IJISCS (International Journal of Information System and Computer Science)

Full Paper
eISSN 2598-246X
pISSN 2598-0793

DISEASE PREDICTION FROM COVID-19 MEDICAL DATA USING DATA MINING ALGORITHM

Nafis Md. Zawad
Department of Computer Science and Engineering,
American International University Bangladesh, Dhaka
408/1, Kuratoli, Dhaka, Bangladesh

*Corresponding author
Email:
nafis.zawad96@gmail.com

Article history:
Received November 16, 2022
Revised December 6, 2022
Accepted December 10, 2022

Keywords:
Covid-19,
Data Mining Algorithm,
Disease Prediction,
Medical Data,
Random Forest.

Abstract

The study was designed to introduce a technique for disease prediction by using a data mining algorithm. Here in this paper, a significant discussion has been made on the Novel Corona Virus and the creation of a model for disease prediction. The novel Coronavirus (COVID-19) pandemic has created chaos in the world. People from both developed and developing countries are facing many death tolls due to insufficient ways to prevent COVID-19. It is observed that the environment requires a quick and effective way to control the spread of COVID-19 across the globe. The use of non-clinical methods like data mining techniques can be an effective way to combat the spreading of Covid-19. To minimize the immense pressure on the healthcare system, improved ways of patients' detection and diagnosis of the nature of the Covid-19 pandemic need to be ensured. In this study, an epidemiological dataset, and data mining models were applied for forecasting the extent of Covid-19 patients. To construct the models, the decision tree and logistic regression were used. Besides, a random forest algorithm was applied to the dataset by using 'Python Programming Language'. The results reveal that the model created with a 'Random Forest Data Mining Algorithm' is more effective in predicting the likelihood of Covid virus-infected patients with the correctness (accuracy) of up to eighty percent (80%).

1.0 INTRODUCTION

The World Health Organization (WHO) has labeled the COVID-19 virus a pandemic, with 630,832,131 confirmed cases of infected patients and 6,584,104 deaths globally on 11 November 2022 (World Health Organization (WHO), 2022). COVID-19 is triggered by the SARS Coronavirus 2 (SARS-CoV-2) and was deemed a pandemic by WHO on March 11, 2020. The treatment for COVID-19 could be postponed owing to the virus's potential genetic mutations (Keeling et al., 2020). The pandemic crisis affects millions of people in social, economic, and medical settings that lead to remarkable changes in macro-level human interactions, health systems, commerce, employment and educational environments. The global pandemic is a challenge to human civilization, and urgent intervention is needed. To protect people from dying, the scientific community has been brainstorming solutions to better limit the epidemic and deter possible pandemics. Here it has been tried to come up with a solution to predict a person being infected with COVID-19 based on the symptoms and age. So, the goal of the study is to help the healthcare system with the non-clinical approach and prevent more people from dying due to this pandemic.

1.1 Background of the Study and Review of Literature

Novel Corona Virus

The world's biggest modern outbreak was uncovered in the mid-twentieth century, and since then, the world's populace has been haunted by the Severe Acute Respiratory Syndrome Coronavirus, or literally, as the citizens affectionately term it, "COVID-19." The negative effects of this disease have harmed the psychological stability, peace and tranquillity of the population of the whole world. People are combining their skills to develop numerous ideas and techniques for constructing a network of machines with sensors and sensors in the cloud to monitor various data streams. Previous studies to see whether an individual had this drug in their blood were costly and time-consuming, but they were not found to be much effective. According to this rationale, since it is important to reduce death rates and classify high-risk individuals through blood tests at a quicker pace, experts are discussing novel ways of recognizing high-risk patients. The viruses in the families of '*Coronaviridae*' and '*Betacoronaviridae*' live as infectious agents in many animal species; when they mutate, they can be viruses causing disease in mammals. These respiratory infection outbreaks have been triggered by SARS in 2002, MERS in 2016, and a new respiratory infection, rhinovirus, in 2018. COVID-19 is an illness transmitted by the SARS-CoV2 virus. Unlike the previous events, this virus induces extreme lower respiratory collapse and can affect the central nervous system (CNS) in the primary stages of infection (Li et al., 2020).

The term 'SARS-CoV-2' stands for the 'Severe Acute Respiratory Syndrome Corona-Virus Two' that is thought of as the agent of the Novel Corona Virus. The virus was believed to have originated in 2019 in Wuhan, Hubei Province of China (Lai et al., 2020), (Jibril & Sharif, 2020), (Wölfel et al., 2020). The coronavirus has already arrived at a critical point and held a pandemic feature that killed a huge number of people throughout the world (Lai et al., 2020), (Rothe et al., 2020). According to the Centers for Disease Control and Prevention (CDC), 2.4 percent of adolescents living with Covid-19 is equivalent to 7.9 percent of adults above the age of 18 (Guan et al., 2020).

Symptoms

The Symptoms of Covid-19 develop typically in four to seven days in the majority of people but that can extend from two to fourteen days when they get infected by the COVID-19 virus. Major symptoms include fever (80 to 90%), cough (60% to 70%), gastrointestinal disorders (40% to 50%), smell difficulties (30% to 40%), and shortness of breathing (20%) of the infected cases (Pan et al., 2020). It is observed that not all these symptoms are common in those affected people. First, it is predicted that an individual will become sick with the disease but will not show symptoms for years. Often after an individual becomes ill the entire scope of the condition becomes apparent. Some manifestations of this disease only benefit such that they may not present with fever or respiratory problems, and in these cases, the patient can take longer to recover. The most prominent signs include a burning fever, shortness of breath, and shortness of breath. Beneficiaries of our treatment rebound in around a week whether they are not suffering respiratory problems. But patients suffering from fever need up to two weeks for healing. The most common signs of coronavirus are fever, dry cough and tiredness. Apart from these symptoms, body aches, sore throat, diarrhea, conjunctivitis, absence of taste, and changes in the colors of fingers and toes are less common signs of the virus infection (Fadugba et al., 2021). The manifestation of these disorders can take four distinct types based on the indications (Diagnosis and Treatment Protocol for Novel Coronavirus Pneumonia (Trial Version 7), 2020; Wang et al., 2021). It is evident that the most difficult symptoms are those of almost non-existent signs, people are hardly aware of the problems and continue to breathe normally. Patients moderately infected have a fever and also hold x-ray signs of pneumonia. This disease has affected about 80% of all patients in the mild to severe range.

Data Mining

The advancement in information technology has resulted in the development of a massive array of databases, archives, and documents in various fields. Database and computer management research studies have developed a tool for manipulating and preserving the necessary data for future decision-making. Data mining is a technique for collecting useful information and trends from large volumes of data. In other words, this technique is known as an information processing, intelligence retrieval, and data-pattern study. It is a computerized technique which can be applied to shift through a large amount of data to extract constructive facts. This approach aims at detecting unidentified trends. If these are identified, their success as a metric is used to determine what types of goods the business can pursue in the future. Investigating the issues is done first and then the similarities inside them.

Exploration: Data cleaning and data conversion into different formats are done at the first stage of data exploration. In this process important variables are determined and the quality of data is ensured (Saeed, 2021).

Pattern Recognition: Once the raw data has been properly interpreted and refined into a more systematic type, the next move is to form a pattern of relations between the findings. Sort the data sets (or inputs) that can be used for prediction (Saeed, 2021).

Pattern Deployment: Patterns are applied based on expected results.
Data mining provides impressive results in disease identification where the proper tools and techniques are utilized. As a consequence, data mining is a promising field in healthcare forecasting. Data mining has played an important role in the healthcare industry, especially in disease prediction. Researchers must develop hybrid models to enhance the prediction to obtain the highest forecast precision. So, the growing size of the dataset would improve the reliability of the findings (Saeed, 2021).

Data mining algorithms

There are various types of data mining algorithms. A few of the algorithms used for prediction are presented below:

- **Logistic Regression (LR)**

"Binary logistic regression analysis is a statistical method that can be applied mainly in retrospective data to explore and model the relationship between a random dichotomous variable and one or more random independent variables (continuous or categorical)"(Pounis, 2018, pp.68-70). Logistic regression, which is best known in its ANOVA form, is commonly used to study correlations between independent and categorical dependent variables. Logistic regressions (LR) refer to a kind of likelihood ratio test in which the dependent variable contains two values like zero and one, yes and no, and true and false. Therefore, it is called binary logistic regression (Ayon et al., 2022). When a dependent variable has two or more values, then multinomial logistic regression is employed for the statistical analysis. A mathematical model of the set of independent variables for logistic regression is applied for producing an expected predicted output by predicting the dependent variable.

- **Support Vector Machine (SVM)**

The Support Vector Machine is one of the 'supervised learning algorithms' used for classification and regression. Testing and training the data with some instances are required for the classification tasks support vector machine. The main objective of this algorithm is to create a model for predicting the target values (Islam et al., 2018), (Mavroforakis & Theodoridis, 2006); (Ho, 1998).

- **Decision Tree (DT)**

A 'Decision Tree' is a popular data mining strategy because of its capacity to manage the categorical as well as the continuous data, and its flexibility and lucidity. This strategy divides the tree into some phases involving the growth and pruning of the phases (Yahaya et al., 2018); (Kumar et al., 2022). A tree is constructed in the first step by dividing data into smaller sets as long as each partition becomes pure (Cao & Xu, 2009);(Kohavi & Quinlan, 1999). However, the development process of the decision tree is more costly in terms of computation than the pruning phase (Hussain et al., 2019).

- **Naive Bayes (NB)**

Naive Bayes is applied for differentiating the dataset instances based on listed features and attributes (Muhammad et al., 2019). This algorithm is thought to be a 'probabilistic classifier' that employs the Bayes principle to perform classification tasks (Gandhi, 2018);(Singh et al., 2022).

- **Random Forest (RF)**

The 'Random Forest' algorithm is a method for the classification and regression tasks in the data mining process. During the process, the algorithm produces a large number of decision trees, which generate outputs (Haque et al., 2018). This algorithm is the strongest technique for any decision tree that has evident links to its training dataset (Muhammad et al., 2019); (Singh et al., 2022).

- **K-Nearest Neighbour (K-NN)**

The K-Nearest Neighbour is a 'supervised' and 'non-parametric data mining classifier' which is used for the regression and classification tasks (Altman, 1992). The input variables in both tasks are the K closes training dataset in the function room. K-NN relies on labelled input data to learn a feature that produces acceptable performance when unlabelled data is inputted (Harrison, 2018). The performance of K-NN classification is a class membership in which data instances are categorized by a majority vote of its neighbours, with the data instance being allocated to the class most popular among its K-nearest neighbours, while the output of K-NN regression is the property value of data instance and this value is the average of the value of K-nearest neighbours (Everitt, 2011).

The prediction may be accomplished using the regression technique. To model the relationship between one or more independent variables and one or more dependent variables, regression analysis may be used. Independent variables are proven attributes in data mining, and answer variables are what we want to forecast. Unfortunately, several real-world issues are unpredictable. Sales amounts, market values, and product failure rates, for example, are all impossible to estimate due to dynamic relationships with various predictor variables. As a consequence, more sophisticated forecasting techniques (such as logistic regression, decision trees, or neural nets) might be needed (Desuky, 2022). The same model styles are often used for both regression and classification. For example, the CART (Classification and Regression Trees) decision tree algorithm may be used to construct both classification trees (to define categorical answer variables) and regression trees (to forecast continuous response variables). Classification and regression models may also be produced by neural networks.

By the use of a decision tree, HCV polyprotein cleavage sites have been predicted and these predictions have proved to be satisfactory. The final results of the process aren't as good as one would be desired, but the decision tree is what helps produce those results. In the future, they will add more components to the models, including the secondary structure, to study their effects on the human body. The Naive Bayes and Decision Tree are the classifier algorithms used to construct their models. Those patients who work as healthcare workers are supposed to stay alive are first and foremost those who have a good prognosis. In this study, the age of the patient was a strong predictor of how long it would take to find a cure. When patients age between the ages of 66 and 87, they have a higher likelihood of suffering from serious complications. The models were tested and compared, so they were judged to all be equal. The estimated model-accuracy percentage ranges between 53.6% and 71.58% depending on the dataset used (Al-Turaiki et al., 2016).

Concerning removing measurements used to evaluate the proximity of neighbours in KNN and its subsidiaries, the normal measurement is Euclidean distance. In any case, Mahalanobis distance turns out to be more reasonable if the information is slanted or the highlights are corresponded, as it mulls over information dissemination. Jaafar et al., (2018) report that Euclidean distance crumbles KNN precision if the information is uneven. In this manner, they propose Mahanalobis distance for a more precise order. Yi et al., (2018) propose a characterization framework dependent on KNN appropriate for automated frameworks. Since robots work in certifiable conditions, where highlights are emphatically associated, they also use Mahalanobis distance. To moderate the computational intricacies associated with Mahalanobis distance, because of figuring the reverse covariance lattice of information, they utilize guideline segment examination (PCA) for information decrease. Fan et al., (2019) additionally use Mahalanobis distance with KNN concerning a structure to improve the security of force frameworks, where highlights are normally profoundly connected. Data mining was utilized for forecasting and making conclusions on numerous illnesses. Ferreira et al., (2012) utilized it to promote the determination of neonatal jaundice. In the study, a dataset comprising 70 variables was gathered for 227 sound babies. Numerous data mining strategies were used, that included J48, Naive Bayes and straightforward calculation. The best prescient models were obtained by utilizing Naive Bayes, multi-facet perceptron, and basic calculation for coronary illness analyses, Venkatalakshmi & Shivsankar, (2014) looked at the presentation of DT calculation and NB. Another study had a dataset of 294 records along with 13 ascribes and showed that the presentation of the 02 calculations is found to be equivalent. Frequent Pattern (FP) growth and decision trees (DT) were utilized in determining and visualizing the bosom malignant growth (Majali et al., 2015).

Table 1-A: Existing method with advantages and disadvantages

Methods	Advantages	Disadvantages
Logistic Regression	LR mainly finds the probability that a new object belongs to a certain class (Prasad et al., 2019).	The linear decision is limited for logistic regression.
Random Forest	RF provides noticeable growth in the 'classification accuracy of a model by constructing a group of trees that produce individual findings (Villavicencio et al., 2021).	RF is not suitable for regression tasks.
Decision Tree	Missing values in the dataset do not affect the process of building a DT to a substantial degree.	This algorithm is not sufficient to employ regression and predict continuous values.

The existing methods that are provided have some advantages and disadvantages. We have to consider those points and provide a suitable model with fewer disadvantages.

Table 1-B: Comparisons between previous research performance measures

Methods	Accuracy	Precision Rate	Recall Rate	f-score
Proposed Method	80%	.80	1.0	.89
Villavicencios proposed method (Villavicencio et al., 2021)	98.81%	.988	.988	.988
Ahmad's proposed method (Ahamad et al., 2020)	89%	-	-	-
Iwendi's proposed method (Iwendi et al., 2020)	94%	1.0	.75	.94
Khanday's proposed method (Khanday et al., 2020)	94%	.93	.94	.93
Zhang's proposed method (Zhang et al., 2021)	75%	-	-	-

From the existing models that the other researchers have built, the majority of them have better accuracy than our model. Due to not having a proper dataset might have caused our model to have a slightly low accuracy than others. But we believe that with the proper dataset we might have better accuracy than this accuracy.

1.2 Objectives of the study

The prime goal of the study is to find an efficient way to predict if the patient is infected with COVID-19 or not. For this, a model had to be created using various data mining algorithm which is used in machine learning. To fulfill the objective, various steps had to be taken. Those steps are to:

1. Understand the problem in the existing models and find out the issues that caused it.
2. State problems for determining a COVID-infected patient.
3. Propose a model for the prediction of the disease.
4. Propose solutions for the problem.
5. Develop an efficient model for the solution.

1.3 Justifications of the Study

This paper intends to provide an efficient technique to predict disease (COVID-19). As the novel coronavirus pandemic is one of the biggest issues in recent times, it has affected all sectors such as societal, economic, health systems, commerce, employment and educational environments. By conducting the research, the researcher will be able to find out an efficient way to find the solution to predict coronavirus-infected patients. In the end, the researcher proposed a model which efficiently predicted the coronavirus disease and analyzed the symptoms of the affected patients to improve the model. In the proposed model, the researcher ensured the efficiency and accuracy of detecting coronavirus disease which showed a beneficial improvement in the research. Also, it provided a solution for the recent major problem of the world.

2.0 THEORETICAL DISCUSSIONS OF THE STUDY

2.1. Theoretical models related to the study

Model selection

The researcher has implemented the model by using python language in the Jupyter notebook. Jupyter Notebook is an open-source web application that allows programmers to generate and exchange documents with live code, calculations, visualizations, and text. The staff at Project Jupyter are in charge of the Jupyter Notebook (Perkel, 2018).

Computational notebooks are computational computing laboratory notebooks. Rather than pasting DNA gels alongside lab protocols, for example, researchers insert coding, records, and text to record their computational methods. For data scientists, this format may be a source of inspiration for more research (Perkel, 2018). Notebooks are a type of collaborative computing, an environment in which users execute code, see what occurs, alter, and repeat in an iterative dialogue between researcher and results. They aren't the only place for such discussions; IPython, the collaborative Python interpreter on which Jupyter's predecessor, IPython Notebook, was developed, is another. Notebooks, on the other hand, encourage users to record such interactions, resulting in more efficient relations between subjects, theories, evidence, and results (Perkel, 2018).

The project needs the following packages and libraries: Datetime, Numpy, Pandas, SciPy, Scikit Learn, and Matplotlib. The project was built on the Jupyter notebook platform with the CPU runtime.

The researcher trained several ML classification models using the pre-processed dataset. This research includes the following models: Decision Tree Classifier, Logistic Regression Classifier, and Random Forest Classifier. Since the dataset we used can be unbalanced, we can use the F1 Score as the primary metric.

The following are the implementation steps:

- Pre-processing of data
- The Random Forest algorithm is being fitted to the Training sample.
- Predicting the outcome of the model
- The result's consistency was checked (Creation of Confusion matrix)
- Visualizing the outcome of the test sample.

These are the steps followed in making the model.

Logistic Regression

The relationship between categorical dependent variables and independent variables can be determined with a method called Logistic Regression (Ng, 2019; Singh et al., 2022). A conditional logistic regression is when both values of the dependent variable are 0and 1 or true and false or yes and no (Ng, 2019; Singh et al., 2022). Multinomial logistic regression is utilized where more than two values are contained within the dependent variable. To simulate a change in the dependent variables, a sequence of explanatory variables is used. Mathematically, the LR alteration is written as:

$$LR(p) = \ln\left(\frac{p}{1-p}\right) \qquad (1)$$

Decision Tree

The decision tree is a popular strategy in data mining because of its capacity to manage categorical data and also continuous data. Because of the simplicity of the algorithm, it is easy to comprehend (Muhammad et al., 2020; Kumar et al., 2022). A tree is constructed in the first step by dividing data into smaller sets until all the small partitions turn pure. But the splitting process produces many split forms of the data and which is based on the type of data found in the dataset. (Kohavi & Quinlan, 1999). However, the development process of the decision tree seems more costly than the process of pruning (Kumar et al., 2022).

Random Forest

A random forest is an indicator that is composed of randomized base regression trees. $\{rn(x, \Theta m, Dn), m \geq 1\}$, where $\Theta 1, \Theta 2, \ldots$ are i.i.d. outputs of a randomizing variable Θ. The aggregated regression approximation is generated by combining these random trees (Hasan et al., 2021; Muhammad et al., 2020; Singh et al., 2022).

$$\bar{r}n(X, Dn) \;=\; E\Theta\ [rn(X, \Theta, Dn)] \qquad (2)$$

In the equation $E\Theta$ is the expectancy concerning the random parameter, conditionally on X, and the dataset Dn. To simplify the representation, after removing the dependence of the projections in the sample and it can be written as $\bar{r}n(X)$ instead of $\bar{r}n(X)$ (X, Dn). Generally, the above assumption is calculated using Monte Carlo, that is, by generating M random trees and averaging the discrete outcomes (Singh et al., 2022). A structure of a random forest is given below:

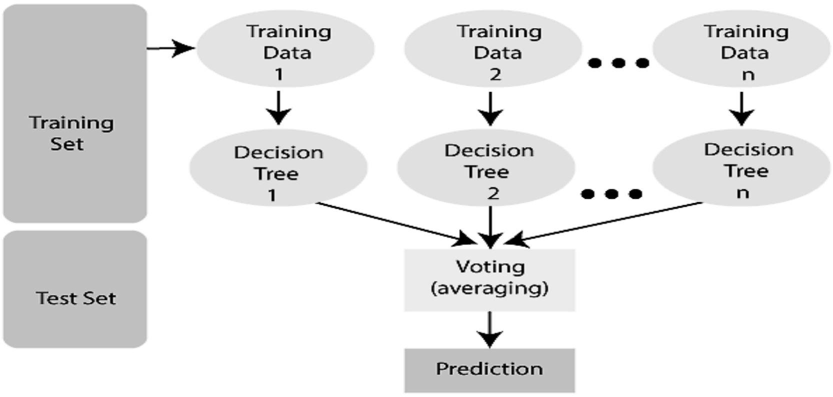


Figure 3: General process of random forest

The randomizing variable is used to decide how the subsequent three cuts are made when constructing the individual trees, such as the coordinate to break and the location of the split. For this paper, we are going to take random forest due to its efficiency in computational time.

The rationale for selecting random forest

The Random Forest algorithm has the following advantages (Mahamunkar & Netak, 2022).

- It avoids over-fitting by averaging or combining the outcomes of multiple DTs.
- Random forests perform more effectively than single decision trees over a variety of data elements.
- The difference between a random forest is lower than the difference between a single DT.
- Random Forests are much more adaptable and have a higher level of precision.
- The RF algorithm does not need data scaling. It retains good precision even though data is provided without scaling.
- Where a significant majority of the data is unavailable, Random Forest algorithms achieve decent precision.

The Random Forest algorithm has the following drawbacks

- The biggest drawback of Random Forest algorithms is their complexity.
- Random forests are much more difficult and time-consuming to create than decision trees.
- The Random Forest algorithm necessitates more computing power.
- When the researcher has a huge set of decision trees, it becomes less intuitive.
- In contrast to other algorithms, the prediction process using random forests takes a long time.

2.2. Theoretical model/system used

In this study, the researcher used the random forest (RT) model because random forest as a classifier usually takes less time and holds better efficiency than the other classifiers. The random forest uses an increase in the decision tree classifier regarding a set of trees. The leading cause of applying RF is to predict and determine a disease well (like sepsis) in course of underfitting the huge dataset (Singh et al., 2022).

3.0 METHODOLOGY

3.1 Research methodology

This paper is based on a solution-based approach, focusing on creating a model and an efficient way to predict covid 19 from a patient. The study follows both open-source data collection on the internet. The process was fully participatory ensuring the efficiency of the model. The research is considered a systematic review research methodology. The necessary steps to perform the research are presented below:

Problems of clinical methods

A blood examination or an antibody test may be used to assess if anyone is afflicted with the Virus. A diagnostic examination will diagnose an infection if anyone is already infected. The Federal Drug Administration (FDA) has approved several diagnostic tests for Covid-19, including molecular tests. The FDA-approved molecular studies are accurate, and the findings are returned in a matter of minutes to several days, depending on the test. Another method of diagnostic examination is antigen checking, which uses a swab to see whether viral proteins are found in a sample obtained from within the nose. These experiments are frequently easier and can have findings quicker than certain molecular tests, often within minutes in a doctor's office. The other kind of test is an antibody test, also known as a serology test. When a person is infected with a virus, the body produces antibodies that aid the immune system in fighting the infection. An antibody screening identifies antibodies to the virus using a blood sample. If an antibody examination detects antibodies in the blood, it is likely that the person has previously been infected with the virus. Antibody testing cannot tell you whether you have a new infection and cannot be used to detect a Covid-19 infection.

- **Limitations of molecular test**

A molecular diagnosis can be made quickly and provides highly sensitive, accurate, and usually quantitative identification of the SARS-CoV-2 Virus RNA. So, it is difficult, expensive, and time-consuming to implement. Each DNA-based research package cost more than $100. The processing time is 2 to 3 days until the lab is prepared. The test kit study, on the other hand, requires 2 to 3 hours. Furthermore, the much-hyped ‘molecular diagnostics’ have not been found available to the end-users, but rather to extremely trained diagnostic laboratory staff. Certain immunoanalytical experiments have a strong false-negative rate by using the RT (primer-PCR) and PCR (polymerase chain reaction) processes. Incorrect RT-PCR findings may be affected by poor compilation, handling, transport, distillation, and healthy processing (Afzal, 2020). The existence of the RNA derived from the swabs also influences the results. Other conditions, such as filtered RNA degradation, purification resistance, the involvement of nucleic acid cross-linking reagents, or genomic mutations, may trigger false-negative results. Furthermore, it is critical to note that if blood samples are mishandled during collecting, sorting, and pipetting, false-positive findings will occur. Although the likelihood of these unfavourable outcomes, these diagnostic tests are currently the most accurate, responsive, and easily accessible methods for early and large-scale diagnosis of Serious Acute Respiratory Syndrome-Coronavirus-2 (Afzal, 2020).

- **Limitations of Antigen Test**

These simple antigen tests make obtaining findings simpler, but they are not without limitations. While rapid antigen tests that identify SARS-CoV-2 often focus on the nasopharyngeal specimen, the rapid antigen test is more often based on a context of a large number of patients (European Centre for Disease Prevention (ECDC), 2020). These specimens are now needed for professional sampling and the use of personal protective equipment during sampling and processing. At this time, the self-sampling method has not been statistically proven or confirmed. Rapid antigen analyses, unlike RT-PCR tests, do not provide controls and therefore are less reliable. Since some of the quick antigen samples are processed separately, analyzing vast numbers of specimens at the same time is impossible, and multiplex detection of other respiratory pathogens is currently not possible. Another disadvantage to accelerated antigen tests is that specimens are seldom submitted to public health labs for more characterization, such as sequencing.

- **Limitations of Antibody Test**

A laboratory professional must understand the underlying problem with all serological testing for antibodies, particularly COVID-19 antibodies, which contributes to high analytical error rates. This form of test's reliability is subjective, varied, spontaneous, insidious, and often unreliable. It is caused by different groups and subclasses of distinct antibodies that are produced throughout an activated immune response (in this case COVID-19 infection). The existence of additional eosinophils as well as other types of normal human antibodies complicates allergic reactions. Any of these passes undetected through the body, and if they do, they can interact with or combine with the tested reagent (Ismail, 2020). This suggests that even with a really strong methodology, the right reagents, and the most stringent internal and external measures in operation, there is always a small margin of unreliability. It is deceptive to claim that such a basic and precise tool is as simple and reliable as a pregnancy test. In these samples, inaccuracy attributable to cross-reactivity is not a possibility.

- **Limitations of PCR Test**

The major limitation of RT-PCR research is that it cannot be used to detect the previous infection with SARS-CoV-2, which is essential for recognizing the spread of the infection since pathogens only exist in the body for a limited time. Different techniques are required to identify, monitor, and research previous infections, especially those that may have originated and spread without causing symptoms. When an individual recovers, the virus is eradicated, and these checks will no longer determine whether or not the person was infected.

Analyzing the problem

There are several limitations to detecting the coronavirus inside the human body. The clinical way is much harder and more painful to obtain. This creates a heavy burden on the people of the healthcare system.

Developing a model

Because of these weaknesses, we want to introduce the concept of data mining. And then we will propose a model which will testify to the prediction capabilities.

3.2 Analysis of the Data

Dataset Description

Data is an integral component of every AI model and, in essence, the prime explanation for the current surge in the popularity of machine learning. Because of the availability of data, scalable ML algorithms have become feasible as real goods that can add value to an enterprise, rather than becoming a by-product of the primary processes. A dataset is similar to a database table or a spreadsheet in Microsoft Excel. This is a traditional data framework that is widely used in the area of machine learning. Other types of records, such as photographs, videos, and text, are not considered at this point.

Instance: An instance is a single row of data. It's a domain-specific phenomenon.

Feature: A feature is a single column of results. It is a part of an observation which is often referred to as a data instance attribute. Some features may be predictors in a model, whereas others may be outcomes or features to be projected.

Data type: A feature has a data type. They may have a true or numerical value, as well as a categorical or ordinal value. Lists, days, periods, and more complicated forms may exist, but when dealing with conventional machine-learning approaches, they are usually limited to real or categorical values.

Dataset: A dataset is a list of instances and when dealing with machine learning techniques, it usually requires a few datasets for various purposes.

Secondary data released by (Hungund, 2020) in Kaggle is used for creating the model to detect and predict virus dissemination throughout the COVID-19 epidemic. The dataset combined with data from various sources to include individual-level data rather than composite data as presented by most data warehouses.

3.3 Details about the Dataset

As it has already used the cleaned covid data with symptoms that were provided in the Kaggle (Hungund, 2020). In the cleaned dataset there are 27 columns and 3,16,801 rows with instances. This cleaned data contains all the possible combinations of data from the raw dataset. But this dataset contains dummy variables after combination. The combined data was applied in chatbot, supervised learning, and unsupervised learning. As the dataset had nominal values, the train data and test data had to be scaled using a standard scalar. So, for this paper, we took 2,37,600 instances with 12 columns.

The dataset is divided into tables, with data of String and Numeric types. The dataset also includes categorical variables. Since the ML model allows all data were taken as input to be in numeric form and the dataset all have a numeric data type, it was helpful for the implementation. But only the country column was removed from the dataset (Indhumathi & Kumar, 2022). The null and duplicated values were not found in the dataset.

Each case in the database reflects a person who tested positive for COVID-19 and was collected from various sources. There were initially 3,16,801 instances in this dataset. To secure patients' anonymity, each case is deindividualized and anonymized. The cases are labeled with these:

Country: A list of the countries that the individual has lived in.

Age: According to WHO, each person's age group is classified.

Symptoms: Fever, tiredness, difficulty coughing, dry cough, and sore throat are the five main symptoms of COVID-19, according to the WHO. Pains, Nasal Congestion, Runny Nose, Diarrhea and Other Symptoms were also included in the dataset.

Severity: The degree of seriousness is divided into Mild, Moderate and Serious Severe

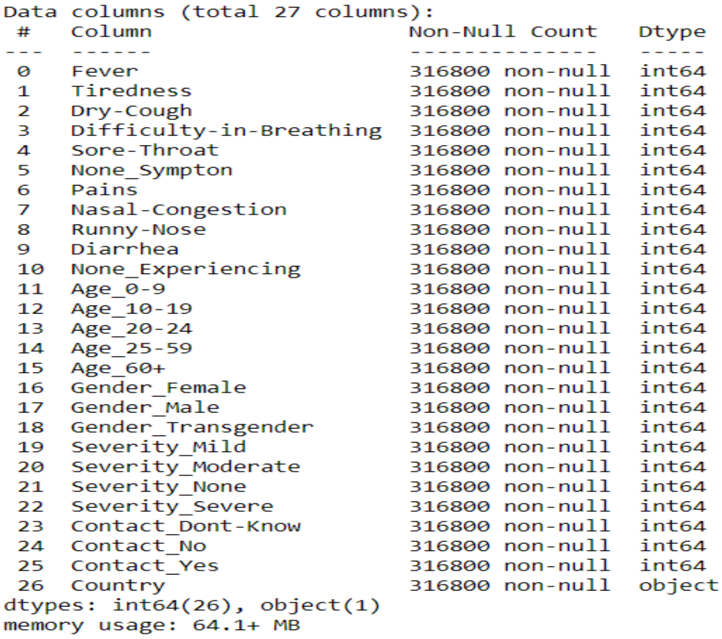
Contact: Has the person approached another COVID-19 patient?

Each variable is listed below and this dataset will be referred to as the Kaggle dataset in the paper. Here the list of columns is stated including datatypes:

As it is seen that there are 27 columns and all the datatypes have integer values and an object. There is a trade-off with the consistency of this dataset and we hoped to reconcile this situation soon. The use of a Kaggle-based dataset was useful due to its large size. Unfortunately, it lacked clear detail about the complications and underlying conditions that each case was dealing with. But this dataset was useful because of its many features, which included various symptoms and ages of the patients infected with covid 19 disease. This dataset was used to train machine learning models to reflect the predictive capabilities of datasets of varying quality.

Insights and descriptive statistics

Fever, tiredness, dry cough, sore throat, body pain, nasal congestion, runny nose, and diarrhea were the recurrent symptoms observed in patients whose data were included in this dataset and are shown in the graph. But there are also data on patients with no symptoms and this was not taken into the consideration.



```
Data columns (total 27 columns):
 #   Column                    Non-Null Count   Dtype 
---  ------                    --------------   ----- 
 0   Fever                     316800 non-null  int64 
 1   Tiredness                 316800 non-null  int64 
 2   Dry-Cough                 316800 non-null  int64 
 3   Difficulty-in-Breathing   316800 non-null  int64 
 4   Sore-Throat               316800 non-null  int64 
 5   None_Sympton              316800 non-null  int64 
 6   Pains                     316800 non-null  int64 
 7   Nasal-Congestion          316800 non-null  int64 
 8   Runny-Nose                316800 non-null  int64 
 9   Diarrhea                  316800 non-null  int64 
 10  None_Experiencing         316800 non-null  int64 
 11  Age_0-9                   316800 non-null  int64 
 12  Age_10-19                 316800 non-null  int64 
 13  Age_20-24                 316800 non-null  int64 
 14  Age_25-59                 316800 non-null  int64 
 15  Age_60+                   316800 non-null  int64 
 16  Gender_Female             316800 non-null  int64 
 17  Gender_Male               316800 non-null  int64 
 18  Gender_Transgender        316800 non-null  int64 
 19  Severity_Mild             316800 non-null  int64 
 20  Severity_Moderate         316800 non-null  int64 
 21  Severity_None             316800 non-null  int64 
 22  Severity_Severe           316800 non-null  int64 
 23  Contact_Dont-Know         316800 non-null  int64 
 24  Contact_No                316800 non-null  int64 
 25  Contact_Yes               316800 non-null  int64 
 26  Country                   316800 non-null  object
dtypes: int64(26), object(1)
memory usage: 64.1+ MB
```

Figure 1: The variables of the dataset including datatype

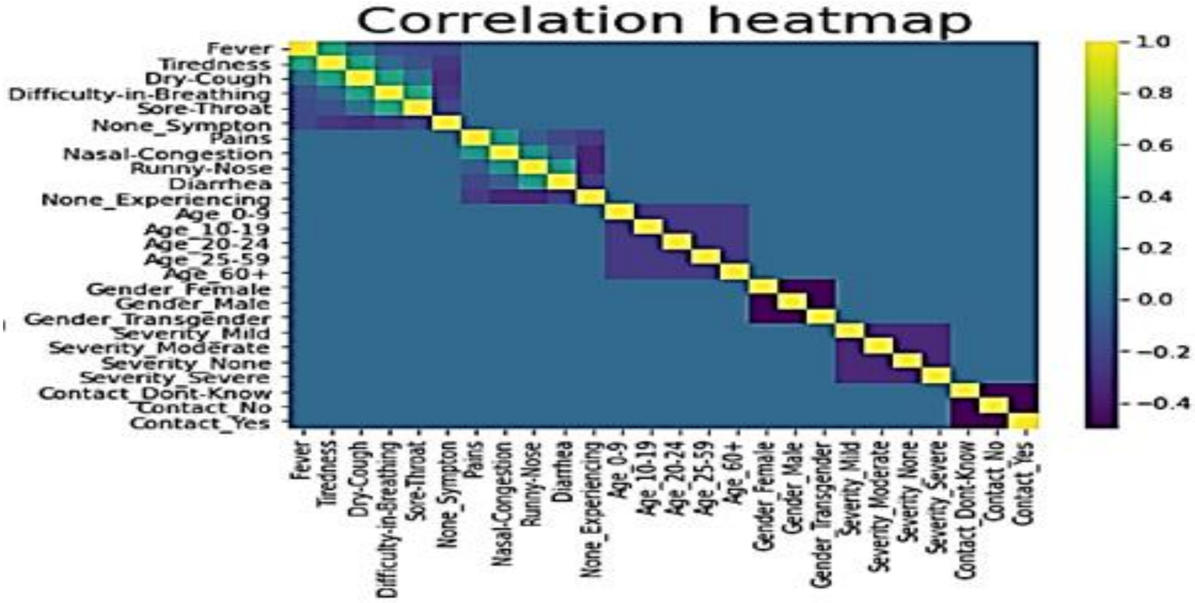


Figure 2: Correlation heatmap of the dataset [Developed by the researcher]

Correlation between dataset features offers important knowledge about the features and their effect on the goal value. Pearson's temperature chart (figure) depicts the association between the dataset's characteristics, which explicitly shows a comparatively stronger positive correlation with the patient's symptoms, whether the patient was aged 0-60 or above. This suggests that the patients have a higher chance of having affected by the disease. There is also a clear positive association between the symptoms.

3.4 Proposed Method

The proposed method consists of different steps. Those steps will be followed in order to effectively find the results of the models. Those results will identify if the person has been infected with covid 19 or not. Cleaning and pre-processing are mandatory for training the dataset to implement as input in the data mining algorithms. From the dataset, features are extracted from the trained data. In this paper, we have identified the correlation heatmap and other analyses. Machine learning algorithms namely decision trees, random forest and logistic regression are implemented. From the implementation, we will get the accuracy of the proposed model.

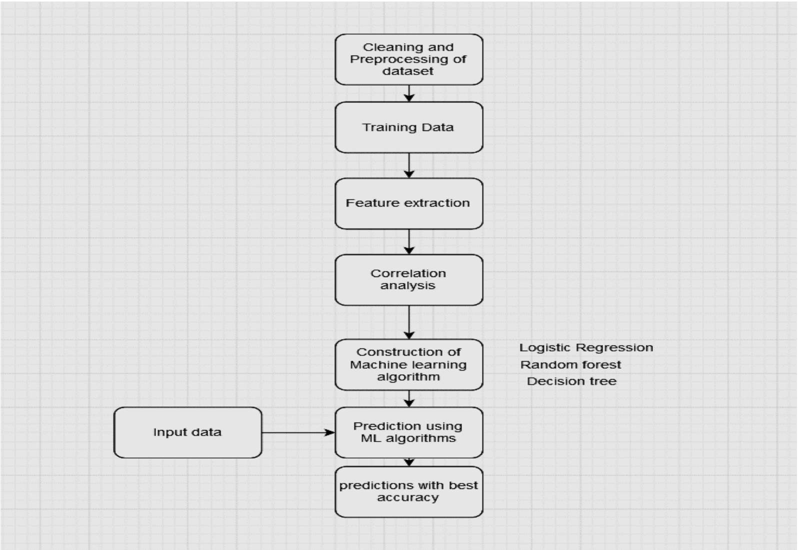

Cleaning and Preprocessing of dataset
Training Data
Feature extraction
Correlation analysis
Construction of Machine learning algorithm
Logistic Regression
Random forest
Decision tree
Input data
Prediction using ML algorithms
predictions with best accuracy

Figure 3: Framework of the proposed method

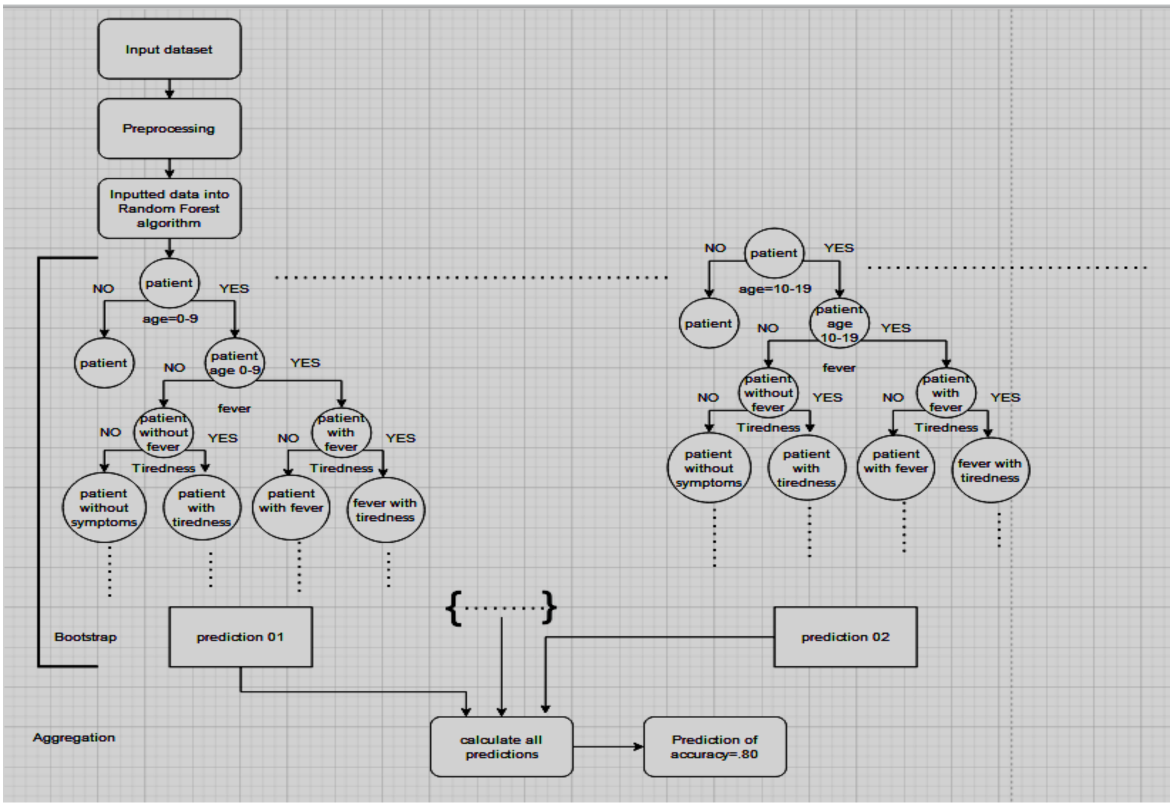


Figure 4: Framework of supervised learning

Here is the detailed process of how the model we proposed come into existence. Random forest algorithm is chosen as a supervised learning model in our proposed model. Random forest will split into a different decision tree and each decision tree will provide a prediction. After calculating all the predictions, an average or most common prediction will provide the best prediction of the model.

3.5 Evaluation of the Study

Evaluation Metrics

The researcher evaluates the models that were implemented and compares the performance of the three classifiers. As the cleaned dataset collected had an issue of having nominal data across the whole dataset, so the accuracy of all three models had a very similar result. The following research aims to reliably forecast the outcome of a specific patient based on a variety of factors, including age and so on. Since this is a critical forecast, accuracy is critical. As a result, for this study's assessment of the model, three evaluation criteria were taken. In the equations, the following expressions are used: TP, True Positive; TN, True Negative; FP, False Positive; and FN, False Negative (Alzahrani & Kanan, 2022).

Accuracy

A dataset of (TP + TN) data points, the precision is proportional to the ratio of the classifier's cumulative accurate predictions sum of the expressions to the entire data points. Accuracy is also known as a significant metric for evaluating the classification model's efficiency (Alzahrani & Kanan, 2022; Singh et al., 2022).

Accuracy is calculated as shown in the equation below:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad 0.0 < accuracy < 1.0 \tag{3}$$

Precision

Precision is defined as the ratio of True Positive samples to the number of TP and FP samples. Precision is another important criterion for correctly identifying patients in an imbalanced class dataset (Kumar et al., 2022; Seo et al., 2021; Singh et al., 2022). The equation for the Precision is stated below:

$$Precision = \frac{TP}{TP+FP} \tag{4}$$

Recall

Recall is equal to the ratio of True Positive samples to the sum of True Positive (TP) and False Negative (FN) samples. It is a crucial criterion for identifying the proportion of patients in an imbalanced class dataset who have been correctly identified out of all the patients who would have been appropriately expected. (Kumar et al., 2022; Seo et al., 2021; Singh et al., 2022).

The recall is calculated as shown in the equation below:

$$Recall = \frac{TP}{TP+FN} \tag{5}$$

F1 score

The F1 Score is proportional to the harmonic mean of the Recall and Precision values. The F1 Score provides an accurate evaluation of the model's performance in categorizing COVID-19 patients by achieving the appropriate balance between Precision and Recall. (Kumar et al., 2022; Seo et al., 2021; Singh et al., 2022).

The F1 score is calculated as shown in the equation below:

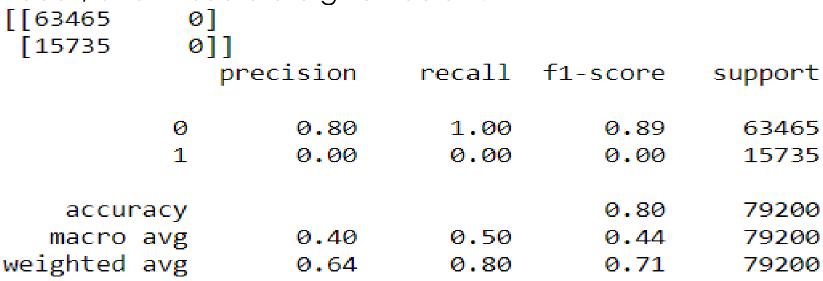
$$F1\,Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

Performance

In this study, the researcher evaluates the performance of the three models which are logistic regression, decision tree, and random forest. As the dataset had 50/50 percent values which caused a problem getting the accuracy. All the models had the same accuracy of 66.85%. So, to improve our model, the researcher considers the accuracy of the models following the age of the patients that are found on the dataset.

a) Prediction for ages between 0-09

The accuracy, recall, and F1 score are given below:



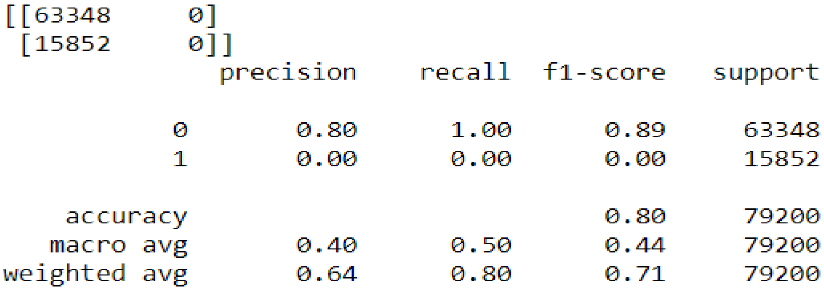
```
[[63465     0]
 [15735     0]]
              precision    recall  f1-score   support

           0       0.80      1.00      0.89     63465
           1       0.00      0.00      0.00     15735

    accuracy                           0.80     79200
   macro avg       0.40      0.50      0.44     79200
weighted avg       0.64      0.80      0.71     79200
```

Figure 5: Accuracy, precision, recall, F1 score of patients aged between 0-9

Random forest algorithms have produced an accuracy of 80.13%,

b) Prediction for ages between 10-19

The accuracy, recall, and F1 score are given below:



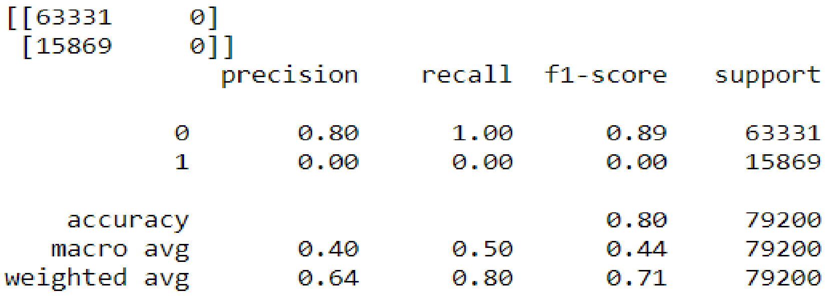
```
[[63348     0]
 [15852     0]]
              precision    recall  f1-score   support

           0       0.80      1.00      0.89     63348
           1       0.00      0.00      0.00     15852

    accuracy                           0.80     79200
   macro avg       0.40      0.50      0.44     79200
weighted avg       0.64      0.80      0.71     79200
```

Figure 6: Accuracy, precision, recall, F1 score of patients aged between 10-19

Even though all three algorithms have produced an accuracy of 79.98%, we will consider taking only the random forest classifiers value. Because of the efficiency of the random forest and as it took less time than other algorithms, we will consider it as the result of the prediction.

c) Prediction for ages between 20-24

The accuracy, recall, and F1 score are given below:



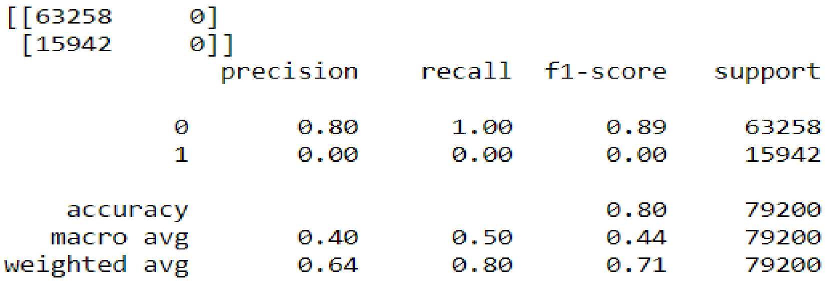
```
[[63331     0]
 [15869     0]]
              precision    recall  f1-score   support

           0       0.80      1.00      0.89     63331
           1       0.00      0.00      0.00     15869

    accuracy                           0.80     79200
   macro avg       0.40      0.50      0.44     79200
weighted avg       0.64      0.80      0.71     79200
```

Figure 7: Accuracy, precision, recall, F1 score of patients aged between 20-24

Even though all three algorithms have produced an accuracy of 79.96%, we will consider taking only the random forest classifiers value. Because of the efficiency of the random forest and as it took less time than other algorithms, we will consider it as the result of the prediction.

d) Prediction for ages between 25-59

The accuracy, recall, and F1 score are given below:



```
[[63258     0]
 [15942     0]]
              precision    recall  f1-score   support

           0       0.80      1.00      0.89     63258
           1       0.00      0.00      0.00     15942

    accuracy                           0.80     79200
   macro avg       0.40      0.50      0.44     79200
weighted avg       0.64      0.80      0.71     79200
```

Figure 8: Accuracy, precision, recall, F1 score of patients aged between 25-59

Even though all three algorithms have produced an accuracy of 79.87%, we will consider taking only the random forest classifiers value. Because of the efficiency of the random forest and as it took less time than other algorithms, the researcher considers it as the result of the prediction.

4.0 RESULANTS

4.1 Results

Here the researcher has discussed the results of the patients according to the age from 0-59 years. After evaluating all the accuracy of the model, it has been found that the patients who have those symptoms have covid cases at an accuracy of up to 80%. So, it seems that based on age, the accuracy of the model shows the patients that he/she has been infected by the coronavirus.

4.2 Discussions

Based on the model implementation, the researcher found that the Random Forest algorithm has provided an efficient solution to the problem. As covid is a problematic disease that is quite hard to identify using the clinical test. Sometimes, in hospitals, it is quite difficult to get the results of covid if they are positive or negative. Villavicencio et al., (2021) used the 'Support Vector machine' and 'Pearson VII Universal Kernel', and got a result with an accuracy of 98.81%. The mean absolute error was 0.012. Ahamad et al., (2020) in the dataset had the predictive features of Covid-19 the same as the ones we have used. The SVM algorithm showed 93% accuracy for patients aged 0-20. XGBoost decision tree had the highest accuracy of 86% to 90% for patients aged 21-96. And also, all other algorithms had at least an accuracy of 80%. Iwendi et al., (2020) have proposed a model, implemented by the 'Random Forest (RF) algorithm and it was boosted by the 'AdaBoost Algorithm'. The algorithm implemented on the model had an accuracy of 86% and the results found that most of the affected patients were between 20 to 70 years of age. Khanday et al., (2020) in the dataset used 212 clinical reports and the reports were labeled into four classes. The results after classifications show that both logistic regression and multinomial Naïve Bayesian classifier had an accuracy of 96.2% in predicting the Covid-19 infection within the persons. Zhang et al., (2021) used the features of RT-PCR test results by relying on baseline demographics, comorbidities, vitals, and lab values.

The approach they had is quite similar to the one we have done but the results changed due to the variables of the dataset. They had more of a clinical approach and implemented an XGBoost classifier in the model. But all in all, both of the results had the same findings of identifying the Covid-19 patients. All of the studies done by the researchers had the best results based on the datasets and models that were implemented by the data mining algorithms. More or less all the study shown in this paper reveals that a few of the data mining algorithms showed better results than the Random Forest algorithm but even though all of them acknowledged that the Random Forest algorithm had better training time for the big dataset. The accuracy of all the models shown in the studies had an accuracy of more than 80%. This proves that Random Forest has better efficiency in predicting Covid-19 in patients. To find the disease beforehand, we used machine learning to get an efficient result and we got a successful result with an accuracy of 80% in predicting covid 19.

5.0 CONCLUSIONS

In the study, the researcher managed to extract the progress and identified some drawbacks. In addition, scaling up the project in future and demonstrating a pathway for a robust implementation phase of this project have been done. Here the researcher has investigated the base of the open-source dataset found on Kaggle. The model created has given an acceptable accuracy rate. And the model proposed is focused on the age of the patients rather than only the symptoms of the patients. Moreover, a part of an efficient way to predict the disease has been successfully developed. The key limitation of the study is to get a proper dataset that is available on the open-source platform. As stated earlier in the paper, this model efficiently worked on the proper dataset, and the dataset found in Kaggle has nominal data that has contributed to the development of the model providing inadequate information on the process. The model proposed by the researcher can be improved if the proper dataset can be found on the open-source platform. As COVId-19 is a new disease that recently has turned into a pandemic. Still, the dataset of COVID-19 is not much available. If a proper medical dataset are found in future, it could be used to get more efficient results. In the paper, the researcher has discussed the problems of predicting COVID-19 disease by using the clinical method and developed an idea to implement the non-clinical way of using a machine learning algorithm to predict patients whether have been infected with the coronavirus or not. Although the proposed model may have some application limitations, this would provide a feasible solution to the present conditions related to Covid-19.

Acknowledgment

I would like to acknowledge the contribution of Prof. Dr. Dip Nandi for his sincere guidance and suggestions in writing this paper. I also express my thanks to ASM Ashiqur Rahman Shovon and Md. Tarek Lehaz Tabid for their technical assistance and cooperation.

REFERENCES

[1] Afzal, A. (2020). Molecular diagnostic technologies for COVID-19: Limitations and challenges. *Journal of Advanced Research*, *26*, 149–159. https://doi.org/10.1016/J.JARE.2020.08.002

[2] Ahamad, M. M., Aktar, S., Rashed-Al-Mahfuz, M., Uddin, S., Liò, P., Xu, H., Summers, M. A., Quinn, J. M. W., & Moni, M. A. (2020). A machine learning model to identify early stage symptoms of SARS-Cov-2 infected patients. *Expert Systems with Applications*, *160*, 113661. https://doi.org/10.1016/J.ESWA.2020.113661

[3] Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *American Statistician*, *46*(3), 175–185. https://doi.org/10.1080/00031305.1992.10475879

[4] Al-Turaiki, I., Alshahrani, M., & Almutairi, T. (2016). Building predictive models for MERS-CoV infections using data mining techniques. *Journal of Infection and Public Health*, *9*(6), 744. https://doi.org/10.1016/J.JIPH.2016.09.007

[5] Alzahrani, A., & Kanan, A. (2022). Machine Learning Approaches for Developing Land Cover Mapping. *Applied Bionics and Biomechanics*, *2022*. https://doi.org/10.1155/2022/5190193

[6] Ayon, S. I., Islam, M. M., & Hossain, M. R. (2022). Coronary Artery Heart Disease Prediction: A Comparative Study of Computational Intelligence Techniques. *IETE Journal of Research*, *68*(4), 2488–2507. https://doi.org/10.1080/03772063.2020.1713916

[7] Cao, R., & Xu, L. (2009). Improved C4.5 algorithm for the analysis of sales. *2009 6th Web Information Systems and Applications Conference, WISA 2009*, 173–176. https://doi.org/10.1109/WISA.2009.36

[8] Desuky, A. S. (2022). Two Enhancement Levels for Male Fertility Rate Categorization Using Whale Optimization and Pegasos Algorithms. In *Advances in Medical Technologies and Clinical Practice* (pp. 234–256). https://doi.org/10.4018/978-1-6684-5092-5.CH011

[9] Diagnosis and treatment protocol for novel coronavirus pneumonia (Trial version 7). (2020). *Chinese Medical Journal*, *133*(9), 1087–1095. https://doi.org/10.1097/CM9.0000000000000819

[10] European Centre for Disease Prevention (ECDC). (2020). *Options for the use of rapid antigen tests for COVID-19 in the EU/EEA and the UK Key messages*.

[11] Everitt, B. S. , L. S. L. M. S. D. (2011). *Miscellaneous Clustering Methods*. 215–255. https://doi.org/10.1002/9780470977811.CH8

[12] Fadugba, S. E., Shaalini, V. J., & Ibrahim, A. A. (2021). Analysis and applicability of a new quartic polynomial one-step method for solving COVID-19 model. *Journal of Physics: Conference Series*, *1734*(1). https://doi.org/10.1088/1742-6596/1734/1/012019

[13] Fan, H., Chen, Y., Huang, S., Zhang, X., Guan, H., & Shi, D. (2019). Post-fault Transient Stability Assessment Based on k-Nearest Neighbor Algorithm with Mahalanobis Distance. *2018 International Conference on Power System Technology, POWERCON 2018 - Proceedings*, 4417–4423. https://doi.org/10.1109/POWERCON.2018.8602125

[14] Ferreira, D., Oliveira, A., & Freitas, A. (2012). Applying data mining techniques to improve diagnosis in neonatal jaundice. *BMC Medical Informatics and Decision Making*, *12*(1), 143. https://doi.org/10.1186/1472-6947-12-143/TABLES/2

[15] Gandhi, R. (2018, May 5). *Naive Bayes Classifier. What is a classifier?* Towards Data Science. https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c

[17] Guan, W., Ni, Z., Hu, Y., Liang, W., Ou, C., He, J., Liu, L., Shan, H., Lei, C., Hui, D. S. C., Du, B., Li, L., Zeng, G., Yuen, K.-Y., Chen, R., Tang, C., Wang, T., Chen, P., Xiang, J., … Zhong, N. (2020). Clinical Characteristics of Coronavirus Disease 2019 in China. *New England Journal of Medicine*, *382*(18), 1708–1720. https://doi.org/10.1056/NEJMOA2002032/SUPPL_FILE/NEJMOA2002032_DISCLOSURES.PDF

[18] Haque, M. R., Islam, M. M., Iqbal, H., Reza, M. S., & Hasan, M. K. (2018). Performance Evaluation of Random Forests and Artificial Neural Networks for the Classification of Liver Disorder. *International Conference on Computer, Communication, Chemical, Material and Electronic Engineering, IC4ME2 2018*. https://doi.org/10.1109/IC4ME2.2018.8465658

[19] Harrison, O. (2018, September 11). *Machine Learning Basics with the K-Nearest Neighbors Algorithm | by |*. Towards Data Science. https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

[20] Hasan, N., Chaudhary, K., & Alam, M. (2021). Unsupervised machine learning framework for early machine failure detection in an industry. *Https://Doi.Org/10.1080/09720529.2021.1951434*, *24*(5), 1497–1508. https://doi.org/10.1080/09720529.2021.1951434

[21] Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(8), 832–844. https://doi.org/10.1109/34.709601

[22]Hungund, B. (2020). *COVID-19 Symptoms Checker*. https://www.kaggle.com/datasets/iamhungundji/covid19-symptoms-checker?select=Cleaned-Data.csv

[23] Hussain, S., Muhammad, L. J., Ishaq, F. S., Yakubu, A., & Mohammed, I. A. (2019). Performance evaluation of various data mining algorithms on road traffic accident dataset. *Smart Innovation, Systems and Technologies*, *106*, 67–78. https://doi.org/10.1007/978-981-13-1742-2_7

[24] Indhumathi, K., & Kumar, K. S. (2022). Seasonal Infectious Disease Prediction based on Electronic Patient Health Records using Boosted Random Forest Algorithms. *2022 2nd*

International Conference on Advance Computing and Innovative Technologies in Engineering, ICACITE 2022, 2025–2032. https://doi.org/10.1109/ICACITE53722.2022.9823453

[25] Islam, M. M., Iqbal, H., Haque, M. R., & Hasan, M. K. (2018). Prediction of breast cancer using support vector machine and K-Nearest neighbors. *5th IEEE Region 10 Humanitarian Technology Conference 2017, R10-HTC 2017, 2018-January*, 226–229. https://doi.org/10.1109/R10-HTC.2017.8288944

[26] Ismail, A. A. A. (2020). Serological tests for COVID-19 antibodies: Limitations must be recognized. *Annals of Clinical Biochemistry: International Journal of Laboratory Medicine*, *57*(4), 274–276. https://doi.org/10.1177/0004563220927053

[27] Iwendi, C., Bashir, A. K., Peshkar, A., Sujatha, R., Chatterjee, J. M., Pasupuleti, S., Mishra, R., Pillai, S., & Jo, O. (2020). COVID-19 patient health prediction using boosted random forest algorithm. *Frontiers in Public Health*, *8*. https://doi.org/10.3389/FPUBH.2020.00357

[28] Jaafar, H., Ramli, N. H., & Abdul Nasir, A. S. (2018). An Improvement To The k-Nearest Neighbor Classifier For ECG Database. *IOP Conference Series: Materials Science and Engineering*, *318*(1), 012046. https://doi.org/10.1088/1757-899X/318/1/012046

[29] Jibril, M. L., & Sharif, U. S. (2020). Power of Artificial Intelligence to Diagnose and Prevent Further COVID-19 Outbreak: A Short Communication. *ArXiv Preprint*. https://doi.org/10.48550/arxiv.2004.12463

[30] Keeling, M. J., Hollingsworth, T. D., & Read, J. M. (2020). Efficacy of contact tracing for the containment of the 2019 novel coronavirus (COVID-19). *Journal of Epidemiology and Community Health*, *74*(10), 861–866. https://doi.org/10.1136/JECH-2020-214051

[31]Kohavi, R., & Quinlan, R. (1999). *Decision tree discovery*. https://www.semanticscholar.org/paper/Decision-tree-discovery-Kohavi-Quinlan/487203d0d87cc706ed90e40d3bc181e5779f1b87

[32] Kumar, V., Lalotra, G. S., & Kumar, R. K. (2022). Improving performance of classifiers for diagnosis of critical diseases to prevent COVID risk. *Computers & Electrical Engineering*, *102*, 108236. https://doi.org/10.1016/J.COMPELECENG.2022.108236

[33] Lai, C. C., Shih, T. P., Ko, W. C., Tang, H. J., & Hsueh, P. R. (2020). Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges. *International Journal of Antimicrobial Agents*, *55*(3), 105924. https://doi.org/10.1016/J.IJANTIMICAG.2020.105924

[34] Li, Y. C., Bai, W. Z., & Hashikawa, T. (2020). The neuroinvasive potential of SARS-CoV2 may play a role in the respiratory failure of COVID-19 patients. *Journal of Medical Virology*, *92*(6), 552–555. https://doi.org/10.1002/JMV.25728

[35] Mahamunkar, G. S., & Netak, L. D. (2022). Comparison of Various Deep CNN Models for Land Use and Land Cover Classification. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *13184 LNCS*, 499–510. https://doi.org/10.1007/978-3-030-98404-5_46

[36] Majali, J., Niranjan, R., Phatak, V., & Tadakhe, O. (2015). Data Mining Techniques For Diagnosis And Prognosis Of Cancer. *IJARCCE*, *4*(3), 613–615. https://doi.org/10.17148/IJARCCE.2015.43147

[37] Muhammad, L. J., Abba Haruna, A., Mohammed, I. A., Abubakar, M., Badamasi, B. G., & Musa Amshi, J. (2019). Performance evaluation of classification data mining algorithms on coronary artery disease dataset. *2019 9th International Conference on Computer and Knowledge Engineering, ICCKE 2019*, 1–5. https://doi.org/10.1109/ICCKE48569.2019.8964703

[38] Muhammad, L. J., Islam, M. M., Usman, S. S., & Ayon, S. I. (2020). Predictive Data Mining Models for Novel Coronavirus (COVID-19) Infected Patients' Recovery. *SN Computer Science*, *1*(4). https://doi.org/10.1007/S42979-020-00216-W

[39] Ng, E. (2019). *Validation of a Protein Biomarker Panel for Early Hepatocellular Carcinoma Detection at the Point-Of-Care* [Doctoral Thesis, Stanford University]. http://purl.stanford.edu/wq623yg2914

[40] Pan, L., Mu, M., Yang, P., Sun, Y., Wang, R., Yan, J., Li, P., Hu, B., Wang, J., Hu, C., Jin, Y., Niu, X., Ping, R., Du, Y., Li, T., Xu, G., Hu, Q., & Tu, L. (2020). Clinical characteristics of COVID-19 patients with digestive symptoms in Hubei, China: A descriptive, cross-sectional, multicenter study. *American Journal of Gastroenterology*, *115*(5), 766–773. https://doi.org/10.14309/AJG.0000000000000620

[41] Perkel, J. M. (2018). Why Jupyter is data scientists' computational notebook of choice. *Nature*, *563*(7729), 145–146. https://doi.org/10.1038/D41586-018-07196-1

[42] Pounis, G. (2018). Statistical analysis of retrospective health and nutrition data. *Analysis in Nutrition Research: Principles of Statistical Methodology and Interpretation of the Results*, 103–144. https://doi.org/10.1016/B978-0-12-814556-2.00005-1

[43] Prasad, R., Anjali, P., Adil, S., & Deepa, N. (2019). Heart disease prediction using logistic regression algorithm using machine learning. *International Journal of Engineering and Advanced Technology*, *8*(3 Special Issue), 659–662.

[44] Rothe, C., Schunk, M., Sothmann, P., Bretzel, G., Froeschl, G., Wallrauch, C., Zimmer, T., Thiel, V., Janke, C., Guggemos, W., Seilmaier, M., Drosten, C., Vollmar, P., Zwirglmaier, K., Zange, S., Wölfel, R., & Hoelscher, M. (2020). Transmission of 2019-nCoV Infection from an Asymptomatic Contact in Germany. *New England Journal of Medicine*, *382*(10), 970–971. https://doi.org/10.1056/NEJMC2001468/SUPPL_FILE/NEJMC2001468_DISCLOSURES.PDF

[45] Saeed, A. A. (2021). *Predictions of a-Decay Half-Lives for Neutron-Deficient Nuclei with the Aid of Machine Learning*. Kwara State University.

[46] Seo, S., Kim, Y., Han, H. J., Son, W. C., Hong, Z. Y., Sohn, I., Shim, J., & Hwang, C. (2021). Predicting Successes and Failures of Clinical Trials With Outer Product–Based Convolutional Neural Network. *Frontiers in Pharmacology*, *12*, 1423. https://doi.org/10.3389/FPHAR.2021.670670/BIBTEX

[47] Singh, Y. V., Singh, P., Khan, S., & Singh, R. S. (2022). A Machine Learning Model for Early Prediction and Detection of Sepsis in Intensive Care Unit Patients. *Journal of Healthcare Engineering*, *2022*. https://doi.org/10.1155/2022/9263391

[48] Venkatalakshmi, B., & Shivsankar, M. V. (2014). Heart Disease Diagnosis using Predictive DataMining. *International Journal of Innovative Research in Science, Engineering and Technology*, *3*(3), 1873–1877. http://www.cs.waikato.ac.nz/ml/weka

[49] Villavicencio, C. N., Macrohon, J. J. E., Inbaraj, X. A., Jeng, J. H., & Hsieh, J. G. (2021). Covid-19 prediction applying supervised machine learning algorithms with comparative analysis using weka. *Algorithms*, *14*(7). https://doi.org/10.3390/A14070201

[50] Wang, G.-Q., Zhao, L., Wang, X., Jiao, Y.-M., & Wang, F.-S. (2021). Diagnosis and Treatment Protocol for COVID-19 Patients (Tentative 8th Edition): Interpretation of Updated Key Points. *Infectious Diseases & Immunity*, *1*(1), 17. https://doi.org/10.1097/ID9.0000000000000002

[51] Wölfel, R., Corman, V. M., Guggemos, W., Seilmaier, M., Zange, S., Müller, M. A., Niemeyer, D., Jones, T. C., Vollmar, P., Rothe, C., Hoelscher, M., Bleicker, T., Brünink, S., Schneider, J., Ehmann, R., Zwirglmaier, K., Drosten, C., & Wendtner, C. (2020). Virological assessment of hospitalized patients with COVID-2019. *Nature 2020 581:7809*, *581*(7809), 465–469. https://doi.org/10.1038/s41586-020-2196-x

[52] World Health Organization (WHO). (2022, November 11). *WHO Coronavirus (COVID-19) Dashboard*. https://covid19.who.int/

[53] Yahaya, B. Z., Muhammad, L. J., Abdulganiyyu, N., Ishaq, F. S., & Atomsa, Y. (2018). An Improved C4.5 Algorithm Using L' Hospital Rule for Large Dataset. *Indian Journal of Science and Technology*, *11*(47), 1–5. https://doi.org/10.17485/IJST/2018/V11I47/132538

[54] Yi, C., Zhen, J., Li, Y., Yi, Y., Yin, P., & Min, H. (2018). A novel method to improve transfer learning based on mahalanobis distance. *2017 IEEE International Conference on Robotics and Biomimetics, ROBIO 2017*, *2018-January*, 2279–2283. https://doi.org/10.1109/ROBIO.2017.8324758

[55] Zhang, J., Jun, T., Frank, J., Nirenberg, S., Kovatch, P., & Huang, K. lin. (2021). Prediction of individual COVID-19 diagnosis using baseline demographics and lab data. *Scientific Reports 2021 11:1*, *11*(1), 1–8. https://doi.org/10.1038/s41598-021-93126-7