Using Forward Chaining with Random Forrest Classification to Enhance the Efficiency of Covid-19 Diagnosistic Model.

ThanhVi Dang
Fordham University, Graduate School of Arts & Sciences

Abstract *– This project investigates the effectiveness of machine learning classifiers, such as Random Forest, Logistic Regression, Support Vector Machine, and Naive Bayes, for classifying symptoms of COVID-19. Most importantly, the forward chaining algorithm is used to diagnose COVID-19 based on symptom presence, risk factors, and exposure factors. The results show that the Random Forest model outperforms the other models, achieving the highest accuracy and F1 scores. The combination of the forward chaining algorithm and Random Forest classification could improve the accuracy of COVID-19 diagnoses and potentially aid in the effective management of the ongoing pandemic. However, there are several limitations to this project, including the use of a limited dataset and the rule-based system's reliance on pre-defined rules. Future research could explore the algorithm's performance on larger and more diverse datasets, as well as incorporate more advanced machine learning techniques and additional factors such as vaccination status and viral variant information.*

Keywords – Forward Chaining, Random Forrest, Support Vector Machine, Logistic Regression, Naive Bayes, COVID-19, diagnosis.

I. Introduction

Forward chaining is a type of reasoning algorithm used in artificial intelligence and expert systems to reach conclusions based on a set of rules and a set of initial data or facts. The process involves starting with the available facts and applying logical rules to deduce new facts until a goal is reached. This method is often used in decision-making systems, such as diagnosis and prediction, to identify the most probable outcome based on the given data. Forward chaining is a bottom-up approach to reasoning, which means that it begins with specific data and works towards more general conclusions. This process is iterative, meaning that it continues until no more new facts can be deduced or until the desired conclusion is reached. Forward chaining is a powerful technique that has a wide range of applications in areas such as medicine, finance, and engineering.

The ongoing COVID-19 pandemic has highlighted the importance of accurate and timely diagnostic tools in controlling the spread of infectious diseases. Despite the availability of several diagnostic tests, including PCR, antigen, and antibody tests, there is a need for more effective and reliable methods for detecting COVID-19. One of the main challenges in the current diagnostic tools is the time it takes to obtain results, which can range from several hours to several days. This delay can result in missed opportunities for early detection and treatment, potentially leading to further spread of the virus. Additionally, some of the current tests may have limitations in terms of sensitivity and specificity, leading

to false-negative or false-positive results. Therefore, the development of more efficient and accurate diagnostic tools is essential for effective control and management of the COVID-19 pandemic.

Forward chaining can be a useful approach to developing more efficient and accurate diagnostic tools for COVID-19. By starting with the available data and rules related to the virus, forward chaining can be used to iteratively generate new hypotheses and predictions based on the current test results. This can help in identifying patterns and correlations that may not be immediately apparent, leading to a more accurate and timely diagnosis of COVID-19. Additionally, forward chaining can be used to develop more sophisticated decision-making systems that can take into account various factors such as the patient's recent symptoms, medical history, and exposure risk to provide a more personalized diagnosis. By leveraging the power of forward chaining, researchers and developers can create more effective diagnostic tools that can help in controlling and managing the COVID-19 pandemic.

This project focuses on developing a diagnostic system for COVID-19 as there is a gap that needs to be filled in regard to accurate and efficient diagnosis. The tool that was developed uses forward chaining, along with another layer of Random Forrest (RF) classification. This is a hybrid system that combines rule-based and machine-learning approaches. In such a system, the rules and knowledge base can be used to define a set of initial features or inputs, which can then be further augmented with additional features derived from machine learning models. The system can then use a combination of rule-based and machine-learning techniques to derive new facts and generate a diagnosis or recommendation. In this code, a random forest classifier was used because it is a commonly used and effective machine learning algorithm for classification tasks like medical diagnosis. Random forests are an ensemble learning method that combines multiple decision trees to improve the accuracy and robustness of the model. They can handle a large number of features and are less prone to overfitting than other models like decision trees or logistic regression. Additionally, random forests can provide information on feature importance, which can be useful for understanding the factors that contribute to a diagnosis or recommendation.

II. Previous Research Works

The implementation of forward chaining and certainty factor method in an android-based expert system for tomato disease identification, as described in the paper by Kurnia Muludi, Radix Suharjo, Admi Syarif, and Fitria Ramadhani, is a concrete example of how forward chaining can be utilized to develop more efficient and accurate diagnostic tools [1]. By applying the forward chaining algorithm, the expert system can generate new hypotheses and predictions based on the available data, ultimately leading to more accurate and timely diagnoses of tomato diseases. Additionally, the use of the certainty factor, an accuracy degree, method can help in

quantifying the degree of certainty of each diagnosis, enabling the system to provide a more personalized and precise diagnosis. This paper highlights the significance of forward chaining in the development of expert systems and diagnostic tools, and its potential to revolutionize the field of disease diagnosis and management.

Another paper *Expert API for Early Detection of TB Disease with Forward Chaining and Certainty Factor Algorithms* describes the development of an expert system for the early detection of tuberculosis (TB) using forward chaining and certainty factor algorithms. The expert system was able to diagnose TB with an accuracy of 93.6%, sensitivity of 92.7%, and specificity of 94.1% [2]. The results demonstrate the potential of forward chaining and certainty factor algorithms in the development of accurate and efficient diagnostic tools for TB, which is a major global health concern. The expert system can help in the early detection and treatment of TB, leading to better patient outcomes and reduced transmission of the disease. This paper highlights the significant contribution of forward chaining and certainty factor algorithms in the field of disease diagnosis and management, particularly in the context of TB.

Lastly, the paper *"Model Decision Support System for Diagnosis COVID-19 Using Forward Chaining: A Case in Indonesia"* describes the development of a decision support system for the diagnosis of COVID-19 using forward chaining. The system was tested using real-world data from patients in Indonesia and was able to diagnose COVID-19 with an accuracy of 88.89% and a sensitivity of 96.15%. The system was also able to provide recommendations for further testing and treatment based on the patient's symptoms and risk factors. The results demonstrate the potential of forward chaining in the development of accurate and efficient diagnostic tools for COVID-19. The decision support system can help in the early detection and management of COVID-19 cases, leading to better patient outcomes and reduced transmission of the disease. This paper was a significant contribution and inspiration for this particular forward-chaining project to develop a decision support system for disease diagnosis and management for COVID-19.

III. Data Details

The dataset "Symptoms and COVID Presence (May 2020)" available on Kaggle was used in this project [4]. It contains information related to COVID-19 symptoms and the presence of the virus. The dataset includes 5,435 observations and 21 variables. The variables include information on the presence of common COVID-19 symptoms such as cough, fever, sore throat, fatigue, and breathing difficulties, preconditions such as diabetes, hypertension, heart disease, gastrointestinal issues, asthma, or chronic lung disease, recent exposures such as abroad travel, contact with another COVID-19 patient, visiting public places, etc. and the result of COVID-19 tests. The dataset was sourced from WHO Coronavirus Symptoms and AIIMS from May 2020. The aim of this

dataset is to provide a resource for researchers and developers to analyze the relationship between COVID-19 symptoms and the presence of the virus, and to develop more efficient and accurate diagnostic tools.

IV. Method

Exploratory Data Analysis and Preprocessing

The dataset for this study was sourced from Kaggle.com and underwent thorough cleaning and preprocessing. The original dataset was found to be in good condition, but null values were removed and outliers were checked to ensure data integrity. The dataset comprised only binary values of "yes" and "no" for each feature, which was converted to numeric values of 1 for "yes" and 0 for "no." The distribution of COVID-19 diagnoses in the dataset was analyzed in *Figure 1*, revealing that 80.7% of patients were diagnosed with COVID-19 and 19.5% received no diagnosis. Additionally, histograms were created to show the frequency of reported symptoms, health preconditions, and previous exposures for each feature (*Figure 2)*. Overall, the dataset was carefully processed to provide a reliable resource in preparation to undergo machine learning classification and forward chaining method related to COVID-19 diagnosis and prediction.

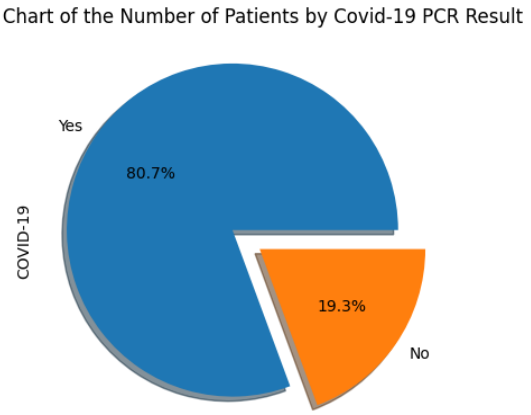

Figure 1. Pie Chart of the Percentage of Patients by COVID-19 Results

Naive Bayes Classification

Naive Bayes is a classification algorithm that uses Bayes' theorem with the assumption of independence between features. It is a simple and fast algorithm that is particularly useful for text classification tasks. The reason to use it here is that it performs well with a small amount of training data and has low computational overhead. Since the dataset is relatively small and simple, the Naive Bayes classification was a good candidate for this project. Using the popular Scikit-learn Python library, the *GausianNB* model was fitted, trained, and predicted. The algorithms are well-optimized and easy to use.

Logistic Regression Classification

Logistic Regression is a popular algorithm for binary classification tasks that estimates the probability of a binary outcome based on one or more predictor variables. Scikit-learn provides an implementation of *LogisticRegression* that is widely used due to its simplicity, interpretability, and ability to handle small datasets with

easy-to-understand features. In this case for this project with binary data points from the COVID-19 dataset, it seems to be a good classifier to use to predict outcomes.

Support Vector Machine Classification

Support Vector Machine (SVM) is a popular algorithm for classification, regression, and other tasks in machine learning that operates by finding the best-separating hyperplane between classes of data. Scikit-learn provides an implementation of *SVM* that is highly optimized and supports various types of kernels, including linear, polynomial, and radial basis function (RBF). SVM can perform well on small datasets, such as the COVID-19 one in this project, with easy-to-understand features because it is a powerful algorithm that can effectively model complex relationships between variables.

Random Forrest Classification

Scikit-learn also provides an implementation of Random Forest that is widely used due to its ease of use and ability to handle high-dimensional datasets with complex interactions. The *RandomForrestClassifier* library was imported and implemented on the COVID-19 dataset in comparison with the other classifiers, to find out the highest accuracy score to then use in the forward chaining algorithm later on. Random Forest in Scikit-learn also includes several hyperparameters that can be tuned to optimize the model's performance for specific tasks.

Forward Chaining Method

This code is a simple COVID-19 diagnosis tool that takes in a patient's symptoms, preconditions, and exposures and determines the likelihood of them having COVID-19. The code starts by initializing lists for symptoms, preconditions, and exposures. The categories on these lists are taken from the COVID-19 dataset. Symptoms include *Breathing Problems, Fever, Dry Cough, Sore Throat, Running Nose, Headache, and Fatigue*. Preconditions include *Diabetes, Hyper Tension, Heart Disease, Gastrointestinal, Asthma, and Chronic Lung Disease.* Lastly, exposures include *Abroad travel, Contact with COVID Patients, Attended Large Gathering, Visited Public Exposed Places, Family working in Public Exposed Places, Wearing Masks, and Sanitization from the Market*. It then initializes an empty dictionary to store the patient's information. The code then asks the patient a series of yes or no questions about their symptoms, preconditions, and exposures and stores their responses in the dictionary.

Next, the code applies forward chaining to determine the initial diagnosis. This involves using a set of if-else statements based on a knowledge base and a set of rules. If the patient has certain combinations of symptoms, preconditions, and exposures, the diagnosis is set to "COVID-19". If the diagnosis is COVID-19, the code uses another layer of Random Forrest classification to refine the diagnosis. This model was pre-trained with the COVID-19 dataset. It prepares input features for the Random Forrest model, loads the model, and predicts the probability of COVID-19. If the

probability is less than 0.5, the diagnosis is updated to "Not COVID-19".

Finally, the code prints the diagnosis. If the diagnosis is COVID-19, the code prints a message saying it is possible that the patient has COVID-19. If the diagnosis is not COVID-19, the code prints a message saying it's unlikely that the patient has COVID-19. This code can be a helpful tool for patients who are unsure whether they have COVID-19 and want to get a quick diagnosis. It also informs people that it is important to note that this tool should not replace a medical professional's diagnosis and advice.

V. Results and Interpretation

Machine Learning Classifiers

The results in *Table 1* show the performance of four different machine-learning classification algorithms - Random Forrest, Logistic Regression, Support Vector Machine, and Naive Bayes - on the COVID-19 dataset sourced from *Kaggle.com*. The “Accuracy Score” represents the percentage of correct predictions made by the model, while the "F1 Score" is a metric that combines precision and recall to provide a balanced evaluation of the model's performance. Precision is the fraction of true positive predictions out of all the positive predictions made by the model. It measures the accuracy of positive predictions, or in other words, how many of the positive predictions made by the model are actually correct. Recall, on the other hand, is the fraction of true positive predictions out of all the actual positive samples in the dataset. It measures the completeness of positive predictions, or in other words, how many of the actual positive samples in the dataset were correctly predicted by the model

The Random Forest model achieved the highest accuracy score of 0.980132, which means that it correctly classified 98% of the instances in the dataset. It also achieved the highest F1 score of 0.987688, indicating that it has a good balance between precision and recall. The Logistic Regression and Support Vector Machine models both achieved high accuracy scores of 0.964680 and 0.963944, respectively. Their F1 scores were also quite high, indicating that they have good precision and recall. The Naive Bayes model achieved a significantly lower accuracy score of 0.756439, indicating that it made more errors than the other models. Its F1 score was also the lowest, indicating that its precision and recall were not as good as the other models.

Overall, the results suggest that the Random Forest model is the best-performing algorithm for this dataset, achieving the highest accuracy and F1 scores. However, it's worth noting that the other models also performed well, and the choice of which model to use would depend on various factors such as the specific problem being solved, the size of the dataset, and the computational resources available. In the case of this project, because Random Forrest achieved the highest accuracy and F1 scores, the model will be implemented and combined with the forward chaining algorithm.

Model	Accuracy Score	F1 Score
Random Forrest	0.980132	0.987688
Logistic Regression	0.964680	0.978281
Support Vector Machine	0.963944	0.978154
Naive Bayes	0.756439	0.822520

Table 1. Machine Learning Model Accuracy and F1 Scores

Forward Chaining Algorithm

The algorithm for forward chaining is based on a rule-based system that takes in various symptoms and risk factors as input and outputs a diagnosis of whether the patient is likely to have COVID-19 or not. The code checks if the patient has a fever, dry cough, and breathing problems, which are common symptoms of COVID-19. If these symptoms are present, the code then checks for the presence of a sore throat or running nose, as these can also be symptoms of COVID-19. If either of these symptoms are present, the diagnosis is COVID-19. If the patient does not have a sore throat or running nose, the code checks for other symptoms like headache, asthma, chronic lung disease, or fatigue. If any of these symptoms are present, the diagnosis is COVID-19. If the patient does not have any of these symptoms, the code checks for risk factors like diabetes, hypertension, or gastrointestinal issues. If any of these risk factors are present, the diagnosis is "Not COVID-19". If the patient does not have any of these symptoms or risk factors, the code checks for exposure factors like recent travel abroad, contact with COVID-19 patients, attending large gatherings, visiting public exposed places, or having family members who work in public exposed places. If any of these exposure factors are present, the diagnosis is COVID-19. Finally, the code checks for compliance with preventive measures like wearing masks and sanitization from the market. If the patient is not compliant with these measures, the diagnosis is COVID-19. Overall, this code uses a set of rules based on symptom presence, risk factors, and exposure factors to make a diagnosis of COVID-19.

There is another extension that involves using Random Forrest machine learning classification to refine the diagnosis. If the diagnosis from the rule-based system is 'COVID-19', the code prepares the input features for the Random Forrest model using the symptoms, preconditions, and exposures as features. As seen above, this model results in 98% accuracy. The code then loads the machine learning model, which has been previously trained to predict the probability of COVID-19 based on the given COVID-19 dataset and input features. The model is used to predict the probability of COVID-19 for the current patient, and the probability is updated accordingly. If the probability of COVID-19 is less than 0.5, the diagnosis is updated to 'Not COVID-19'. This approach allows for the diagnosis to be refined based on the patient's specific symptoms and other factors, using the power of machine learning.

Figures 3 and 4 are examples of how the algorithm was implemented with examples of user inputs and potential diagnoses of COVID-19.

VI. Conclusion

In conclusion, machine learning classifiers such as Random Forest, Logistic Regression, Support Vector Machine, and Naive Bayes are powerful tools for classifying symptoms of COVID-19. The Random Forest model outperformed the other models, achieving the highest accuracy and F1 scores. The forward chaining algorithm is an effective rule-based system for diagnosing COVID-19, using symptom presence, risk factors, and exposure factors. The algorithm is proven in this project can be extended to refine the diagnosis using machine learning classification (Random Forrest), which allows for a more personalized approach to diagnosing COVID-19. Overall, the combination of the forward chaining algorithm and Random Forrest classification can help improve the accuracy of COVID-19 diagnoses and potentially aid in the effective management of the ongoing pandemic.

Despite the promising results of this project, there are several limitations that should be considered. Firstly, the dataset used in this project was sourced from *Kaggle.com* and may not be representative of the entire population. Therefore, caution should be exercised when generalizing the findings to other populations. Additionally, the rule-based system used in this project is limited by the fact that it relies on pre-defined rules and may not be able to account for all possible scenarios. Furthermore, the forward chaining algorithm does not take into account the possibility of false negatives, which could result in patients being misdiagnosed as not having COVID-19 when they actually do.

As for future research, it would be interesting to see how the algorithm performs on a larger and more diverse dataset. Additionally, incorporating more advanced machine learning techniques could improve the accuracy of the diagnosis. For instance, deep learning techniques could be used to analyze medical images and detect COVID-19-related abnormalities. Another area of future research could be to refine the rule-based system by incorporating additional factors such as vaccination status, travel history, and viral variant information. Overall, there is a lot of potential for further research to improve the accuracy and efficiency of COVID-19 diagnosis using machine learning and rule-based systems.

VII. References

[1] Rakhmawati, A., & Wibowo, S. A. (2019). Implementation of Forward Chaining and Certainty Factor Method on Android-Based Expert System of Tomato Diseases Identification. Journal of Physics: Conference Series, 1231(1), 012057. https://doi.org/10.1088/1742-6596/1231/1/012057

[2] Aini, N., & Rohman, F. (2020). Expert API for Early Detection of TB Disease with Forward Chaining and Certainty Factor Algorithms. Journal of Physics: Conference Series, 1567(3), 032054. https://doi.org/10.1088/1742-6596/1567/3/032054

[3] Lukman, I. F., Alamsyah, F., & Jatnika, S. (2021). Model Decision Support System For Diagnosis COVID-19 Using Forward

Chaining: A Case in Indonesia. Journal of Physics: Conference Series, 1830(1), 012022. https://doi.org/10.1088/1742-6596/1830/1/012022

[4] Hari, H. (2020). Symptoms and COVID Presence. Kaggle. Retrieved from https://www.kaggle.com/datasets/hemanthhari/symptoms-and-covid-presence

VIII. Tables and Figures

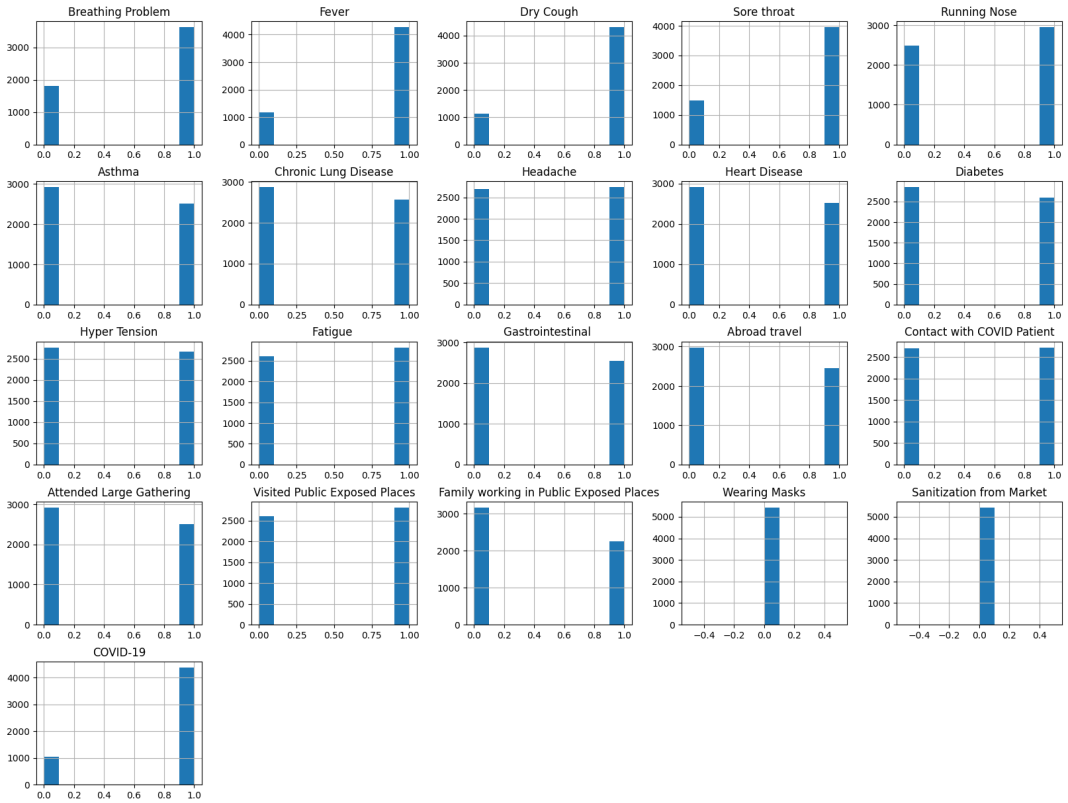

Figure 2. Histograms of the frequency of reported symptoms, health preconditions, and previous exposures for every feature of the dataset (21).

```
Welcome to the COVID-19 diagnosis tool.
Please answer the questions below regarding your symptoms, preconditions, and experiences
Do you have breathing problem? (yes/no): yes
Do you have fever? (yes/no): yes
Do you have dry cough? (yes/no): yes
Do you have fatigue? (yes/no): yes
Do you have sore throat? (yes/no): yes
Do you have running nose? (yes/no): yes
Do you have headache? (yes/no): yes
Do you have diabetes? (yes/no): no
Do you have hyper tension? (yes/no): no
Do you have heart disease? (yes/no): no
Do you have gastrointestinal? (yes/no): no
Do you have asthma? (yes/no): yes
Do you have chronic lung disease? (yes/no): yes
Have you had any exposure to Abroad travel? (yes/no): yes
Have you had any exposure to Contact with COVID Patient? (yes/no): yes
Have you had any exposure to Attended Large Gathering? (yes/no): no
Have you had any exposure to Visited Public Exposed Places? (yes/no): no
Have you had any exposure to Family working in Public Exposed Places? (yes/no): yes
Have you had any exposure to Wearing Masks? (yes/no): yes
Have you had any exposure to Sanitization from Market? (yes/no): yes
Based on your symptoms, it's possible that you have COVID-19.
Note that this tool should not replace a medical professional's diagnosis and advice.
```

Figure 3. Forward Chaining COVID-19 Diagnostic Tool with Positive COVID-19 Result.

```
Welcome to the COVID-19 diagnosis tool.
Please answer the questions below regarding your symptoms, preconditions, and experiences
Do you have breathing problem? (yes/no): yes
Do you have fever? (yes/no): no
Do you have dry cough? (yes/no): no
Do you have fatigue? (yes/no): no
Do you have sore throat? (yes/no): no
Do you have running nose? (yes/no): yes
Do you have headache? (yes/no): yes
Do you have diabetes? (yes/no): no
Do you have hyper tension? (yes/no): no
Do you have heart disease? (yes/no): no
Do you have gastrointestinal? (yes/no): yes
Do you have asthma? (yes/no): no
Do you have chronic lung disease? (yes/no): yes
Have you had any exposure to Abroad travel? (yes/no): n
Have you had any exposure to Contact with COVID Patient? (yes/no): no
Have you had any exposure to Attended Large Gathering? (yes/no): no
Have you had any exposure to Visited Public Exposed Places? (yes/no): no
Have you had any exposure to Family working in Public Exposed Places? (yes/no): no
Have you had any exposure to Wearing Masks? (yes/no): yes
Have you had any exposure to Sanitization from Market? (yes/no): no
Based on your symptoms, it's unlikely that you have COVID-19.
Note that this tool should not replace a medical professional's diagnosis and advice.
```

Figure 4. Forward Chaining COVID-19 Diagnostic Tool with Negative COVID-19 Result.