Maximum Value Attribute-based Decision Tree and Random Forest for COVID-19 Prediction

Khurram Gulzar[1], Basharat Ahmad Hassan[2] Muhammad Abbas[3], Zubair Ahmad[3], Muhammad Anis[3] and Hasnat Ullah[3]

[1*]Computer Science, Qurtuba University of Science and Information Technology, Phase 3, Hayatabad, Peshawar, 25000, KPK, Pakistan.

[2]Department of Computer Science, Faculty of Computer Science & Information Technology, The University of Agriculture, Peshawar, 25130, Khyber Pakhtunkhwa, Pakistan.

[3]Department of Computing, Abasyn University, Peshawar, Ring Road, Peshawar, 25000, Khyber Pakhtunkhwa, Pakistan.

*__Correspondence__: khurram57011@gmail.com

Citation | Gulzar. K, Hassan. B. A, Abbas. M, Ahmad. Z, Anis. M, Ullah. H, "Maximum Value Attribute-based Decision Tree and Random Forest for COVID-19 Prediction", IJIST, Vol. 7, Issue. 4 pp 2940-2954, November 2025

Received | October 21, 2025 **Revised |** November 14, 2025 **Accepted |** November 21, 2025 **Published |** November 28, 2025.

The Corona Virus Disease 2019 (COVID-19) is the most threatening disease of the present century that disturbed the whole world from an economic and life perspective. The increased number of positive COVID-19 patients put the health sector under stress to tackle the outbreak of this virus. In the current decade, the usage of Machine Learning (ML) in medical science has increased, particularly in the detection of Heart failure, Pneumonia, Dengue, Breast cancer, and Diabetes. Similarly, the COVID-19 symptoms can be utilized for an early prediction of COVID-19 to reduce the spread rate of infection in society. Several ML techniques detected the COVID-19 disease, and ensemble-based methods like Decision Tree and Random Forest perform well in terms of accuracy as compared to other standard classifiers. However, the execution time and iterations are the major areas of concern for these ensemble-based methods, as early and timely detection of COVID-19 can reduce its infection rate. In this study, the main focus is on the identification of fatal Coronavirus using ML techniques. For that purpose, Rough Set Theory (RST) based Maximum Value Attribute (MVA) is integrated with classical Decision Tree (DT) and Random Forest to efficiently predict COVID- 19 in terms of time and iterations. The proposed model can detect the result of COVID-19 as negative or positive on the eight basic relevant clinical symptoms. Accordingly, the performance of classical DT and RF classifiers is enhanced by integrating MVA. ML models are implemented to evaluate the model performance over clinical symptoms of 136294 COVID-19 patients. The information is extracted from the open-source GitHub website. Based on the symptoms of the COVID-19 data set, four ML models, DT, RF, Maximum Value Attribute-based Decision Tree (MVA-DT), and Maximum Value Attribute-based Random Forest (MVA-RF) were implemented in Jupyter Notebook via Python repository to forecast the result of COVID-19. Standard performance parameters of the classification process are considered to test model reliability against the prediction of COVID-19. From the time and iteration perspective, the proposed MVA-DT out- performed the other three models, and the MVA-RF technique predicted COVID-19 disease comparatively better, having 95.82% accuracy, 81.90% precision, 59.28% recall, and 68.77% F1 score.

Keywords: COVID-19 Prediction, MVA, Symptoms, Rough Set Theory

Introduction:

COVID-19 carries a serious life threat to worldwide health. The outbreak of this unique virus has spread all over the world rapidly from Wuhan, a city in China, in December 2019. COVID-19 spreads from person to person and disturbs the respiratory system of the human body, which can also lead to death. The virus was declared a pandemic by the World Health Organization on 11 March 2020. The disease reached the rest of the countries outside of China rapidly through tourists, putting the world population in danger, particularly old people who have chronic diseases. COVID-19 infection challenges medical setups globally in many circumstances, including demands for beds in hospitals and creating an acute shortage in terms of electromedical devices. The paramedical units also suffered a lot from this microbe. The increased demand for ventilators for COVID-19 patients, capacity for timely medical decisions, and efficient use of health- care resources are major concerns of developing countries where the health system is not up to the mark. Therefore, timely detection of this malady is necessary to overcome issues related to the scarcity of equipment and human resources [1].

In order to reduce the spread of COVID-19 in public, Machine Learning (ML) models are utilized. To cope with COVID-19 disease, it is foremost necessary to predict it promptly, as there is the possibility of an increase in the spread rate of the virus with the passage of time. Machine Learning field is practiced in discovering COVID-19 outcomes by using techniques like Auto Regressive Moving Average (ARMA), Support Vector Regressor (SVR), Linear Regression (LR), Auto Regressive Integrated Moving Average (ARIMA), Lin- ear Regressor polynomial (LRP), Bayesian Ridge (BR), Extreme Gradient Boosting (XGB), Random Forest Regressor (RFR) and Holt Winter exponential smoothing (HW). Despite implementing these ML techniques on COVID-19 datasets, there are accuracy and time-consumption issues in diagnosing infection. Moreover, the training of algorithms takes more iterations due to the complex procedures of classifiers to predict the disease [2].

Comparative analysis is established between different Machine Learning techniques that detected the mortality rate of COVID-19 and patients admitted to the Intensive Care Unit (ICU) [3]. By performing implementation, it is proven that ensemble-based methods like Decision Tree and Random Forest delivered better accuracy by comparing with other ML methods like LR, SVR, Naive Bayes, K-Nearest Neighbor, Discriminant Analysis, Gaussian process, and neural model. Similarly, Gradient Boosting, SVR, LR, LRP, BR, XGB, RFR, Naive Bayes, K-Nearest Neighbor, and Neural Network are Machine Learning techniques deployed for detecting COVID-19 disease. Apart from accuracy, there are other issues of execution time and iteration steps that need to be improved because detecting COVID-19 promptly can easily reduce the infection rate in different areas [4]. In this study, the major issue of excessive iterations and time is handled by integrating a rough set-based model that chooses only the most significant attributes while ignoring the irrelevant ones during data analysis. Hence, the time and iterations are surely reduced.

A rough set is a mathematical approach for intelligent data analysis and data mining. A rough set approach is used to discover structural relationships within noisy and imprecise data. The rough set can also handle vague, inconsistent, and uncertain information. The Maximum Value Attribute (MVA) technique was introduced, utilizing the rough value set for the selection of the most optimal attribute(s). The MVA technique is one of those clustering techniques where the domain knowledge is enough for the decision-making. The MVA technique analyzes the data set and ranks the attributes by the maximum cardinality principle [5]. In this study, the MVA technique is now practiced for the selection of the root node. Considering the classical root node selection by Decision Tree and Random Forest,

the MVA is expected to reduce the computational steps significantly. The MVA technique allows us to get rid of additional requirements like searching for dependence and the weight of data.

Finally, Maximum Value Attribute-based Decision Tree (MVA-DT) and Maximum Value Attribute-based Random Forest (MVA-RF) are two integrated techniques that are introduced in this study to reduce execution time and iteration steps for identifying COVID-19 in large freely accessible open-source information available on the GitHub website. Integration of MVA with DT and RF makes virus prediction in a simplified mathematical approach by replacing the selection procedure of the root node of the tree. We compare the performance of classical DT and RF with the proposed MVA-DT and MVA-RF for COVID-19 prediction.

Fig 1 shows that the existing Machine Learning algorithms for predicting the disease of COVID-19 require more time to generate better results. Accuracy is also compromised for complex datasets. Decision Tree and Random Forest overcome accuracy issues; however, the time and iterations to predict COVID-19 need to be reduced. These issues are overcome by the Maximum Value Attribute (MVA) based Decision Tree and Random Forest. The predictive integrated model not only reduces the number of computational steps but also produces better results in health care setup in less amount of time.

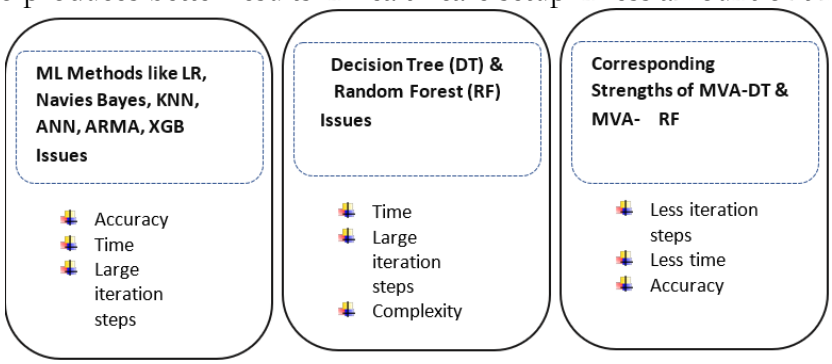

Figure 1. Issues of COVID-19 Predictors

The rest of the article is organized as follows: the overview of the research background and related literature is discussed in Section 2. The working methodology of the Proposed Integrated Predictive MVA-DT and MVA-RF is presented in Section 3. Experimental results and discussion on the COVID-19 data set are presented, analyzed, and summarized in Section 4. In the end, the summary of the conclusion is mentioned in section 5, and the future direction of this research work is illustrated in section 6.

Related Work:

The existing studies are summarized in Table 2 and ?? in terms of some relevant literature review parameters. Several related proposed and compared techniques are mentioned with their pros and cons. The table highlights that several ML techniques like Gradient Boosting, SVR, LR, DT, RF, Naive Bayes, K-Nearest Neighbor, ANN, and XGB have been experimented on different COVID-19 data sets for prediction. The DT, RF, and LR outperformed other techniques by considering performance parameters of accuracy, error rate, recall, F1 score, and precision. Mostly, the clinical symptoms, audio, images, and laboratory COVID-19 datasets are utilized to test the performance of ML techniques to reduce the spread rate of infection in public. Thousands of COVID- 19 datasets are available on GitHub, Kaggle, and the Johns Hopkins University website. Researchers gathered COVID-19 information from these websites because these open- source websites have verified, relevant, and precise datasets.

It is also evident from the table that DT and RF classifiers can efficiently make clear rules for the detection of COVID-19 by analyzing datasets. These supervised ML techniques

can make the relationship between different attributes of the data set that is helpful for the prediction of COVID-19 disease. Moreover, they perform well on small categorical datasets of a particular country or city. However, they also carry some limitations. Firstly, these techniques have not been experimented on a large volume of COVID-19 datasets, which may affect accuracy. Secondly, the ML techniques need to be tested on the COVID-19 data set of different countries because they have different dynamics and factors of the COVID-19 disease. Lastly, the iterations and time consumption can be reduced by suggesting alternative strategies for tasks like root node selection, or integrating simple statistical and mathematical concepts, like rough set theory, can significantly reduce the additional iterations and excessive time taken.

Table 1. Summary of Related Work

Citation	Proposed Technique	Compare Technique	Pros	Cons
[1]	MLP-ICA	ANFIS	Perform well on a small dataset	Performance changes if dynamic changes
[2]	ARMA	ARMA, SVR, LRP, BR, LR, HW, and XGB	ARMA performs well in a small dataset.	ARMA takes more time for training
[3]	RF	XG-Boost, RF, and Multinomial LR	RF is helpful to predict the severity of COVID-19.	Performance should be tested on a large-scale dataset.
[4]	DT	LR, Naïve Bayes, ANN, SVM	DT detects relevant features in predicting COVID-19	Performance should be tested on large scale dataset
[5]	MVA	MDA, MSA, ITDR	MVA performed well on a small categorical dataset.	MVA needs to be extended to large datasets
[6]	Gradient Boosting	SHAP	Perform well in predicting a symptom-based dataset	Self-reported symptoms can be misleading
[7]	Grad-Boost technique	Shapley	Deliver better in categorical data	Medical symptoms can be misleading
[8]	DT and RF	LR, SVR, Naïve Bayes, KNN	DT and RF can easily make relevance among attributes.	Implementation on a larger dataset is challenging.
[9]	ANN	ANN, DT, PLS-DA and KNN	A healthcare professional can utilize an ANN in a public organization	Medical biomarker levels can mislead
[10]	MRMR	SVM	MRMR efficiently reduce redundancy of features	The ML Model needs to be tested on a symptoms-based dataset
[11]	DNN	CNN	CNN performs well in a small dataset	CNN is not suitable for a larger dataset
[12]	LR	DT, MLP, XGB CNN	LR performs well in a small dataset	Public sources of data can be misleading

[13]	DT	Kaplan Meier survival Curve Analysis, K-Means cluster	DT has a good hold on a small dataset from an accuracy perspective	DT needs to be practiced on a large dataset
[14]	LR	Lasso, ES, vector assistance	LR efficiently forecast future scenario of COVID-19	LR needs to be practised on a large dataset
[15]	MLP	LR, SVM, and MLP	Perform well on a small dataset	Performance should be tested on a large-scale dataset

Table 2. Summary of Related Work

Citation	Proposed Technique	Compare Technique	Pros	Cons
[16]	XG-Boost	LASSO, XG-Boost	XGB tree easily predicts COVID-19 by analysis of clinical features	Performance should be tested on a large-scale dataset
[9]	LR, RF	Naive Bayes, SVM, RF, and LR	Perform well on a small dataset.	Performance should be tested on a large-scale dataset.
[17]	Neural Network	SVM Bayesian, Network, and NN	Perform well on a small dataset	Performance should be tested on large scale dataset
[18]	RF	Random Forest, SVM, and ANN	RF predicts COVID-19 efficiently as it decides on the majority	More features of COVID-19 should be tested
[19]	SVM	K-NN, RF, SVM, DT, and K-NN	Perform well on a small dataset	Performance should be tested on large scale dataset
[20]	RF	RF, RF, SVM, Neural Network, and DTr	K-Fold validation is proven to be reliable for measuring the consistency of RF.	Performance needs improvement
[21]	J48 DT	Hoeffding DT	DT easily generates rules on the symptoms dataset	Performance comparisons are not very significant
[22]	Deep learning	ANN	DNN can timely predict COVID-19 to control the spread rate	Medical researchers need experimentation
[23]	DT	SVM, Naive Bayes	DT make relationship between factors precisely	Need implementation in other counties
[24]	RF	DT, Naive Bayes, LR, KNN, SVM	RF can analyze the relationship of COVID-19 features	More symptoms need to be tested
[25]	ANN	KNN	ANN is suitable for short-term prediction	Performance should be tested on large scale dataset

[26]	DT	LR, SVM, GB, Neural network	DT predict mortality rate of COVID-19 efficiently	Study to be extended to other regions
[27]	DT	DT, LR	DT helps make connections among different attributes	Cross validation be utilized to remove outliers
[28]	Prophet Algorithm	Prophet algorithm, Polynomial Multi-Layer Perceptron	It can estimate missing information by analyzing other datasets	Contact tracing among person-to-person is lacking
[29]	Gradient Boosting	RF, Gradient Boosting, and K-NN	GB performs well with a dataset of different dynamics	Performance needs improvement

Proposed Maximum Value Attribute-based Decision Tree (MVA-DT) and Random Forest (MVA-RF):

The proposed methodology of rough set integration with classification techniques is shown in a flowchart as represented in Figure 2. Data set collection, pre-processing of data, testing, and training are the different phases of the strategy. Secondly, the implementation of MVA-based classical DT and RF for predicting COVID-19 is the most important phase of the proposed methodology. Moreover, the comparative analysis of classical DT, MVA-DT, RF, and MVA-RF decides the optimal predictive technique based on performance classification. The proposed integrated predictive models are tested on COVID-19 symptoms, and the models will be ranked by analyzing several relevant performance parameters.

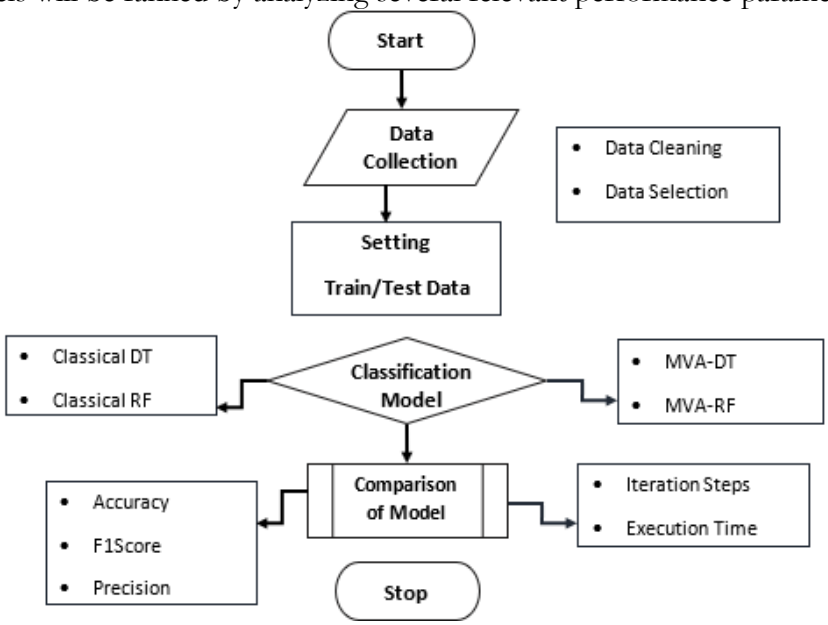

Figure 2. Research Process Flow Chart

Data Set Description and Experimental Setup:

The data set contained clinical features of COVID-19 patients collected from an open-source GitHub [6] with nine important clinical attributes like cough, fever, headache, shortness of breath, sore throat, gender, age above 60 or not, test indication, and COVID-19 result as the target label. Jupyter Notebook (6.4.8 version) is utilized as an implementation tool kit. The experiments are conducted on a Windows 8.1 Pro Lenovo (64-bit operating system) with an Intel(R) Celeron(R) CPU N2840 @2.16 GHz. The COVID-19 symptoms-based data set is distributed with 90-10 %, 80-20 %, and 70-30 % ratio. After data set distribution, performance factors of accuracy, precision, recall, F1 score, time, and iteration are extracted for the

prediction of the COVID-19 pandemic disease. Built-in Graphical features are called for a better understanding of data sets and results. The working procedure of the Classical Decision Tree and the Maximum Value Attribute-based Decision Tree is not similar. The only difference lies in the selection of the root node. Mathematical procedures to acquire the root node of DT and MVA- DT are different. Classical Decision Tree, Random Forest, MVA-DT, and MVA-RF are experimented on a data set of a total of 136294 patients with clinical symptoms as shown in Table 3. The attributes with their values are: Cough (Yes / No), Fever (Yes / No), Sore Throat (Yes / No), Shortness of Breath (Yes / No), Headache (Yes / No), Age 60 and Above (Yes / No), Gender (Male / Female), Test Indication (Other, Abroad, Contact with confirmed), Corona Result (Positive / Negative) - Target Label.

After gathering the data set, analysis was performed, and concluded that there exist several null values that were initially removed in Excel before uploading to Jupyter Notebook. The data set description is overviewed in Anaconda software via the Jupyter Notebook tool. After uploading the COVID-19 information, it was split into testing and training for model testing. The target label is selected with the attribute name corona result, and the remaining attributes are utilized for model training and testing to examine the performance.

Machine Learning Models Deployment:

The dataset of 278648 patients with COVID-19 symptoms is acquired from the GitHub website [7]. Data cleaning was mandatory as it contained many missing and null values. Excel was utilized to remove such values by applying filters on each attribute. After data cleaning, the clinical symptoms-based COVID-19 dataset was reduced to 136294 patients. After this data preprocessing, the testing and training data are separated, and DT, RF, MVA-DT, and MVA-RF are implemented in the Jupiter Notebook. Different functions for dataset visualization are imported by calling Pandas and scikit- learn libraries for the prediction of COVID-19. In the proposed model, the root node selection strategy of classical DT and RF is replaced with the rough set-based MVA technique. Fig 3 shows the MVA role in deciding the root node by working on the Maximum cardinality principle to rank attributes. Moreover, it also highlights the significant reduction of computational steps to find the root node.

Table 3. Features of the COVID-19 dataset and attributes utilized by the ML predictive models in this study

Attribute	Total n = 136294 n	Total (%) %	COVID-19 n =125668 n %	Negative (%)	COVID-19 Positive n=10626 n	Positive (%)
Gender: Male	69153	50.74	63140	50.24	6013	56.59
Gender: Female	67141	49.26	62528	49.76	4613	43.41
Age 60 and Above	23701	17.39	21648	17.23	2053	19.32
	112593	82.61	104020	82.77	8573	80.68
Cough Yes	24851	18.23	19774	15.74	5077	47.78
No	111443	81.77	105894	84.26	5549	52.22
Fever Yes	12661	9.29	8212	6.53	4449	41.87
No	123633	90.71	117456	93.47	6177	58.13
Sore Throat Yes	1473	1.08	117	0.09	1356	12.76

No	134821	98.92	125551	99.91	9270	87.24
Shortness of Breath Yes	1061	0.78	86	0.07	975	9.18
No	135233	99.22	125582	99.93	9651	90.82
Headache Yes	2075	1.52	81	0.06	1994	18.77
No	134219	98.48	125587	99.94	8632	81.23
Test Indication Other	114358	83.91	110248	87.73	4110	38.68
Abroad	14534	10.66	13185	10.49	1349	12.70
Contact with Result	7402	5.43	2235	1.78	5167	48.62
confirmed Corona						
Positive Negative	10626125668	7.8092.20	Target Labels			

Root node by Decision Tree:

The DT algorithm named ID3 is traditionally utilized where the root (parent) node is selected by Information Gain (IG) and Entropy criteria.

Information Gain (IG):

Information Gain is a calculation of uniqueness (difference) in entropy from the beginning to after the target label is divided by its outcome, and it checks the uncertainty of its sub-values. IG plays a pivotal role in the selection of the root node, as its value is subtracted from the entropies of the rest of the attributes.

$$I.G = \frac{P}{-P+N} log2\ \frac{P}{P+N} - log2 \frac{N}{P+N} \quad (1)$$

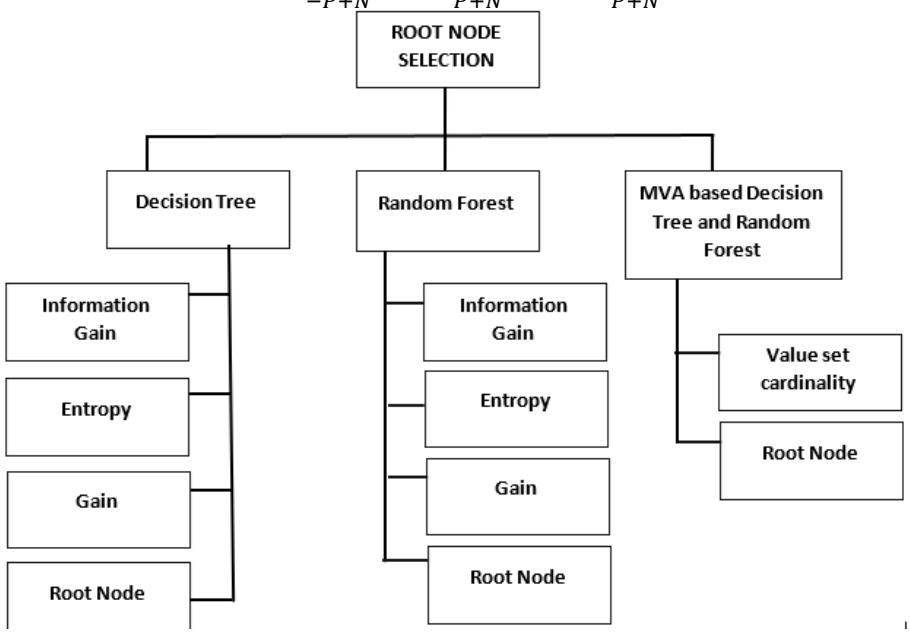

Figure 3. Root Note Selection

Entropy (E):

Entropy refers to the computation of uncertainty of attributes (A) except the target attribute, which is covid result, in the data set used for COVID-19 prediction. This entropy of

all attributes is subtracted from the IG of the target label, and we achieve the gain on which our root node is selected.

$$E(A) = \sum_{i=1}^{v} \frac{Pi+Ni}{P+N} I(Pi + Ni) \quad (2)$$

Where, A = current attribute, P = Probability of Yes, and N = Probability of No

Gain:

The difference between IG and entropy is referred to as Gain. The attribute that has the highest gain is selected as the root node of a tree.

$$Gain = IG - E(A) \quad (3)$$

Root node by Maximum Value Attribute (MVA):

The MVA technique is used for the selection of root nodes in predicting COVID-19. The working steps of the MVA technique are simple, with less execution time and fewer iteration steps. MVA is a rough set-based recent technique where domain knowledge is enough for decision-making. It can select a better-suited root node with fewer computational steps. The MVA technique works on the concept of a rough value set. The MVA technique allows us to get rid of the additional computational steps of searching for dependence and the weight of the data.

The working procedure of the MVA technique in finding the root node of a tree is less complicated as compared to classical DT. In MVA, we analyze the data and calculate the total outcome (cardinality) of each attribute except the target attribute. The attribute that contains the highest outcome is selected as our root node. The cardinality of an attribute will be determined by the Cardinality principle using the equation.

Here, "V" is the value set and "a" is the value of that attribute. The MVA technique is integrated with classical DT and RF. The pseudo-code of MVA-DT and MVA-RF is also presented subsequently.

$$Cardinality\ V_a = |V_a| \quad (4)$$

Algorithm 1: Calculate the cardinality of all attributes except the target label

Require: [Input:] COVID-19 Data set

Ensure: [Output:] Prediction of Corona Result (Positive/Negative)

Select the best attribute as the root node from several available attributes by determining the cardinality of all attributes using MVA. The cardinality of the attribute will be found using equation 4.

if *Cardinality* of *one* attribute > *the other* attributes *except the* target *label,* **then**

The attribute with the highest cardinality is selected as the final root node of a Decision Tree

else

It is selected as the left or Right Child Node of a Decision Tree

After achieving the Root Node, MVA-DT and MVA-RF repeat steps 1 and 2 on each subset until you find other child leaf nodes in all the branches of the tree.

end if

```
for i:= p o s s i b l e  val ue s
begin
{ Cardinality Check }
end ;
Write ( ' Outcome - o f - a l l - a t t r i b u t e s ' ) ;
Write ( ' Root - Node ' ) ;
```

Experimental Results and Discussions:

Comparative analysis of classical DT, RF, proposed integrated MVA-DT, and MVA-RF is established by implementing these predictive models. These models are tested on a COVID-19 clinical symptoms-based data set to validate the performance of models with

different data distribution phases. Testing and training of the COVID-19 data set are distributed with 90/10, 80/20, and 70/30 ratios to analyze the efficiency of the applied models used in the study. These proposed integrated predictive models are critically viewed under some performance parameters, especially on accuracy, precision, recall, and F1 score. Moreover, the comparison is also extended to execution time and iteration steps. The experiment procedure is represented in Figure 4, which consists of applied models, data set specification, distribution of data, and classification performance criteria.

At the start of implementation, the root node of the classical Decision Tree is obtained. For that reason, the information gain of the target label, i.e. corona result, is calculated.

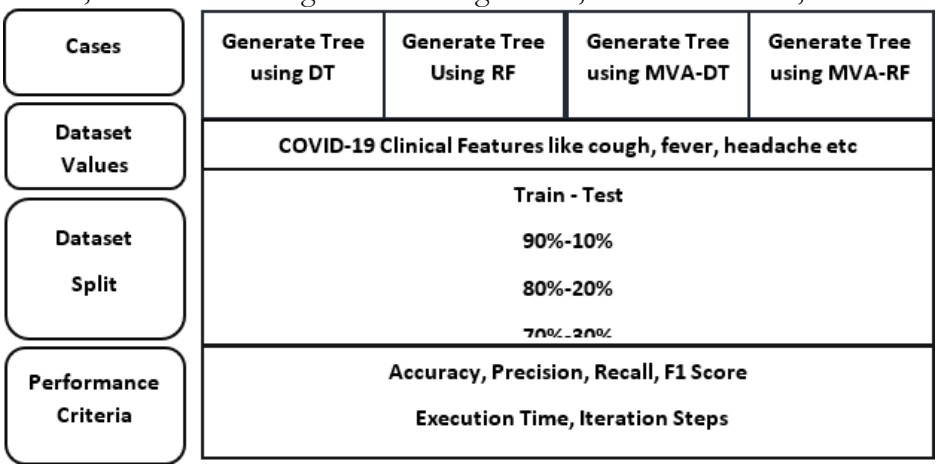

Figure 4. Experimental Procedures of all implemented techniques

At the start of implementation, the root node of the classical Decision Tree is obtained. For that reason, the information gain of the target label, i.e. corona result, is calculated.

The next step is to find the entropy of the rest of the attributes that are eight in our data set. After getting the entropy of all attributes, each attribute's entropy is subtracted from the information gain of the target attribute. By performing this, a gain of all attributes is gathered. We select that attribute as our final root node that contains a maximum value, which is known as "Gain. The mathematical values in the model implementation are mentioned below in Table 4.

Table 4. Root Node measurements by DT

Attributes	IG of Target	Entropy of Attributes except Target	Gain (IG − Entropy)
Attribute (e.g., corona result)	IG value	Entropy value	IG − Entropy
Cough	-	0.028	0.3666
Fever	-	0.048	0.3465
Sore Throat	-	0.033	0.3616
Shortness of Breath	-	0.023	0.3712
Headache	-	0.522	0.3427
Corona result (Target)	0.3949	-	-
Age 60 and Above	-	0.000	0.3949
Gender	-	0.000	0.3949
Test Indication	-	0.112	0.2829 (Max Gain)

Now, to compute the root node using the Maximum Value Attribute-based Decision Tree, we analyze the data and calculate the total outcome (cardinality) of each attribute except the target attribute. The attribute that contains the highest outcome is selected as our root node. After model implementation, the mathematical values are mentioned below in Table 5. In subsequent sections, the experimental results on classical Decision Tree (DT), Random Forest (RF), MVA-DT, and MVA-RF are presented. The summary of results with the average

output is also illustrated accordingly. All techniques are deployed on freely accessible open-source data of 136294 patients having COVID-19 symptoms. The data set comprises nine important clinical attributes like cough, fever, headache, shortness of breath, sore throat, gender, age above 60 or not, test indication, and covid result, etc. Out of these nine attributes, covid result is selected as the target label. These features contained values as Yes and No, on which classical DT predicted the outcome of COVID-19, whether someone would encounter the COVID-19 virus or not. Eight attributes are used for testing and training. Initially, the classical DT is implemented, and the obtained average results are represented in Table 6. Similarly, the classical RF is implemented to test the models by considering the performance parameters. The obtained results, along with the average values of RF, are represented in Table 7. Now the integration of MVA with classical DT is implemented, and the obtained average results of the proposed MVA-DT are represented in Table 8. Finally, the integration of MVA with classical RF is also implemented, and the obtained average results of the proposed MVA-RF are represented in Table 9.

Table 5. Root Node measurements by MVA-DT

Attribute	Value set cardinality	MVA
Cough Fever	2	
Sore Throat Shortness of Breath Headache	2	
Corona result	2	
Age above 60 and Above Gender	2	
Test Indication	2	Max info outcome

Table 6. Results achieved by Classical Decision Tree

Performance Factors	70-30 % Train - Test	80-20 % Train - Test	90-10 % Train - Test	Average
Accuracy (%)	95.85%	**95.85 %**	95.71 %	95.80 %
Precision (%)	81.77 %	**81.54 %**	82.33 %	81.88 %
Recall (%)	57.88 %	**59.93 %**	57.88 %	59.01 %
F1 Score (%)	68.70 %	**69.08 %**	67.98 %	68.58 %
Time Execution	1 m 16.28 s	**1 m 16.28 s**	1 m 16.28 s	1 m 16.28 s
Iteration Steps	670	**670**	670	670

Table 7. Results achieved by Classical Random Forest

Performance Factors	70-30 % Train - Test	80-20 % Train - Test	90-10 % Train - Test	Average
Accuracy (%)	95.85 %	**95.88 %**	95.73 %	95.82 %
Precision (%)	81.77 %	**81.56 %**	82.38 %	81.90 %
Recall (%)	59.37 %	**60.40 %**	58.07 %	59.28 %
F1 Score (%)	68.79 %	**69.40 %**	68.12 %	68.77 %
Time Execution	1 m 16.99 s	**1 m 16.99 s**	1 m 16.99 s	1 m 16.99 s
Iteration Steps	67	67	67	67

Table 8. Results achieved by the Maximum Value Attribute-based Decision Tree

Performance Factors	70-30 % Train - Test	80-20 % Train - Test	90-10 % Train - Test	Average
Accuracy (%)	95.85%	**95.85 %**	95.71 %	95.80 %
Precision (%)	81.77 %	**81.54 %**	82.33 %	81.88 %
Recall (%)	57.88 %	**59.93 %**	57.88 %	59.01 %
F1 Score (%)	68.70 %	**69.08 %**	67.98 %	68.58 %
Time Execution	1 m 12.37 s	**1 m 12.37 s**	1 m 12.37 s	1 m 12.37 s
Iteration Steps	37	**37**	37	37

Table 9. Results achieved by Maximum Value Attribute-based Random Forest

Performance Factors	70-30 % Train - Test	80-20 % Train - Test	90-10 % Train - Test	Average
Accuracy (%)	95.85%	**95.88 %**	95.73 %	95.82 %
Precision (%)	81.77 %	**81.56 %**	82.38 %	81.90 %
Recall (%)	59.37 %	**60.40 %**	58.07 %	59.28 %
F1 Score (%)	68.79 %	**69.40 %**	68.12 %	68.77 %
Time Execution	1 m 13.65 s	**1 m 13.65 s**	1 m 13.65 s	1 m 13.65 s
Iteration Steps	370	**370**	370	370

The findings of implemented classical DT, classical RF, proposed MVA-DT, and MVA-RF on COVID-19 information are represented in tabular form with average performance factors of accuracy, precision, recall, F1 Score, time execution, and mathematical computational steps (iteration steps). The results of the proposed integrated model for COVID-19 prediction are summarized in Table 10. The MVA-RF technique predicted COVID-19 disease comparatively better in terms of 95.82% accuracy, 81.90% precision, 59.28% recall, and 68.77% F1 score. However, the comparison in terms of other factors, including execution time and iteration steps, also illustrates that the MVA-DT predicted the result of COVID-19 disease in less execution time of 1 m 12.37 s with 37 iterations. These results show that utilizing the integrated MVA-DT and MVA-RF to detect COVID-19 outcomes in less time with limited iteration steps without impacting accuracy. In terms of time and iteration steps, the proposed MVA-DT is efficient as it outperformed the remaining three techniques that predicted COVID-19. However, from an accuracy perspective, the proposed MVA-RF is comparatively better and detects COVID-19 with more confidence and reliability.

Table 10. Result summary of Proposed Integrated Techniques

Applied Models	Avg Accuracy	Avg Precision	Avg Recall	Avg F1 Score	Avg Execution Time	Avg Iteration
DT	95.80 %	81.88 %	59.01 %	68.58 %	1 m 16.28 s	67
RF	95.82 %	81.90 %	59.28 %	68.77 %	1 m 16.99 s	670
MVA-DT	**95.80 %**	**81.88**	**59.01 %**	**68.58 %**	**1 m 12.37 s**	**37**
MVA-RF	**95.82 %**	**81.90**	**59.28 %**	**68.77 %**	**1 m 13.65 s**	**370**

Conclusion:

A Rough set-based unsupervised Maximum Value Attribute (MVA) technique is integrated with classical classifiers like Decision Tree (DT) and Random Forest (RF) to efficiently predict the COVID-19 disease. The proposed integrated approach is comparatively analyzed with classical DT and RF over a publicly available COVID-19 symptoms-based data set extracted from GitHub. From the experiments based on the simulation results, it is concluded that MVA-DT detected better results of COVID- 19 utilizing less time and iterations, and MVA-RF outperformed in terms of accuracy and other factors. In this study, the investigations and research conducted lead toward following major contributions:

The performance of classical DT and RT classifiers is enhanced by integrating MVA.

We integrated the Maximum Value Attribute technique with the classical Decision Tree and Random Forest to predict the result of COVID-19 more precisely.

Integrated predictive models MVA-DT and MVA-RF efficiently select the root node of the decision tree in a more simplified manner. As a result, complexity, time, and iterations are reduced, which is the main contribution.

The integrated MVA-DT and MVA-RF predicted COVID-19 efficiently by considering other performance parameters like accuracy, F1 score, Precision, and recall.

Discussion:

The experimental results of this study demonstrate that the integration of Rough Set

Theory-based Maximum Value Attribute (MVA) significantly enhances the efficiency of classical ensemble models. While traditional Decision Tree (DT) and Random Forest (RF) models rely on computationally intensive metrics like Information Gain and Entropy for root node selection, the proposed MVA-DT and MVA-RF models simplify this process by using the maximum cardinality principle. This modification led to a notable reduction in iteration steps from 67 iterations in classical DT to just 37 in MVA-DT and a decrease in execution time without compromising the high accuracy of 95.82% achieved by the MVA-RF model.

When comparing these findings with existing literature, the MVA-based approach addresses several limitations noted in previous COVID-19 prediction studies. For instance, research by Khakharia et al. (2021) utilizing ARMA models reported that while effective for small datasets, the models were time-consuming during the training phase. Similarly, standard Decision Tree implementations in studies by Muhammad et al. (2021) showed strong feature detection but faced performance challenges when scaled to larger datasets. In contrast, our proposed MVA models were tested on a substantial dataset of 136,294 patients, maintaining an average accuracy of over 95.8% while specifically targeting the "time-consumption" and "complex iteration" issues highlighted in those earlier works.

Furthermore, the predictive reliability of our MVA-RF model, which yielded a precision of 81.90% and an F1 score of 68.77%, compares favorably with other ensemble methods documented in the literature. Studies using Gradient Boosting and XG-Boost have been praised for their performance on clinical features, but often require extensive hyperparameter tuning and iterations. Our study suggests that by replacing the root node selection strategy with MVA, we can achieve comparable accuracy to these complex models but with a "simplified mathematical approach. Although the recall remains a point for future improvement due to class imbalances in symptom-based data, the current MVA integration provides a more viable solution for real-time medical diagnostic tools where early and timely detection is critical.

Future Work:

The proposed Integrated Predictive Model can be helpful for health management authorities as precise COVID-19 prediction will enable them to be focused and alert for upcoming new viruses. Accordingly, the destruction can be reduced by adopting necessary precautionary measures. Moreover, the proposed integrated predictive model, MVA-DT and MVA-RF, has the potential to improve performance as more relevant data is available and fewer resources can be utilized by practicing in the public sector by medical practitioners. This model can efficiently tackle Congo and other new variants of COVID-19, like Omicron and BF-7.

References:

[1] D. P. Kavadi, R. Patan, M. Ramachandran, and A. H. Gandomi, "Partial derivative Nonlinear Global Pandemic Machine Learning prediction of COVID 19," *Chaos, Solitons & Fractals*, vol. 139, p. 110056, Oct. 2020, doi: 10.1016/J.CHAOS.2020.110056.

[2] A. Khakharia *et al.*, "Outbreak Prediction of COVID-19 for Dense and Populated Countries Using Machine Learning," *Ann. Data Sci. 2020 81*, vol. 8, no. 1, pp. 1–19, Oct. 2020, doi: 10.1007/S40745-020-00314-9.

[3] A. A. Alrajhi *et al.*, "Data-Driven Prediction for COVID-19 Severity in Hospitalized Patients," *Int. J. Environ. Res. Public Heal. 2022, Vol. 19, Page 2958*, vol. 19, no. 5, p. 2958, Mar. 2022, doi: 10.3390/IJERPH19052958.

[4] L. J. Muhammad, E. A. Algehyne, S. S. Usman, A. Ahmad, C. Chakraborty, and I. A. Mohammed, "Supervised Machine Learning Models for Prediction of COVID-19 Infection using Epidemiology Dataset," *SN Comput. Sci. 2020 21*, vol. 2, no. 1, pp. 11-, Nov. 2020, doi: 10.1007/S42979-020-00394-7.

[5] J. Uddin, R. Ghazali, M. M. Deris, U. Iqbal, and I. A. Shoukat, "A novel rough value set categorical clustering technique for supplier base management," *Comput. 2021 1039*, vol. 103, no. 9, pp. 2061–2091, Apr. 2021, doi: 10.1007/S00607-021-00950-W.

[6] Y. Zoabi, S. Deri-Rozov, and N. Shomron, "Machine learning-based prediction of COVID-19 diagnosis based on symptoms," *npj Digit. Med. 2021 41*, vol. 4, no. 1, pp. 3-, Jan. 2021, doi: 10.1038/s41746-020-00372-6.

[7] Y. Zoabi and N. Shomron, "COVID-19 diagnosis prediction by symptoms of tested individuals: a machine learning approach," *medRxiv*, p. 2020.05.07.20093948, May 2020, doi: 10.1101/2020.05.07.20093948.

[8] S. Subudhi *et al.*, "Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19," *npj Digit. Med. 2021 41*, vol. 4, no. 1, pp. 87-, May 2021, doi: 10.1038/s41746-021-00456-x.

[9] N. Shakhovska, V. Yakovyna, V. Chopyak, N. Shakhovska, V. Yakovyna, and V. Chopyak, "A new hybrid ensemble machine-learning model for severity risk assessment and post-COVID prediction system," *Math. Biosci. Eng. 2022 66102*, vol. 19, no. 6, pp. 6102–6123, 2022, doi: 10.3934/MBE.2022285.

[10] D. K. Sharma, M. Subramanian, P. Malyadri, B. S. Reddy, M. Sharma, and M. Tahreem, "Classification of COVID-19 by using supervised optimized machine learning technique," *Mater. Today Proc.*, vol. 56, pp. 2058–2062, Jan. 2022, doi: 10.1016/J.MATPR.2021.11.388.

[11] H. Coppock, A. Gaskell, P. Tzirakis, A. Baird, L. Jones, and B. Schuller, "End-to-end convolutional neural network enables COVID-19 detection from breath and cough audio: a pilot study," *BMJ Innov.*, vol. 7, no. 2, pp. 356–362, Apr. 2021, doi: 10.1136/BMJINNOV-2021-000668.

[12] P. Patwa *et al.*, "Can Self Reported Symptoms Predict Daily COVID-19 Cases?," May 2021, Accessed: Dec. 09, 2025. [Online]. Available: https://arxiv.org/pdf/2105.08321

[13] Z. Li *et al.*, "Efficient management strategy of COVID-19 patients based on cluster analysis and clinical decision tree classification," *Sci. Reports 2021 111*, vol. 11, no. 1, pp. 9626-, May 2021, doi: 10.1038/s41598-021-89187-3.

[14] R. K. Mojjada, A. Yadav, A. V. Prabhu, and Y. Natarajan, "WITHDRAWN: Machine learning models for covid-19 future forecasting," *Mater. Today Proc.*, Dec. 2020, doi: 10.1016/J.MATPR.2020.10.962.

[15] E. Fayyoumi, S. Idwan, and H. Aboshindi, "Machine Learning and Statistical Modelling for Prediction of Novel COVID-19 Patients Case Study: Jordan," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 5, pp. 122–126, May 2020, doi: 10.14569/IJACSA.2020.0110518.

[16] Y. Gao *et al.*, "Machine learning based early warning system enables accurate mortality risk prediction for COVID-19," *Nat. Commun. 2020 111*, vol. 11, no. 1, pp. 5033-, Oct. 2020, doi: 10.1038/s41467-020-18684-2.

[17] A. H. M. Hassan, A. A. M. Qasem, W. F. M. Abdalla, and O. H. Elhassan, "Visualization & Prediction of COVID-19 Future Outbreak by Using Machine Learning," *Int. J. Inf. Technol. Comput. Sci.*, vol. 13, no. 3, pp. 16–32, Jun. 2021, doi: 10.5815/IJITCS.2021.03.02.

[18] A. Alotaibi, M. Shiblee, and A. Alshahrani, "Prediction of Severity of COVID-19-Infected Patients Using Machine Learning Techniques," *Comput. 2021, Vol. 10, Page 31*, vol. 10, no. 3, p. 31, Mar. 2021, doi: 10.3390/COMPUTERS10030031.

[19] S. T. Ogunjo, I. A. Fuwape, and A. B. Rabiu, "Predicting COVID-19 Cases From Atmospheric Parameters Using Machine Learning Approach," *GeoHealth*, vol. 6, no. 4, p. e2021GH000509, Apr. 2022, doi: 10.1029/2021GH000509.

[20] V. K. Gupta, A. Gupta, D. Kumar, and A. Sardana, "Prediction of COVID-19

confirmed, death, and cured cases in India using random forest model," *Big Data Min. Anal.*, vol. 4, no. 2, pp. 116–123, Jun. 2021, doi: 10.26599/BDMA.2020.9020016.

[21] N. Rochmawati *et al.*, "Covid Symptom Severity Using Decision Tree," *Proceeding - 2020 3rd Int. Conf. Vocat. Educ. Electr. Eng. Strength. Framew. Soc. 5.0 through Innov. Educ. Electr. Eng. Informatics Eng. ICVEE 2020*, Oct. 2020, doi: 10.1109/ICVEE50212.2020.9243246.

[22] M. H. Tayarani N., "Applications of artificial intelligence in battling against covid-19: A literature review," *Chaos, Solitons & Fractals*, vol. 142, p. 110338, Jan. 2021, doi: 10.1016/j.chaos.2020.110338.

[23] S. Roy and P. Ghosh, "Factors affecting COVID-19 infected and death rates inform lockdown-related policymaking," *PLoS One*, vol. 15, no. 10, p. e0241165, Oct. 2020, doi: 10.1371/JOURNAL.PONE.0241165.

[24] K. B. Prakash, "Analysis, Prediction and Evaluation of COVID-19 Datasets using Machine Learning Algorithms," *Int. J. Emerg. Trends Eng. Res.*, vol. 8, no. 5, pp. 2199–2204, May 2020, doi: 10.30534/IJETER/2020/117852020.

[25] H. R. Niazkar and M. Niazkar, "Application of artificial neural networks to predict the COVID-19 outbreak," *Glob. Heal. Res. Policy 2020 51*, vol. 5, no. 1, pp. 50-, Nov. 2020, doi: 10.1186/S41256-020-00175-Y.

[26] X. Guan *et al.*, "Clinical and inflammatory features based machine learning model for fatal risk prediction of hospitalized COVID-19 patients: results from a retrospective cohort study," *Ann. Med.*, vol. 53, no. 1, pp. 257–266, 2021, doi: 10.1080/07853890.2020.1868564;ISSUE:ISSUE:DOI.

[27] D. Darwin, D. Christian, W. Chandra, and M. Nababan, "Comparison of Decision Tree and Linear Regression Algorithms in the Case of Spread Prediction of COVID-19 in Indonesia," *J. Comput. Networks, Archit. High Perform. Comput.*, vol. 4, no. 1, pp. 1–12, Jan. 2022, doi: 10.47709/CNAHPC.V4I1.1234.

[28] M. S. Satu *et al.*, "Short-Term Prediction of COVID-19 Cases Using Machine Learning Models," *Appl. Sci. 2021, Vol. 11, Page 4266*, vol. 11, no. 9, p. 4266, May 2021, doi: 10.3390/APP11094266.

[29] D. Chumachenko, I. Meniailov, K. Bazilevych, T. Chumachenko, and S. Yakovlev, "Investigation of Statistical Machine Learning Models for COVID-19 Epidemic Process Simulation: Random Forest, K-Nearest Neighbors, Gradient Boosting," *Comput. 2022, Vol. 10, Page 86*, vol. 10, no. 6, p. 86, May 2022, doi: 10.3390/COMPUTATION10060086.