

A ROUGH SET BASED CLASSIFICATION OF PANDAMIC PATIENTS: A CASE STUDY OF COVID-19

Asadullah Khan*

*Department of Information Science and Engineering,
Sri Sidhartha School of Engineering, Kartanaka, India*

*For correspondence. (hmsit.asad@gmail.com)

Abstract: COVID-19, caused by a novel strain of corona virus, has emerged as one of the most severe public health crises in modern history. Speedily growing across the world, it claimed lots of lives and infected another more. Like many viruses, the corona virus has undergone continuous mutations, producing variants such as Delta and Omicron that are highly infectious and responsible for large-scale fatalities in several countries. Although the rate of infection has declined in recent months, the pandemic has left a lasting impact on societies, economies, and everyday life. Until recently, governments and scientists worldwide faced the enormous challenge of containing the rapid transmission of the disease. Despite the wide availability of vaccines, the virus continues to evolve into stronger variants with higher transmission rates, as observed in recent surges in China and South Korea. At the beginning, symptoms were mostly mild fever, headache, and dry cough. Over time, however, a significant number of asymptomatic cases were reported, complicating detection and control measures. In the early stages, treatment approaches were largely symptom-based. As with other pandemics, the exponential growth of infections often led to acute shortages of medical resources. This situation was particularly visible during the first and second waves of COVID-19, when even technologically advanced nations with modern healthcare infrastructure struggled, resulting in high mortality. The key challenge in such scenarios is the optimal utilization of limited medical resources to reduce adverse outcomes. A promising way to address this issue is through the application of classification algorithms that can prioritize patients based on severity. In this article, we propose a rough set theory-based (RST) classification algorithm to categorize COVID-19 positive patients into predefined groups, enabling healthcare providers to manage each category more effectively. The proposed approach was validated by the experimental study on a real dataset.

Keywords: Corona Virus; Information table; Decision table; α -neighborhood; Neighborhood relation.

1. Introduction:

World Health Organization (WHO) [1] said “coronaviruses are a large family of viruses that can infect both humans and animals”. In human body viruses cause respiratory illnesses changing from the common cold to more severe diseases. In 2019 there was an outbreak of coronavirus that resulted the disease COVID-19 with symptoms like fever, dry cough, and fatigue. In addition, several less common symptoms were reported, such as body aches, nasal congestion, headache, sore throat, conjunctivitis, skin rashes, loss of smell or taste, and discoloration of fingers or toes. These were insignificant symptoms which appeared gradually. A particularly concerning development has been the emergence of a large number of asymptomatic cases across various countries, which has complicated detection and control measures.

It is estimated that approximately 80% of COVID-19 patients recover without the need for specialized treatment, while about 20% of patients become seriously ill, often developing breathing difficulties that require oxygen therapy or ventilator support. Another critical feature of COVID-19 is its extremely high transmission rate [2, 3]. In 2020 WHO [1] identified its spread and risk declared as a Public Health Emergency of International Concern and named COVID-19. Since then within a few COVI-109 spread to almost every country, resulting huge life loss.

Because coronaviruses mutate over time, several variants of COVID-19 have emerged. Based on their characteristics and potential risk, the WHO [4] has classified some of these variants as “Variants of Interest” and placed them under close monitoring. Among these, several well-known variants have drawn worldwide attention and are discussed in detail below.

Alpha (B.1.1.7): This variant was identified initially among patients of south eastern England. Primary research demonstrated that the variant could increase 70% transmissibility posing a maximum mortality risk. [4].

Beta (B.1.351): This was easier to spread than the original strain and detected mostly in South Africa and Nigeria. [4].

Gamma (P.1): This variant has been observed to be more transmissible than earlier strains and has the ability to reinfect individuals who have previously recovered from COVID-19. It was primarily identified among patients in Brazil. [4].

Delta (B.1.617.2): This detected first in India during October 2020 and was responsible for a huge increase in cases during the April-June, 2021. It transmitted to many countries then, causing millions of deaths. [4].

Omicron (B.1.1.529): This variant normally does not cause severe illness in comparison others and fully vaccinated people have stronger immunity against it. [4].

Historical evidence from previous pandemics shows that when infection rates rise exponentially, countries with poorly resourced healthcare systems tend to suffer the most severe consequences. The same was observed during COVID-19, particularly in the second quarter of last year, when countries such as India, Indonesia, and Bangladesh faced devastating impacts from the Delta variant. Even highly developed nations such as Italy, Spain, and the United States—despite having advanced medical infrastructure—struggled to cope with the overwhelming number of patients during the first wave.

Although vaccines have been developed, low- and middle-income countries with limited resources faced acute shortages due to restricted vaccine procurement, leaving them highly vulnerable to widespread infection [5]. More recently, reports of breakthrough infections among vaccinated individuals, along with evidence of transmission from vaccinated people, have become a serious concern. Furthermore, in some regions, delays or reductions in COVID-19 testing contributed to increased infection rates and, consequently, higher mortality [5]. Early studies, however, demonstrated that timely testing combined with immediate quarantine significantly reduced both transmission and fatality rates [6]. Such measures remain essential until a substantial percentage of the population is vaccinated—a goal that remains distant in many parts of the world.

To resolve these issues, RST-based classification [7, 8] approach was proposed. RST-based classification was used for discrete datasets, as it depends on equivalence relations. Though continuous data can also handle with compromising the compositionality, leading to the extended techniques [9, 10, 11]. In this context, investigators in [12] suggested an NRSC approach for medical diagnosis.

In this article, an algorithm is proposed for the rapid detection and diagnosis of COVID-19. A key strength of the proposed approach is its potential to assist severely affected countries in utilizing their limited medical resources more efficiently, thereby reducing fatality rates. COVID-19 patients can be identified through their medical records, which typically contain both numerical and categorical symptom values. Accordingly, COVID-19 datasets consist of patient records where symptoms serve as attributes of mixed data types.

To analyze these datasets, we define a metric [13] for measuring symptomatic similarity among patients based on their records. Using this metric, we construct neighborhood relations, which in turn generate lower and upper approximation spaces as well as boundary regions. These regions give decision rules which are required to support the classification process.

The article is organized as follows: Section 2 analyses current advances in this area. Section 3 presents the related terminology. Section 4 presents the methodology of the proposed approach. Section 5 discourses the results, Finally, Section 6 accomplishes the paper with a summarized conclusions and inferences.

2. Related works:

Coronaviruses constitute a large family of viruses capable of causing infections in both animals and humans [1]. The most recently emerged coronavirus has caused the disease COVID-19, which has infected millions of people worldwide and led to high mortality rates. COVID-19 is extremely communicable [2, 3], which prompted the WHO [1] to announce it a pandemic.

The global spread of COVID-19 has posed an unprecedented public health challenge [14]. Nearly every country has been affected, with nations such as the USA, Italy, Spain, the United Kingdom, and Brazil among the hardest hit. Although vaccines have been developed and made available, a significant portion of the global population remains unvaccinated. Until sufficient vaccination coverage is achieved, health authorities continue to advocate for strong social distancing, lockdown measures, and rapid medical interventions for infected individuals. Some countries have faced particular difficulty in controlling the spread and mortality of COVID-19 due to the age distribution of their populations and the prevalence of comorbidities [15].

Several computational models were proposed for the prediction and management of pandemic. For instance, [16] presented a model to forecast COVID-19 trends over a six-week period, while [17] analyzed the relative effects of lockdown stages on case numbers in the UK, USA, and India, with potential applicability to other countries. In [18], an extended SEIR model was introduced to estimate the duration of social distancing required to reduce transmission. Treatment strategies have also been explored: [19] proposed early CPAP therapy to reduce mortality, [20] studied risk factors associated with age and pre-existing chronic conditions, [21] developed a highly sensitive method for direct detection of SARS-CoV-2, and [22] recommended RT-PCR testing to shorten hospital admission periods for COVID-19 patients.

As observed in previous pandemics, countries with limited medical resources suffer disproportionately when infections rise exponentially [5]. This pattern was evident during both the first and second waves of COVID-19, particularly in regions where testing was delayed or conducted at minimal rates, increasing population exposure. Early testing and immediate quarantine have been shown to significantly reduce transmission and fatality rates [6]. In this context, RST-based classification [7, 8] offers a promising approach to manage such challenges effectively.

Even though infection rates are gradually declining, the post-COVID-19 impact remains substantial, affecting the socio-economic life of many countries. For example, [23] reported the first fatal case of rhabdomyolysis associated with COVID-19 vaccination and highlighted autoimmunity as a significant risk factor. Children, both physically and emotionally, are particularly vulnerable during pandemics. Detailed studies on the neurocognitive and asthmatic impacts on children have been presented in [24, 25], while [26] examined the effects of COVID-19 on consultation methods and treatment outcomes for patients with hepatocellular carcinoma in the Asia-Pacific region.

In the 1980s, Pawlak [7] introduced RST as a mathematical framework to address vagueness and uncertainty in datasets. Extensions to handle continuous and complex data have been proposed, such as the neighborhood rough set-based classification (NRSC) algorithm for medical diagnosis [12]. RST has also been applied specifically to COVID-19; for instance, [27] used it to predict symptomatic cases, and [28] employed a hybrid approach combining neighborhood rough set classification with backpropagation neural networks to classify COVID-19 images. Similar works related RST were reported in [29-31]

3. Problem Statement:

In this section, we present the fundamental definitions and notations employed in the proposed algorithm. Let $\mathbf{U} = \{x_1, x_2, \dots, x_m\}$ denote the dataset comprising records of m COVID-19 positive patients, where each record x_i is represented as an n -dimensional feature vector. The feature space has k numeric and $(n-k)$ categorical attributes. Each attribute value is assigned as $\emptyset$ for asymptomatic patients, whereas other values correspond to specific numeric or categorical descriptors as defined by domain experts.

Definition 3.1. Metric Between Two COVID-19 Records:

To formalize the metric function, we adopt the formulation described in [13]. Let $A = \{a_1, a_2, \dots, a_k, a_{k+1}, a_{k+2}, \dots, a_n\}$ denote attribute set, where the first k are numeric and the remaining are categorical. Let x and y represent the records corresponding to two COVID-19-positive patients. These records may be either empty or non-empty, depending on whether the respective patients are asymptomatic or symptomatic. For asymptomatic patients, we define $x = y = \emptyset$. For symptomatic patients, the records are expressed as $x = \{a_1(x), a_2(x), \dots, a_k(x), a_{k+1}(x), \dots, a_n(x)\}$, $y = \{a_1(y), a_2(y), \dots, a_k(y), a_{k+1}(y), \dots, a_n(y)\}$, where $a_i(x)$ and $a_i(y)$ are attribute values a_i in x and y , respectively. The metric distance between the two records, denoted by $d(x, y)$, is defined as $d(x, y) = (1/n) \sum \delta_i(x, y)$, where $\delta_i(x, y) = |a_i(x) - a_i(y)| / (R_i + \varepsilon)$, if a_i is numeric, $\delta_i(x, y) = 0$, if a_i is categorical and $a_i(x) = a_i(y)$, and $\delta_i(x, y) = 1$, if a_i is categorical and $a_i(x) \neq a_i(y)$. And R_i is the range of the numeric attribute a_i . A constant ε is included in the denominator is small positive number that is required to avoid division by zero. By construction, $d(x, y) \in [0, 1]$. In particular, $d(x, y) = 0$ indicates that the two records possess identical attribute values, signifying complete similarity between the corresponding patient profiles.

Definition 3.2: α -Neighborhood:

For all $x_i \in U$ and $0 \leq \alpha \leq 1$, the neighbourhood relation (U, d) is characterized as: $\alpha(x_i) = \{x \mid d(x_i, x) \leq \alpha\}$. Here, $\alpha(x_i)$: the α -neighborhood of x_i , indicating the set of records with at most α distanced from x_i .

Definition 3.3: Neighborhood Decision System (NDS):

Let (U, AUD) denote a decision system, where U represents the universe of COVID-19 patient records, A : the conditional attributes, and D : the decision attributes. For $\alpha < 1$, the A generates an α -neighbourhood relation N . The α -neighbourhood decision system with respect to α is defined as: $NDS = (U, AUD, \alpha)$

Definition 3.4: Lower and Upper Approximations:

Let $B \subseteq A$ and $X \subseteq U$. The lower and upper approximations of X in terms of the neighbourhood relation N with respect to B are defined as: Lower Approximation: $B(X) = \{x \in U \mid B(x) \subseteq X\}$, Upper Approximation: $\overline{B}(X) = \{x \in U \mid B(x) \cap X \neq \emptyset\}$, where $B(x)$ is a sub-vector of $A(x)$ consisting of all attribute values belonging to $B \subseteq A$. The boundary region is defined as a difference of lower approximation from the upper. The neighbourhood lower approximation, the union of lower approximations of each decision class, is labelled the positive region. The boundary region supports to lessen uncertainty in decision making.

4. Proposed Method:

To extract classification rules, a suitable value of α is first selected to generate the α -neighbourhood relation. Let the dataset has m patient records, with n attributes, represented by an $m \times n$ matrix $[x_{ij}]$, where x_{ij} denotes the value of the j -th attribute for the i -th patient ($i = 1, 2, \dots, m; j = 1, 2, \dots, n$). Generally, a supervised dataset is represented as (U, AUD) , where $U = \{x_1, \dots, x_m\}$, $A = \{a_1, \dots, a_n\}$, and $D = \{d_1, \dots, d_r\}$, are the conditional and decision attributes, respectively. The flowchart of the algorithm is given Figure 1.

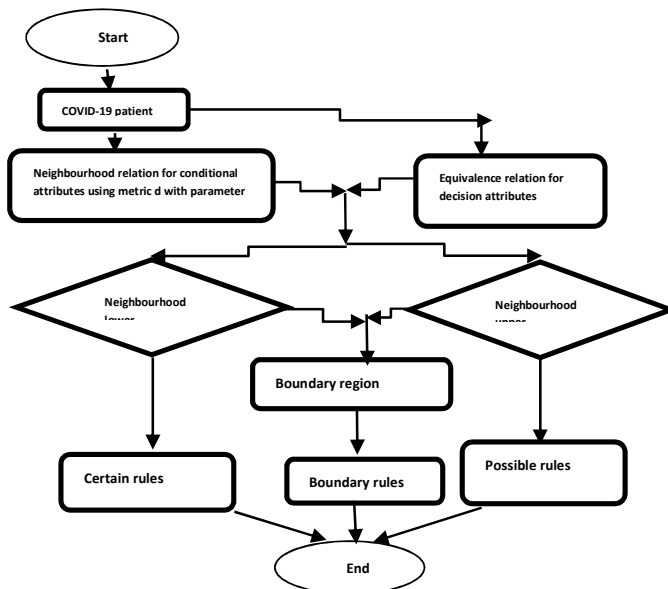


Figure 1: Flowchart of the proposed method

5. Experiments and Discussion:

Experiments were conducted with COVID-19 Surveillance Dataset collected from [32].

The dataset contains 14 instances and 7 conditional attributes representing symptoms, with two decision classes: D1: Cured and Released, D2: Deceased.

Table 1: Dataset Description

Dataset	Dataset char	No of instances	No of conditional attributes (symptoms)	No of decision attributes
COVID-19 Surveillance Datasets	Multivariate	14	7	2

The proposed method was implemented in MATLAB. Two random samples (7 instances each) were used for training and testing, repeated 20 times. Performance was evaluated using Precision, Recall, F-measure, Rand Index, and Fowlkes–Mallows Index, and compared with DT, RFC, NBC, KNNC, and SVM.

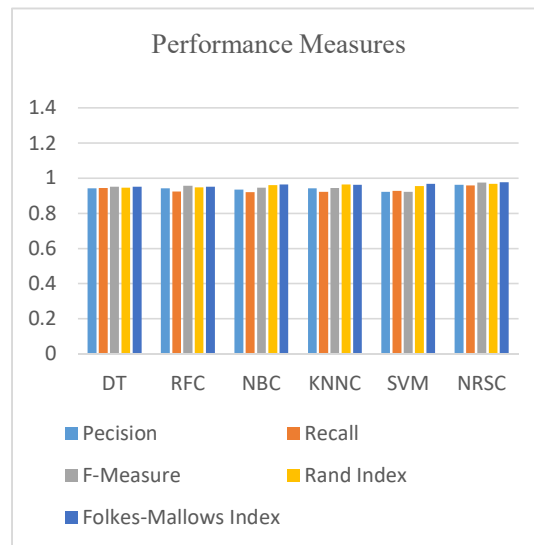


Figure 2: Comparative study on performance measures

Table 2: Comparative Evaluation of different methods in performance metrics

	DT	RFC	NBC	KNNC	SVM	NRSC
Accuracy	84%	92%	785%	93%	83%	97%
Sensitivity	84%	90%	77%	92%	81%	96%
Specificity	87%	91%	74%	92%	83%	97%
Error rate	2.5%	2.1%	2.8%	1.1%	1.9%	2.1%

Results showed that NRSC outclasses the other approaches across many performance metrics.

6. Conclusion:

This study offers an NRSC method to categorize COVID-19 patients efficiently. By incorporating both numeric and categorical attributes and using a unified metric an α -neighborhood relation is constructed. The α -neighborhood relation computes the lower and upper approximation spaces, and detects boundary regions. Lower and upper approximation provides the certain and possible rules to classify patient records efficiently. Experimental results with real dataset validate the superiority of NRSC over the conventional classifiers in accuracy, sensitivity, and specificity, providing a trustworthy decision-support tool for hospitals to optimize patient management.

References:

- [1] <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/q-a-coronaviruses>
- [2] World Health Organization; “Infection prevention and control of epidemic- and pandemic-prone acute respiratory infections in health care. Geneva: World Health Organization; 2014” https://apps.who.int/iris/bitstream/handle/10665/112656/9789241507134_eng.pdf?sequence=1.
- [3] J. Liu, X. Liao, S. Qian, J. Yuan, F. Wang, Y. Liu, Z. Wang, Fu-Sheng Wang, L. Liu, Z. and Zhang, Community transmission of severe acute respiratory syndrome corona virus 2, Shenzhen, China, 2020. Emerging Infectious Diseases, Volume 26 (6).
- [4] <https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/>
- [5] J. Hopman, B. Allegranzi, and S. Mehtar, Managing COVID-19 in Low- and Middle-Income Countries, JAMA, 323(16):1549-1550, 2020.
- [6] <https://jamanetwork.com/journals/jama/fullarticle/2762130>
- [7] Z. Pawlak, Rough sets, International Journal of Computer and Information Sciences, vol. 11, 341–356, 1982.
- [8] R. R. Nowicki, Rough Set Based Classification Systems, Springer, 2019.
- [9] Q. Hu, D. Yu, Z. Xie, J. Liu, Fuzzy probabilistic approximation spaces and their information measures, IEEE Transaction of Fuzzy System, vol. 14, 191–201, 2006.
- [10] Y. Yao, Probabilistic rough set approximations, International Journal of Approximate Reasoning, vol. 49, 255–271, 2005.
- [11] R. Slowinski, D. Vanderpooten, A generalized definition of rough approximations based on similarity, IEEE Transactions on Knowledge and Data Engineering, vol. 12, 331–336, 2000.
- [12] S. Udhaya Kumar, and H. Hannah Inbarani, A Novel Neighborhood Rough Set Based Classification Approach for Medical Diagnosis, Procedia Computer Science, 47, 351-359, 2015.
- [13] F. A. Mazarbhuiya, M. Y. AlZahrani, and G. Lilia, Anomaly Detection using Agglomerative Hierarchical Clustering Algorithm, Information Science and Application, LNEE, Vol. 514, 475-484, 2018.
- [14] S. Nathaniel and M. Fraser, Straining the system: Novel coronavirus (COVID-19) and preparedness for concomitant disasters, 648-649, 2020.
- [15] N. Pearce, et al, Is death from COVID-19 a multistep process?, medRxiv preprint doi: <https://doi.org/10.1101/2020.06.01.20116608>.
- [16] S. Basu, Model Based Covid-19 Case Studies in the UK, the USA and India, medRxiv preprint doi: <https://doi.org/10.1101/2020.05.31.20118760>
- [17] M. Ruzhansky, et al, A Projection Model Of Covid-19 Pandemic For Belgium, medRxiv preprint doi: <https://doi.org/10.1101/2020.05.31.20118406>.
- [18] N. Rakocz, et al, A Statistical Model for Quantifying the Needed Duration of Social Distancing for the COVID-19 Pandemic, medRxiv preprint doi: <https://doi.org/10.1101/2020.05.30.20117796>.
- [19] A. Ashish, et al, Early CPAP reduced mortality in covid-19 patients. Audit results from Wrightington, Wigan and Leigh Teaching Hospitals NHS Foundation Trust, medRxiv preprint doi: <https://doi.org/10.1101/2020.05.28.20116152>

- [20] P. M. McKeigue, et al, Rapid Epidemiological Analysis of Comorbidities and Treatments as risk factors for COVID-19 in Scotland (REACT-SCOT): a population-based case-control study, medRxiv preprint doi: <https://doi.org/10.1101/2020.05.28.20115394>
- [21] S. Wei, et al, Direct diagnostic testing of SARS-CoV-2 without the need for prior RNA extraction, medRxiv preprint doi: <https://doi.org/10.1101/2020.05.28.20115220>.
- [22] S. Paramita, et al, Clinical features of COVID-19 patients in Abdul Wahab Sjahranie Hospital, Samarinda, Indonesia, medRxiv preprint doi: <https://doi.org/10.1101/2020.05.27.20114348>.
- [23] K. M. Ajmera, Fatal Case of Rhabdomyolysis Post-COVID-19 Vaccine, Infection and Drug Resistance, Vol. 2021(14), 3929-3935.
- [24] R. Rizzo, L. Karlov, N. Maugeri, S. Disilvestre, V. Eapen, Impact of the COVID-19 Pandemic on Family Wellbeing in the Context of Neurodevelopmental Disorders, Neuropsychiatric Disease and Treatment, Vol. 2021 (17), 3007—3014.
- [25] A. A. Asseri, Pediatric Asthma Exacerbation in Children with Suspected and Confirmed Coronavirus Disease 2019 (COVID-19): An Observational Study from Saudi Arabia, Journal of Asthma and Allergy, Vol. 2021(14), 1139—1146, 2021
- [26] M. Gandhi, WH. Ling, CH. Chen, JH. Lee, M. Kudos, R. Chanwat, SI. Strasser, Z. Xu, SH. Lai, PKH. Chow, Impact of COVID-19 on Hepatocellular Carcinoma Management: A Multicountry and Region Study, Journal of Hepatocellular Carcinoma, 2021 Volume 2021(8), 1159—1167.
- [27] H. R. Bhapkar, P. N. Mahalle, G. Shinde, and M. Mahmud (2021), In Book: COVID_19: Predicyion, Decision Making and Its Impact.
- [28] S. Nivetha, and H. H. Inbarani, Neighborhood Rough Neural Network Approach for COVID-19 Image Classification, Neural Processing Letters, 2022.
- [29] F. A. Mazarbhuiya, and M. Shenify, An Intuitionistic Fuzzy-Rough Set-Based Classification for Anomaly Detection, Applied Science, MDPI, Vol. 13(9), 2023, pp.1-21.
- [30] Mohamed Shenify, F. A. Mazarbhuiya and A. S. Wungreiphi, Detecting IoT Anomalies Using Fuzzy Subspace Clustering Algorithms, Applied Science, MDPI, 14(3), 2024, pp. 1-30.
- [31] F. A. Mazarbhuiya, M. Shenify, A.S. Wungreiphi, J. Kez, and R. Paul, Detecting Alzheimer's Disease using Rough Set-based Classification Approach, Engineered Science, Vol. 37, 1736, Engineered Science Publisher, 2025, pp. 1-10.
- [32] B. Desgroupes, Clustering Indices - CRAN, University of Paris Quest-Lab'X, , 2017, 1-34.