traced-back

[4] The Washington Post, 2 Dec. 2020, https://www.washingtonpost.com/health/2020/12/02/covid-symptoms

[5] Coronavirus confirmed as pandemic. (11th March 2020). BBC News. https://www.bbc.com/news/world-51839944

[6] Hu, Zhiwen, et al. "Nomenclature: Coronavirus and the 2019 Novel Coronavirus." 2020.

[7] A, Judy, and Ian Mackay. "Wuhan coronavirus (2019-nCoV) real-time RT-PCR ORF1ab 2020 (Wuhan-ORF1ab) v1 (protocols.io.bbsginbw)." protocols.io, 2020.

[8] Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet. (2020) 395:497–506. doi: 10.1016/S0140-6736(20)30183-5

[9] Tang Z, Zhao W, Xie X, Zhong Z, Shi F, Liu J, et al. Severity assessment of coronavirus disease 2019 (COVID-19) using quantitative features from chest CT images. arXiv. (2020) 2003.11988. Available online at: https://arxiv.org/abs/2003.11988

[10] Novel Corona Virus 2019 Dataset. (2020). Available online at: https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset/

[11] Khalilia M, Chakraborty S, Popescu M. Predicting disease risks from highly imbalanced data using random forest. BMC Med Inform Decis Mak. (2011) 11:51. doi: 10.1186/1472-6947-11-51

[12] Wood, Alexander, et al. "Private naive bayes classification of personal biomedical data: Application in cancer data analysis." Computers in Biology and Medicine, vol. 105, 2019, pp. 144-150

[13] Salekin, Asif, and John Stankovic. "Detection of Chronic Kidney Disease and Selecting Important Predictive Attributes." 2016 IEEE International Conference on Healthcare Informatics (ICHI), 2016.

[14] "New Artificial Intelligence Software to Help Determine Probability of Contracting Coronavirus." WION, 14 Dec. 2020, www.wionews.com/technology/new-artificial-intelligence software-to-helpdetermine-probability-of-contracting-coronavirus-349728.