# COVID-19 Symptoms Detection Using Iterative Dichotomiser (ID3) Algorithm: A Greedy Method

Hassan Adamu
Binyaminu Usman Polytechnic, Hadejia, Jigawa State, Nigeria
**17** PUBLICATIONS **15** CITATIONS
SEE PROFILE

Oluwaseyi Jaiyeoba
Purdue University
**2** PUBLICATIONS **0** CITATIONS
SEE PROFILE

Jamilu Awwalu
The Federal University Dutse
**18** PUBLICATIONS **95** CITATIONS
SEE PROFILE

# COVID-19 Symptoms Detection Using Iterative Dichotomiser (ID3) Algorithm: A Greedy Method

*

Hassan Adamu
*School of Computer Sciences*
*Universiti Sains Malaysia*
Penang, Malaysia
hasanadamu@student.usm.my

Azman Samsudin
*School of Computer Sciences*
*Universiti Sains Malaysia*
Penang, Malaysia
azman.samsudin@usm.my

Oluwaseyi Jaiyeoba
*School of Computer Sciences*
*Universiti Sains Malaysia*
Penang, Malaysia
Jaiyeoba@student.usm.my

Jamilu Awwalu
*Federal University Dutse (FUD)*
*Jigawa State, Nigeria*
jamilu.awwalu@fud.edu.ng

***Abstract*—Iterative Dichotomiser 3 (ID3) algorithm is a popular greedy algorithm that has been applied in solving different types of classification and prediction problems. With the current COVID-19 pandemic, authorities are pushed to the limit on resource allocation and laboratory test administration for suspected cases. In this study, an ID3 model is created using the holdout and cross-validation approaches to classify potential COVID-19 cases based on presented symptoms. The two approaches in the ID3 model are compared based on accuracy and time complexity. Findings from the study show that the 50:50 split holdout approach achieved the overall best accuracy across both the holdout and cross-validation approaches, while the cross-validation had the highest time complexity.**

***Index Terms*—Greedy algorithm, Iterative Dichotomiser (ID3), COVID-19 symptoms classification**

## I. Introduction

According to Güttinger [1], Greedy algorithm has been widely used to quickly find approximate solutions to combinatorial optimization problems. It tries to find a localized optimum solution, which may eventually lead to globally optimized solutions. Common variations of greedy algorithms are Dijkstra's algorithm, Iterative Dichotomiser 3 (ID3) algorithm, A* algorithm, and Kruskal's algorithm.

COVID-19 is a global pandemic that has claimed many lives around the world from late last 2019 to date. Just like other diseases, it has particular symptoms, but the ultimate confirmation of the sickness is obtained from laboratory test results. Worldwide, conducting tests appears to be a challenging task due to the high population in some countries or limited access to state of the art molecular laboratories. Particularly in the third world and developing countries, the laboratory tests conducted are highly limited due to technological and financial constraints facing these countries [2].

In order to overcome this challenge, researchers have been working round the clock to come up with solutions that would ease testing procedures. It is clear that testing everyone or all the citizens in a country within this short time is not possible; hence who qualifies to get tested adds more to existing challenges. The process of prioritizing test candidates can easily be shadowed with challenges such as human relations and influence, especially in countries with poor governance and a high rate of bribery and corruption.

This study proposes the use of ID3 to classify potentially infected people based on a set of exposed symptoms, after which they can proceed with the laboratory test. The ID3 algorithm adopts a greedy non-backtracking method to construct a decision tree in a top-down recursive approach [3]. The using of ID3 algorithm in this paper was motivated by the previous studies by Güttinger et al. [1] and Ma et al. [4], in which the greedy algorithm approach was used to analyze and classified medical diseases [1, 3-9]. The ID3 algorithm, just like other algorithms used for machine learning classification tasks, can be implemented using the holdout or cross-validation approach. In this study, both approaches are employed, and their performance is compared based on their accuracy and time complexity.

## II. CORONAVIRUS DISEASE 2019 (COVID-19)

Covid-19 is the latest global threat, which was discovered in December 2019 in Wuhan, Hubei province of China, and has spread rapidly throughout the world to a full-blown pandemic. After virus identification and isolation, the pathogen for this new disease was initially referred to as 2019 novel coronavirus (2019-nCoV)2 but it was however renamed officially as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) by the World Health Organization (WHO). Covid-19 is a respiratory disease caused by the virus SARS-CoV-2 which belongs to a group of viruses known as the coronavirus. The coronavirus also belongs to a family of viruses known as the Coronaviridae. The coronavirus was also responsible for the severe acute respiratory syndrome (SARS) outbreak and the Middle East respiratory syndrome (MERS). Compared with the SARS-CoV that caused an outbreak of SARS in 2003, SARS-CoV-2 has a stronger transmission capacity [5]. The Covid-19 outbreak has posed critical challenges for the public health, research, and medical communities as the disease is very communicable through air. The easy and rapid spread

Identify applicable funding agency here. If none, delete this.

of the disease makes prevention and control a high priority issue. Several countries have already instituted a temporary restriction on travel in order to slow down the spread of the disease. The wearing of mask and practicing of social distance of 1meter minimum has also been encouraged and enforced in some countries [6]. Although the clinical symptoms of COVID-19 are predominately respiratory symptoms, some patients have severe cardiovascular damages. Patients with underlying cardiovascular diseases are at higher risk that could lead to death. Therefore, understanding the symptoms of this disease is important for early diagnosis which could give patients a fighting chance and helps control the spread of the diseases. Using untransformed coordinates and Sybyl-style hydrogen naming, the shape of the SARS-CoV-2 main protease receptor is shown shown in Figure 1.

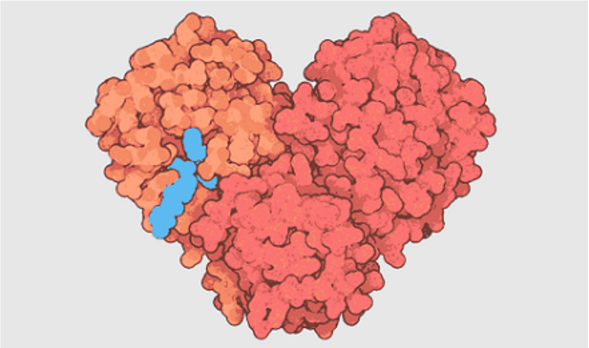


Fig. 1. SARS-CoV-2 (2019-nCoV) coronavirus main protease, with inhibitor in turquoise [7]

## III. USE OF COMPUTED TOMOGRAPHY (CT) FOR SYMPTOMATIC DIAGNOSIS OF COVID-19

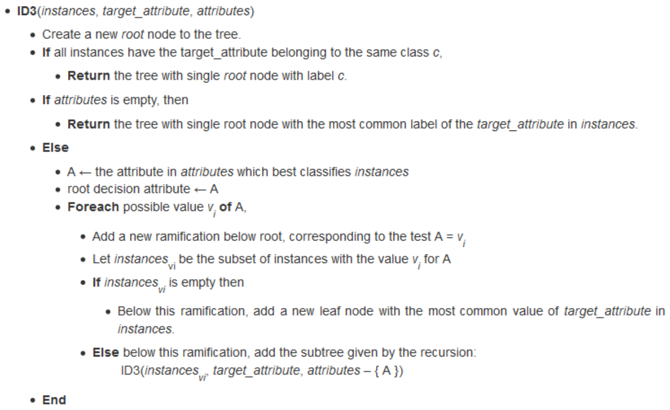
There is great interest in developing symptom-based screen to prioritize who should be tested for Covid-19. However, the reliability of such symptoms has been a subject of debate, but it remains highly important to priorities screening in the face of limited resources. When coronavirus first broke out, details about the disease were sketchy especially among nonhospital patients. To better understand profiling of patients most nations used the option of questionnaires. In the United States for example, the CDC used an optional questionnaire to collect detailed information on confirmed Covid-19 patients [8]. The data was analyzed by age group, sex, hospitalization status, and symptom. It was discovered among 164 confirmed symptomatic patients that 96% reported fever, cough, or shortness of breath. 68% of 57 hospitalized adult patient reported all three symptoms [8]. It has been discovered that Covid-19 can cause a wide range of symptoms. There are over 22 multivariable models to diagnose covid-19 which target suspected cases of covid-19. Most of these models report a C index range of between 0.65 and 0.99. the most frequently diagnostic predictors were flu-like symptoms, age, body temperature, lymphocyte count, and neutrophil count [9]. Most studies used computed tomography images or chest radiographs, Others used spectrograms of cough sounds and lung ultrasound to diagnose patients with symptoms before proper test [10]. Because of the strong infectious rate of covid-19, medical practitioners need a faster way to diagnose patients as proper testing could take days to get results. With CT scan, results could be obtained much faster. In a study of 36 patients with covid-19 who underwent CT scan. The CT scan reported 29/36 covid-19 cases [11].

## IV. ID3 ALGORITHM

### *A. ID3 Algorithm*

The ID3 algorithm is an important research concept for induction research on a data set. It is the most widely used algorithm in decision tree method. The ID3 algorithm was initially introduced by J. Ross Quinlan in 1975 as a Concept Learning System (CLS) algorithm. Over the years the CLS algorithm has developed into an ID3 algorithm that searches through attributes of the training instances and extracts the prefect attribute that best classifies the training set [12]. The ID3 generates decision rules from a set of training examples. ID3 derives it classes from a fixed set of training instances and there is a non-incremental algorithm. Once a class is created, it is used to predict all future instances. By using information theory to choose features, ID3 gives the greatest information gain [12]. To obtain decision rules that best classify the training examples, a test is carried out by selecting the characteristics and then dividing the examples into subgroups using selected characteristics. After grouping, the entropy is calculated to know how importance the feature is [13].

ID3 algorithm are specifically suited for certain data set and expected results. ID3 algorithm is often used when Native Bayes will not satisfy a problem. When a group of features are dependent on each other, the ID3 algorithm is better suited than the Native Bayes. The ID3 algorithm is also useful for categorical data, i.e. data with distinct attributes. For example: hot or cold. Another very useful application of ID3 is when objective values have discrete output values. Example “yes’ or “no”. The Pseudocode of the ID3 algorithm is shown in Figure 2.



```
• ID3(instances, target_attribute, attributes)
    • Create a new root node to the tree.
    • If all instances have the target_attribute belonging to the same class c,
        • Return the tree with single root node with label c.
    • If attributes is empty, then
        • Return the tree with single root node with the most common label of the target_attribute in instances.
    • Else
        • A ← the attribute in attributes which best classifies instances
        • root decision attribute ← A
        • Foreach possible value $v_i$ of A,
            • Add a new ramification below root, corresponding to the test A = $v_i$
            • Let $instances_{vi}$ be the subset of instances with the value $v_i$ for A
            • If $instances_{vi}$ is empty then
                • Below this ramification, add a new leaf node with the most common value of target_attribute in
                  instances.
            • Else below this ramification, add the subtree given by the recursion:
                ID3($instances_{vi}$, target_attribute, attributes – { A })
    • End
```

Fig. 2. Pseudocode of the ID3 Algorithm

### B. Dataset Description

The dataset used in this study, which is a collection of the reported Covid-19 cases and Symptoms put together by different contributors, was obtained from the nCov2019 repository on GitHub [14]. The description of the dataset is shown in TableI, while the dataset attributes are described in Table II.

TABLE I
DATASET DESCRIPTIONS

| Characteristics | Multivariate Attributes |
|---|---|
| **Attributes Type** | Nominal, Continuous |
| **Year** | 2020 |
| **Number of Instances** | 51 |
| **Number of Attributes of Instances** | 7 |
| **Missing Values** | None |

TABLE II
ATTRIBUTES DESCRIPTION

| Attribute | Type |
|---|---|
| Symptoms 1 to 5 | Nominal |
| Contact | Nominal |
| Class Attribute | Continuous |

### C. Experimental Setting

The experiments conducted on the ID3 algorithm, as shown in Figure 3, are two, i.e., the holdout and the cross-validation approach.

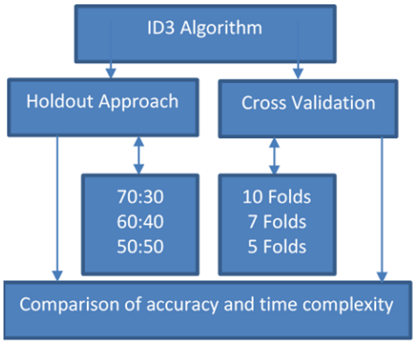


Fig. 3. Methodology

The experiment was conducted in two different approaches; the holdout approach where the dataset is split into training and testing. And the cross-validation approach where resampling procedure is used to evaluate the ID3 generated model on a limited data sample. The results obtained from the experiment of using each approach are presented and discussed in the next section.

## V. RESULTS

The results of the experiments conducted are presented in Table III.

### A. Holdout Approach

Results obtained from the evaluation of the holdout approach on the different splits (70:30, 60:40, and 50:50) are shown in TableIII and Figure 4.

TABLE III
ID3 RESULTS OF HOLDOUT APPROACH

| Holdout(%) | Precision | Recall | F-Measure |
|---|---|---|---|
| 70:30 | 0.628 | 0.444 | 0.488 |
| 60:40 | 0.718 | 0.545 | 0.597 |
| 50:50 | 0.932 | 0.727 | 0.785 |

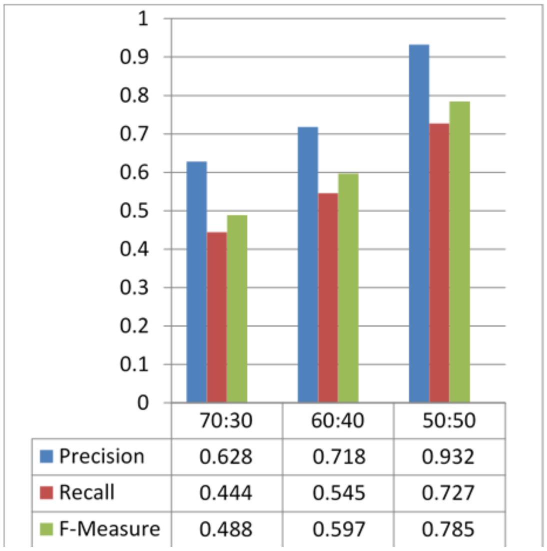


Fig. 4. Graphical Presentation of Obtained Results Based on the Holdout Approach

### B. Cross-Validation

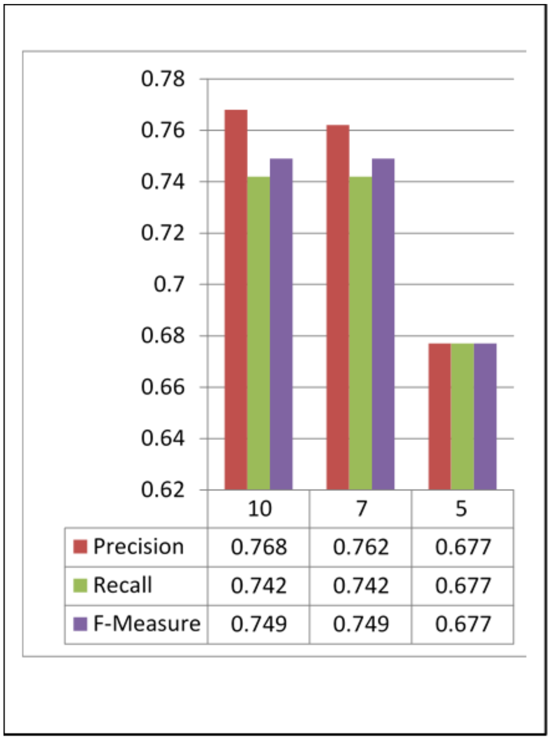
Results obtained from the evaluation of the cross-validation approach on the different folds (10, 7, and 5) are shown in Table IV and Figure 5. A cross-section of the rules generated from the ID3 algorithm on the Covid-19 dataset is shown in Fig 4. The interpretation of the produced rules for the symptoms is based on class values ranging from 10 to 100, where 70 to 100 indicates a high possibility of a Covid-19, and laboratory testing can be conducted on that person. A cross section of the rules generated from the ID3 algorithm on the Covid-19 dataset is shown in Fig 5. The interpretation of the produced rules for the symptoms is based on class values ranging from 10 to 100 where 70 to 100 indicates high possibility of a Covid-19 and laboratory testing can be conducted on that person.

## VI. OBSERVATIONS AND DISCUSSION

Based on combined accuracy values obtained from the holdout and cross-validation experiments, as shown in Figure 5, this section discusses and analyzes the effectiveness of using the ID3 algorithm on Covid-19 symptom's classification.

TABLE IV
ID3 RESULTS OF CROSS-VALIDATION APPROACH

| K-Folds(%) | Precision | Recall | F-Measure |
|---|---|---|---|
| 10 | 0.768 | 0.742 | 0.749 |
| 7 | 0.762 | 0.742 | 0.749 |
| 5 | 0.677 | 0.677 | 0.677 |



| | 10 | 7 | 5 |
|---|---|---|---|
| 🟥 Precision | 0.768 | 0.762 | 0.677 |
| 🟩 Recall | 0.742 | 0.742 | 0.677 |
| 🟪 F-Measure | 0.749 | 0.749 | 0.677 |

Fig. 5. Graphical Presentation of Obtained Results Based on the Cross-Validation Approach

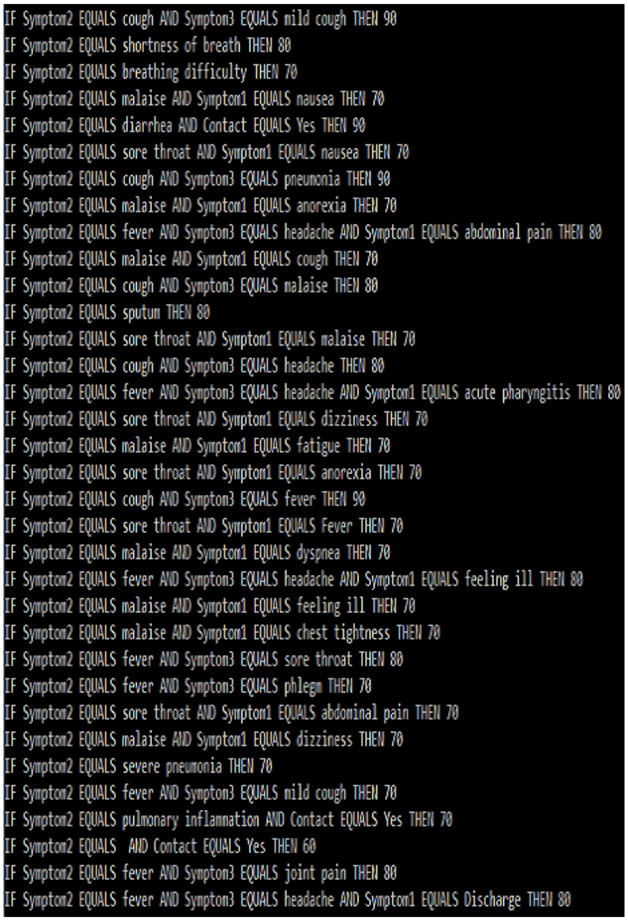
Figure 7 shows the generated rules on each symptom for the ID3 algorithm

## A. Accuracy

Comparing the performance of the ID3 algorithm on the precision and holdout approach in terms of accuracy, as shown in Fig. IV, the holdout approach scored both the highest and lowest on the range of percentage splits experimented. The highest was on the 50:50 split on all metrics, i.e. precision = 0.932, recall = 0.727, and f-measure = 0.785; while the lowest was on the 70:30 split with the precision = 0.628, recall = 0.444, and f-measure = 0.488. However, on all other accuracy values, the cross-validation approach scored higher, except for the precision of the 60:40 split, which was higher than the accuracies obtained by the 5-folds cross-validation.

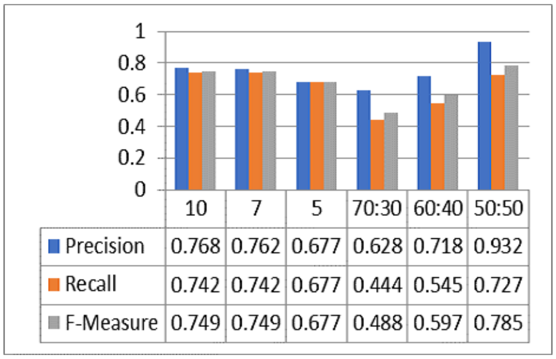
## B. Runtime

In order to have a clearer runtime between the two experimented approaches, a sizeable amount of data was generated with details shown in Table V and result in terms of the ID3 algorithm runtime is shown in Fig 8.



```
IF Symptom2 EQUALS cough AND Symptom3 EQUALS mild cough THEN 90
IF Symptom2 EQUALS shortness of breath THEN 80
IF Symptom2 EQUALS breathing difficulty THEN 70
IF Symptom2 EQUALS malaise AND Symptom1 EQUALS nausea THEN 70
IF Symptom2 EQUALS diarrhea AND Contact EQUALS Yes THEN 90
IF Symptom2 EQUALS sore throat AND Symptom1 EQUALS nausea THEN 70
IF Symptom2 EQUALS cough AND Symptom3 EQUALS pneumonia THEN 90
IF Symptom2 EQUALS malaise AND Symptom1 EQUALS anorexia THEN 70
IF Symptom2 EQUALS fever AND Symptom3 EQUALS headache AND Symptom1 EQUALS abdominal pain THEN 80
IF Symptom2 EQUALS malaise AND Symptom1 EQUALS cough THEN 70
IF Symptom2 EQUALS cough AND Symptom3 EQUALS malaise THEN 80
IF Symptom2 EQUALS sputum THEN 80
IF Symptom2 EQUALS sore throat AND Symptom1 EQUALS malaise THEN 70
IF Symptom2 EQUALS cough AND Symptom3 EQUALS headache THEN 80
IF Symptom2 EQUALS fever AND Symptom3 EQUALS headache AND Symptom1 EQUALS acute pharyngitis THEN 80
IF Symptom2 EQUALS sore throat AND Symptom1 EQUALS dizziness THEN 70
IF Symptom2 EQUALS malaise AND Symptom1 EQUALS fatigue THEN 70
IF Symptom2 EQUALS sore throat AND Symptom1 EQUALS anorexia THEN 70
IF Symptom2 EQUALS cough AND Symptom3 EQUALS fever THEN 90
IF Symptom2 EQUALS sore throat AND Symptom1 EQUALS Fever THEN 70
IF Symptom2 EQUALS malaise AND Symptom1 EQUALS dyspnea THEN 70
IF Symptom2 EQUALS fever AND Symptom3 EQUALS headache AND Symptom1 EQUALS feeling ill THEN 80
IF Symptom2 EQUALS malaise AND Symptom1 EQUALS feeling ill THEN 70
IF Symptom2 EQUALS malaise AND Symptom1 EQUALS chest tightness THEN 70
IF Symptom2 EQUALS fever AND Symptom3 EQUALS sore throat THEN 80
IF Symptom2 EQUALS fever AND Symptom3 EQUALS phlegm THEN 70
IF Symptom2 EQUALS sore throat AND Symptom1 EQUALS abdominal pain THEN 70
IF Symptom2 EQUALS malaise AND Symptom1 EQUALS dizziness THEN 70
IF Symptom2 EQUALS severe pneumonia THEN 70
IF Symptom2 EQUALS fever AND Symptom3 EQUALS mild cough THEN 70
IF Symptom2 EQUALS pulmonary inflammation AND Contact EQUALS Yes THEN 70
IF Symptom2 EQUALS  AND Contact EQUALS Yes THEN 60
IF Symptom2 EQUALS fever AND Symptom3 EQUALS joint pain THEN 80
IF Symptom2 EQUALS fever AND Symptom3 EQUALS headache AND Symptom1 EQUALS Discharge THEN 80
```

Fig. 6. ID3 Generated Rules



| | 10 | 7 | 5 | 70:30 | 60:40 | 50:50 |
|---|---|---|---|---|---|---|
| 🟦 Precision | 0.768 | 0.762 | 0.677 | 0.628 | 0.718 | 0.932 |
| 🟧 Recall | 0.742 | 0.742 | 0.677 | 0.444 | 0.545 | 0.727 |
| ▨ F-Measure | 0.749 | 0.749 | 0.677 | 0.488 | 0.597 | 0.785 |

Fig. 7. Performance of All ID3 Experiments Approach

It is clear from the results presented in Fig 8 that the ID3 algorithm, just like other algorithms, takes more time on the cross-validation approach than on the holdout approach. However, despite that, the ID3 algorithm cross-validation consumed more time than the holdout approach. The holdout 50:50 split achieved higher results than all (10, 7, and 5) folds experimented. This highest score is, however, particular to the 50:50 split, the other holdout splits (60:40, 70:30) performance is below all (10, 7, and 5) folds experimented.

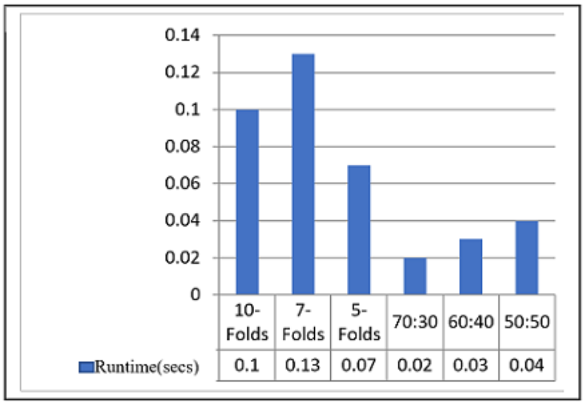


Fig. 8. Runtime of Generated Dataset for All Experiments Approach

TABLE V
DESCRIPTION OF GENERATED DATASET

| **Attributes Type** | Nominal |
|---|---|
| **Size** | 5000 |
| **Number of Attributes** | 10 |
| **Missing Values** | None |

## VII. CONCLUSION

This research reports a comparative study on two approaches of ID3 algorithm classification. The two approaches are holdout and cross-validation which are used to classify Covid-19 symptoms of the pandemic. Findings from this study show that all the cross-validation approaches achieved better accuracy than the 70:30 and 60:40 holdout approach, except the 50:50 split of holdout approach, which scored the overall best accuracy across the holdout and cross-validation approaches. Also, all the cross-validation has higher runtime compared to the holdout approach.

## ACKNOWLEDGMENT

This work was partially supported by Universiti Sains Malaysia (LRGS/MRUN: 203.PKOMP.6777005).

## REFERENCES

[1] D. Güttinger, E. Godehardt, and A. Zinnen, ”Online strategies for optimizing medical supply in disaster scenarios,” in Proceedings of 2011 IEEE International Conference on Service Operations, Logistics and Informatics, 2011, pp. 38-44.

[2] C.-L. Bong, C. Brasher, E. Chikumba, R. McDougall, J. Mellin-Olsen, and A. Enright, ”The COVID-19 Pandemic: Effects on Low-and Middle-Income Countries,” Anesthesia and analgesia, 2020.

[3] S. Kuznetsov, A. Napoli, and S. Rudolph, ”Workshop Notes Seventh International Workshop” What can FCA do for Artificial Intelligence?”,” 2019.

[4] G. Ma, L. Zhang, G. Cui, and Y. Cheng, ”Design of Medical Examination Data Mining System Based on Decision Tree Model,” in Journal of Physics: Conference Series, 2019, p. 022022.

[5] Y.-Y. Zheng, Y.-T. Ma, J.-Y. Zhang, and X. Xie, ”COVID-19 and the cardiovascular system,” Nature Reviews Cardiology, vol. 17, pp. 259-260, 2020.

[6] A. S. Fauci, H. C. Lane, and R. R. Redfield, ”Covid-19—navigating the uncharted,” ed: Mass Medical Soc, 2020.

[7] S. K. Enmozhi, K. Raja, I. Sebastine, and J. Joseph, ”Andrographolide as a potential inhibitor of SARS-CoV-2 main protease: an in silico approach,” Journal of Biomolecular Structure and Dynamics, pp. 1-7, 2020.

[8] R. M. Burke, M. E. Killerby, S. Newton, C. E. Ashworth, A. L. Berns, S. Brennan, et al., ”Symptom Profiles of a Convenience Sample of Patients with COVID-19—United States, January–April 2020,” Morbidity and Mortality Weekly Report, vol. 69, p. 904, 2020.

[9] L. Wynants, B. Van Calster, M. M. Bonten, G. S. Collins, T. P. Debray, M. De Vos, et al., ”Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal,” bmj, vol. 369, 2020.

[10] L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, et al., ”Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT,” Radiology, 2020.

[11] C. Long, H. Xu, Q. Shen, X. Zhang, B. Fan, C. Wang, et al., ”Diagnosis of the Coronavirus disease (COVID-19): rRT-PCR or CT?,” European journal of radiology, p. 108961, 2020.

[12] C. Jin, L. De-Lin, and M. Fen-Xiang, ”An improved ID3 decision tree algorithm,” in 2009 4th International Conference on Computer Science Education, 2009, pp. 127-130.

[13] K. J. Cios and N. Liu, ”A machine learning method for generation of a neural network architecture: A continuous ID3 algorithm,” IEEE Transactions on Neural Networks, vol. 3, pp. 280-291, 1992.

[14] E. Dong, H. Du, and L. Gardner, ”An interactive web-based dashboard to track COVID-19 in real time,” The Lancet Infectious Diseases, vol. 20, pp. 533-534, 2020.