Enhanced Diagnosis of the COVID-19 Behaviour Using the Rough Set Theory and Genetic Algorithms

Mansour Esmaeilpour (ma.esmaeilpour@gmail.com)
IAU: Islamic Azad University  https://orcid.org/0000-0002-2475-518X

Rasha Kashef
Ryerson University

Research Article

Keywords: Rough Set, Data Analysis, Genetic Algorithm, Covid-19, Feature Reduction

Posted Date: November 23rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-734842/v1

License: ☺  This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
Abstract

The outbreak of the coronavirus 2019 (COVID-19) has created an excellent challenge for the care system worldwide. One in every of the foremost vital points of this challenge is that the management of COVID-19 patients needing acute and/or vital metastasis care. The main objective of applying data mining to Covid-19 dataset is essential to propel learning by empowering data-oriented decision making to improve existing clinical practices and learning materials. Current data mining techniques offer patient data analysis for achieving an automated diagnosis of the diseases as an example; however, the results are not very accurate nor reliable, especially with a dynamic virus as the COVID-19. In this paper, we are proposing a multi-stage diagnostic (MSD-Covid19) model to enhance the diagnosis of the COVID-19, and to provide a sustainable automated system to improve the healthcare systems and patient outcomes. The first stage includes a selection of a classification model with no reduction attributes. Tested classification algorithms include Deep learning, Multilayer Perceptron, KNN, Bayesian Auto Regression, Logistic Model Trees (LMT), Hoeffding tree (VFDT), and Fuzzy Unordered Rule Induction Algorithm. In the second stage, a rough set reduction algorithm based on genetic algorithms is employed, and finally, an optimization of the classification is conducted using the reduced attributes. The proposed model is evaluated on a global COVID-19 dataset. Experimental results demonstrate that the proposed MSD-Covid19 has a great contribution to increase the diagnostic accuracy of the COVID-19 disease behaviour.

1. Introduction

A Coronavirus is a large group of viruses that incorporates regular cold infection. COVID-19 is another respiratory virus that started in late 2019 and mid-2020 in Hubei Province and the Chinese city of Wuhan, killing many people. The virus is known as the Severe Acute Respiratory Syndrome (SARS-CoV-2), and the World Health Organization (WHO) has named the virus COVID-19 after more than 1,000 people have died [1]. Even with any emergency, the most significant approach to defeat it is to know the problematic itself and know about its belongings. To perform better, it is essential to find out about this dangerous virus. This information has numerous measurements; one of its most critical measurements is the consciousness of the measure of harm, both humanly and socially. One reason a few governments can show improvement over others is to have the information, data, models and abilities to manage the emergency.

There square measure many ways that to research the COVID-19 information, like time-series analysis, simulation and modelling, benchmarking, data processing, etc. Data mining can be a process of analyzing information to get patterns and data within the datasets in a very human-understandable structure. Generally, Data mining is that the process of analyzing information from totally different views and summarizing it into helpful info that may be wont to increase revenue, cuts costs, or enhance system’s performance. Data Analysis is associate degree approach to resolve issues, issues, or events through a data perspective to know information at a deeper level and type new insights regarding them. To perform a data analysis task, a suitable machine learning method must be selected. Some machine learning methods are appropriate for some data structure, and otherwise. Therefore, a machine learning
technique should be chosen for the data based on a particular purpose, and each machine learning technique may announce different accuracy with various performances. The set of features in dataset represent a fingerprint for each of dataset. These features may be continuous, discrete, or categorical. All data mining models, regardless of whether to foresee the scattering of sentiment, to comprehend the conduct of networks, or to study the spread viruses, contain factors and features that should be considered and studied carefully.

The main objective of applying data mining to the COVID-19 dataset is to propel learning by empowering data-oriented decision making to improve existing clinical practices and learning materials. Several data mining studies [2] have been developed over the last few months to monitor, trace, and diagnosis the spread of the deadly virus, however none of these studies have provided a deeper and accurate analysis of the models adopted, features, and data types. In addition, Current data mining techniques offer patient data analysis however the results are not very accurate, have limited results due to the shape of data, and not reliable due to the dynamic change of the virus. Thus, we propose a multi-stage model to improve the critical care of the COVID-19 patients by providing a sustainable automated model to provide an efficient diagnosis of the COVID-19 patterns, and to discover associated factors or features that are highly correlated to COVID-19. This paper focuses on finding the key factors affecting the performance of the COVID-19 clinical diagnosis. The findings of this paper will positively impact future decisions about the progress of the patients’ outcomes, quality of the clinical process, and the future of the clinical provider. The first stage in the model is to discover the impact of all the features of the COVID-19 dataset in selecting a proper classification model, and the second one is the reduction step. The third stage is to acquire an effect of the selected properties after reduction, which has increased the accuracy.

The rest of this paper is organized as follows. Section 2 describes the related work and background on the Covid-19 pandemic. Sections 3 describe the rough set theory and reduction attributes. Sections 4 introduce the proposed Multi-stage diagnosis algorithm. Section 5 and Sect. 6 discuss the experimental dataset and framework used to evaluate the performance of the proposed algorithm. Finally, Sect. 7 presents conclusions and future directions.

2. Related Work And Background

Data mining is utilised to trace and predict however the COVID-19 complaint can spread and reshape over time. as an example, following a past pandemic, that of the 2015 Zika- virus, Akhtar et al. Set up a neural network to predict its spread, such the developed methods ought to be re-prepared utilizing knowledge from the COVID-19 pandemic [3].

Since the outbreak of COVID-19, try to develop learning methods for COVID-19 screening based on medical images such as CT scans and deep learning. Huang et al. Assessed the severity of pulmonary manifestations of COVID-19 through chest CT using a deep learning method. According to the deep learning algorithm, in patients with different clinical severity, there was a significant difference in the percentage of lung opacity. This automated tool for quantifying lung involvement may be used to
monitor disease progression and understand the temporal evolution of COVID-19. Patients with COVID-19 who underwent chest CT between 1 January 2020 and 3 February 2020 were screened. Patients were divided into mild, moderate, severe and critical types according to the initial clinical, laboratory and CT findings. Percentage of CT of the lung was divided into mild, moderate, severe and critical according to the initial clinical, laboratory and CT findings. Percentage of CT of the whole lung and five lobes was automatically measured by deep learning and compared with CT scan follow-ups. Longitudinal changes in the quantitative parameter of CT were also compared among the four clinical types [4].

After proving computed tomography (CT) as a useful way to diagnose COVID-19 patients. Yang et al. found that publicly available COVID-19 CT datasets were incomplete due to privacy issues. They have a small number of samples, which hinders the research and development of CT-based COVID-19 diagnostic methods [5].

Zhang et al. in a retrospective analysis of CT findings in patients infected with coronavirus 2019 (COVID-19) in a collection that included thirty-four cases, 15 women and 19 men, ranging in age from 7 to 88 Years, which were confirmed by reverse transcriptase-reverse polymerase chain reaction (RT-PCR), were used for the study, and all thin lung CT scans were performed in all patients. Clinical, laboratory and CT imaging were available for evaluation in all patients. Preliminary CT scans have been shown to be important for early detection and assessment of COVID-19 disease progression. [6]

Young et al. developed a deep learning-based CT diagnostic system to help physicians identify patients with COVID-19 [7]. They developed an in-depth learning algorithm by modifying the transfer-learning model for clinical presentation [8]. Shi et al. proposed an infection-conscious randomized forest approach (iSARF) that could automatically classify individuals into groups with different ranges [9]. Narin et al. proposed three different deep learning models, ResNet 50, InceptionV3, and Inception-ResNetV2, to detect COVID-19 infection from X-ray images. It is noteworthy that the existing COVID-19 dataset, which consisted of X-ray images of 50 COVID-19 patients and 50 normal breast images, is available in Kaggle. Evaluation results show that the ResNet50 model was reasonably accurate [10].

Xu et al. in response to the question: Can artificial intelligence technology be used for early detection of COVID-19 patients by computed tomography (CT) images and what is the diagnostic accuracy of the computer? A case study was conducted with the aim of establishing an initial screening model to distinguish COVID-19 pneumonia from influenza-A viral pneumonia and healthy cases with pulmonary CT images using deep learning. By doing this research, they proved that this fully automated method could be a promising complementary diagnostic method for front-line clinicians [11].

Hosseini et al. to control the epidemic and prevent COVID-19 infection. They developed an efficient optimization algorithm, which can solve NP-hard in addition to applied optimization problems. They first proposed the COVID-19 optimization algorithm (CVA) to cover almost all operational areas of optimization problems. They also simulated the distribution of the corona virus in several countries around the world. They then modeled the corona virus distribution process as an optimization problem to minimize the number of countries infected with COVID-19 and thus slow the spread of the epidemic. In
addition, they proposed three scenarios to solve the optimization problem using the factors influencing the distribution process. The simulation results show that one of the control scenarios has better performance than the others. Extensive simulations using several optimization problems show that the performance of CVA compared to volcanic eruption algorithm (VEA), gray wolf optimizer (GWO), particle swarm optimization (PSO) with a maximum of 15%, 37%, 53% and 59% increase in best performance [12].

Data mining technology can be applied to healthcare in order to build predictive models providing predictions in real environments using, for that, real clinical data. The 2019 novel coronavirus (COVID-19) presents several unique features [14, 15]. While the diagnosis is confirmed using polymerase chain reaction (PCR), infected patients with pneumonia may present on chest X-ray and computed tomography (CT) images with a pattern that is only moderately characteristic for the human eye [16]. In late January, a Chinese team published a paper detailing the clinical and paraclinical features of COVID-19. They reported that patients present abnormalities in chest CT images with most having bilateral involvement [17]. Bilateral multiple lobe and sub segmental areas of consolidation represent the standard findings in chest CT image of treatment unit (ICU) patients on admission [18]. Compared, non-ICU patients show bilateral ground-glass opacity and sub segmental areas of consolidation in their chest CT image [19]. In these patients, later chest CT image show bilateral ground-glass opacity with resolved consolidation [20]. Table 1 shows the Comparison table of reviewed methods.
Table 1  
Comparison table of reviewed methods

| Method                                      | Date | Advantages                                                                 | Disadvantages                                                                 |
|---------------------------------------------|------|-----------------------------------------------------------------------------|-------------------------------------------------------------------------------|
| U-Net [4]                                   | 2020 | Detection of pulmonary involvement in COVID-19 by CT automatically with deep learning. | Failure to generalize the algorithm to another data set.                      |
|                                             |      |                                                                             | Requires radiologist supervision.                                             |
| ResNet – 50                                  | 2020 | Create a public access database in GateHub.                                 | Need to create a better monopoly model to train and improve results.          |
| Densenet-169 [5]                             | 2020 | High accuracy of diagnosis.                                                 |                                                                               |
| ResNet50 InceptionV3                        | 2020 | Achieve high-precision detection accuracy with ResNet50.                    | Lack of generalization of the algorithm on a larger data set.                 |
| Inception ResNetV2[6]                       |      |                                                                             | The dataset used was not publicly available.                                  |
| UNet ++ [7]                                  | 2020 | Create an automated system based on UNet ++.                                | The use of a small dataset containing only a few samples of COVID-19 makes it difficult to generalize the results. |
|                                             |      | To diagnose COVID-19.                                                       |                                                                               |
| Rough Set Support Vector Machine (RS-SVM),  | 2020 | The analysis was performed based on data sets provided by the Johns Hopkins Corona Virus Center. | Failure to generalize the algorithm to a larger data set.                    |
| Bayesian Ridge and Polynomial Regression,    |      |                                                                             |                                                                               |
| S IR model, and RNN[8]                      |      |                                                                             |                                                                               |
| COVID-19 optimizer Algorithm (CVA) [12]      | 2020 | Create an efficient optimization algorithm that can solve NP-hard in addition to applied optimization problems. | For further large-scale validation, the study should be performed in several hospitals and several districts. |
| Manual Resnet 34, Decision tree [13]        | 2020 | Quick network training.                                                     | Low detection accuracy.                                                       |
|                                             |      |                                                                             | Failure to generalize the algorithm to a larger data set.                    |

3. The Rough Theory

The primary gain of the Rough set is to reach approximate significance of the earn data. This theory is a powerful mathematical tool for conflict ambiguity and uncertainty that has ways that to reduce attributes that is inapplicable or unnecessary for a dataset. This process of feature reduction in data is based on training the main functionality in the data, and without losing the basic structure of the dataset. As a
result of data reduction, a set of concise and meaningful rules is obtained that facilitates the decision-making process [28]. In the information table (dataset), each subset of attributes provides a homogeneous relationship. If a set B has a set of non-empty sets, and two objects v and u (indistinguishable by B), we will have another set U, for each feature, such that we have A_i in B, that A_i is the dataset attributes:

\[ u \equiv v \text{(mod } B) \text{ if } \rho(u, A_i) = \rho(v, A_i) \]  \hspace{1cm} (1)

The \( \equiv \) symbol represents the equivalence relationship, also called the equilibrium relationship, and is denoted by IND (B). Equivalent classes include U by \([u]\) IND (B) or more simply as \([u]\) B, where B can be a single set as B = \(\{A_i\}\). In this case, we define and have the equivalent relationship as IND (A_i):

\[ \text{IND}(B) = \cap \\{\text{IND}(A_i) : A_i \text{ in } B\} \] \hspace{1cm} (2)

The above relationship states that IND (B) equivalence classes include all possible commonalities of IND (A_i) equivalence classes. Therefore, in general, the IND (P) relationship, also known as the uncertainty relationship, can be used to show the uncertainty on the U reference set.

\[ \text{IND}(P) = \{(x, y) \in U^2 | \forall a \in P, a(x) = a(y)\} \] \hspace{1cm} (3)

The equivalence class of the relation with the characteristics of P with \([x]_p\) and the low approximation of X are obtained based on formula in Eq. 4.

\[ \text{PX} = \{x | [x]_p \subseteq X\} \] \hspace{1cm} (4)

And the equivalence relationship between P and Q on U is defined in positive areas as in Eq. 5.

\[ \text{pos}_p (Q) = \bigcup_{x \in U \setminus Q} \text{PX} \]

5

Finally, the degree of dependence of the property can be represented by Eq. 6 which indicates that: Q depends on P and its degree of dependence is equal to K (0 \( \leq \) k \( \leq \) 1).

\[ K = \gamma(Q) = \frac{|\text{pos}_p (Q)|}{|U|} \] \hspace{1cm} (6)

### 3.1 Reduction Attributes

Searching for an optimal subset involves finding features that are strongly related to decision-making but separate from each other. The choice of the optimal subgroup varies depending on the problem being considered. Usually, feature reduction algorithms are based on exploratory methods or a random search to reduce the degree of complexity and ultimately lead to a reduced subset of features [29]. This algorithm begins with giving an initial value of a reduction candidate with an empty set. Then each
property in the function of uncertainties is evaluated by exploratory measurement. The reduction algorithm generally counts the number of features that appear in the phrase and considers the attribute that seems the most important. The features that have the highest amount of exploration are added to the reduction candidate. The whole phrase deleted features that are included in the undefined function are removed. As soon as all the expressions are removed, the algorithm terminates and returns the reduction candidate [30].

3.2 The Genetic Algorithms

Genetic Algorithms are random and evolutionary process search techniques supported the principles of biological evolution, natural process, and genetic recombination. It simulates the principle of ‘survival of the fittest’ during a population of potential solutions referred to as chromosomes. Each chromosome represents one attainable answer to the matter or a rule a classification. [31]. Genetic algorithms cannot directly handle the info of answer area. Thus, it should categorical them as genotype string-shaped data in genetic area through cryptography. Then, a fixed-length string of binary symbols is employed to express individuals [32].

5. The Proposed Multi-stage Covid-19 Diagnosis (Msd-covid19) Algorithm

In this paper, we employ a data mining-based model to improve the critical care of COVID-19 patients. The proposed method uses feature reduction to increase the accuracy of the model and to identify factors affecting COVID-19 diagnosis. The multi-stage model consists of three steps, which are fully explained below.

5.1 Classification Model with No Reduction Attributes

The first step is to select a classifier that performs the best in diagnosing COVID-19. Eight classification models on the Covid-19 dataset are employed with No reduction attributes. We have used Deep learning, Multilayer Perceptron (MLP), k-Nearest Neighbours (KNN), Bayesian Auto Regressions (BARs), Logistic Model Trees (LMT), Hoeffding Tree (VFDT), and Fuzzy Unordered Rule Induction Algorithm (FURIA). Each model was implemented separately on the COVID-19 dataset with 10-fold cross-validation, and the following results were obtained. Some of the classification algorithms adopted in the paper are described next.

Deep learning could be a branch of machine learning that generates multi-layered representations data, normally victimisation artificial neural networks. It’s improved the progressive in numerous machine learning tasks. This analysis has utilized Deep learning [21] that supports absolutely connected feed-forward networks, convolutional networks, and perennial networks.

Multilayer Perceptron (MLP), as a nonlinear learning and modelling tool, has been with success employed in a broad vary of applications. In theory, such applications are 3 basic categories: operate estimation
Bayesian Auto Regressions (BARs) square measure linear variable time-series models ready to capture the joint dynamics of multiple statistic. Bayesian illation treats the power unit parameters as random variables. It provides a framework to estimate “posterior” likelihood distribution of the situation of the model parameters by combining data provided by a sample of ascertained knowledge and previous data derived from a spread of sources, like alternative macro or small datasets, theoretical models, alternative economics phenomena, or contemplation [23]

The Fuzzy Unordered Rule Induction Algorithm (FURIA) algorithm is proposed by Huhn and Hullermeier; since it is based on fuzzy rules, models the decision limits by making them more flexible, with fuzzy intervals, generating non-ordered rules in place of the regular list of rules [24].

The Hoeffding tree (VFDT) is associate degree progressive call tree induction rule capable of learning from large data, presumptuous that the distribution generating examples don’t modification over time. This idea is supported mathematically by the Hoeffding certain, that quantifies the amount of observations required to estimate some statistics among a prescribed preciseness. Victimization the Hoeffding certain, one will show that its output is sort of the image of that of a non-incremental learner victimization infinitely several examples [25].

The Logistic Model Trees (LMT) is a classification method with associate supervised learning algorithmic that mixes logistic regression and decision tree learning. Logistic model trees square measure supported the sooner plan of a model tree [26].

The k-Nearest Neighbours algorithm (KNN) is a non-parametric methodology planned by Cowl used for classification and regression. In each case, the input is the k nearest instances within the feature space. The output of k-NN is employed for classication or regression [27].

5.2. Stage2 (Feature Reduction Using Rough Set Based on GA)

So as to beat the challenges in feature reduction with large number of features, genetic algorithm are employed to reach the minimal features of data set by consolidating its exceptional capacity for generally using the rough set theory. This is an evolutionary model for function optimization based on the biological progress. Existing reduction algorithms, start from rough set principle to achieve reduction set of the features. S = (U, R) is a data table and S is a subset of U and it’s a set of features, as well as, R is condition features. Decision information is represented by S = (U, R ∪ {D}) where D ∈ R that called decision feature. A set of features D is depends on the set of condition features R, where C ⇒ D. D depends on R where k (0 ≤ k ≤ 1) as follow.

\[ k = \gamma (R, D) = |(POSRD)||U|. \] (7)
5.2 Stage3 (Reclassification Analysis with reduction attributes and Genetic algorithm)

This step constructs the discernibility matrix of attributes based on genetic algorithm. The method employs each attribute to the initial strings to reach minimal attribute set as reduction subset. The rough set theory is defined as fitness function.

The Pseudo code of the proposed algorithm is explained in Fig. 1.

6. Experimental Datasets

Experimental data contain an open public dataset of chest CT images, and metadata of patients. Those patients are positive or suspected of COVID-19 or other viral and bacterial pneumonia. The exploit data is a public dataset from hospitals and physicians. All of the chest CT images are released publicly in GitHub report [33]. The data consists of 1242 images. Data pre-processing has been completed as follows: Patient information was extracted during the visit to the medical centre, and metadata information was extracted from a CT image. Each image has the following attributes Sex, Age, Survival, Intubated, Intubation Present, Went_icu, In_icu, Needed_Supplemental_O2, Extubated, Temperature, PO2_Saturation, Leukocyte_count, Lymphocyte_count, View, Modality, and Finding. Table 2 describes each features of dataset.
Table 2
COVID-19 Dataset features

| Attribute        | Description                                                                 |
|------------------|-----------------------------------------------------------------------------|
| Sex              | Male (M), Female (F)                                                        |
| Age              | Age of the patient in years                                                 |
| Survival         | Yes (Y) or no (N)                                                           |
| Intubated        | Yes (Y) if the patient was intubated (or ventilated) at any point during this illness or No (N) |
| Intubated Present| Yes (Y) if the patient is intubated Present now(ventilated) at any point during this illness or No (N) |
| Went_icu         | Yes (Y) if the patient was in the ICU (intensive care unit) or CCU (critical care unit) at any point during this illness or No (N) |
| Needed_supplemental_O2 | Yes (Y) if the patient required supplemental oxygen at any point during this illness or No (N) |
| Extubated        | Yes (Y) if the patient was successfully extubated or No (N)                 |
| Temperature      | Temperature of the patient in Celsius at the time of the image              |
| PO2 saturation   | partial pressure of oxygen saturation in % at the time of the image         |
| Leukocyte_count  | white blood cell count in units of 10^3/uL at the time of the image         |
| Lymphocyte count | lymphocyte cell count in units of 10^3/uL at the time of the image          |
| View             | Posteroanterior (PA), Anteroposterior (AP), AP Supine (APS), or Lateral (L) for X-rays; Axial or Coronal for CT scans |
| Modality         | CT, X-ray                                                                   |
| Finding          | Type of pneumonia (12 classes)                                              |

A Sample X-ray is shown Fig. 2.

7. Experimental Analysis

In the first stage of the algorithm, some of classification algorithms are adopted including Deep learning, Multilayer Perceptron (22 hidden layer), Multilayer Perceptron classifier (one hidden layer), Bayesian auto Regression, KNN, FURIA, VFDT, LMT. Table 3 shows the classification accuracy with no reduction attributed used. We have used the True Positive (TP) Rate, False Positive (FP) Rate, Precision, Recall, and F-Measure [12] as metrics of accuracy.
Table 3
Classification Accuracy of Multiple classifiers (No attributes reduction)

| Data analysis model                                      | TP Rate | FP Rate | Precision | Recall | F-Measure |
|-----------------------------------------------------------|---------|---------|-----------|--------|-----------|
| Deep Learning                                             | 0.990   | 0.991   | 0.812     | 0.990  | 0.892     |
| Multilayer Perceptron classifier (one hidden layer)       | 0.940   | 0.748   | 0.881     | 0.940  | 0.909     |
| Multilayer Perceptron (22 hidden layer)                   | 0.992   | 0.974   | 0.815     | 0.992  | 0.895     |
| Bayesian auto Regression                                  | 0.996   | 0.896   | 0.827     | 0.996  | 0.904     |
| KNN                                                       | 0.913   | 0.817   | 0.828     | 0.913  | 0.869     |
| FURIA                                                     | 0.956   | 0.878   | 0.824     | 0.956  | 0.885     |
| VFDT                                                      | 1.000   | 1.000   | 0.812     | 1.000  | 0.896     |
| LMT                                                       | 0.956   | 0.696   | 0.856     | 0.956  | 0.903     |

Results using the rough set model and genetic algorithms to reduce the number of redundant features, thus improving the classifier accuracy is shown in Table 4.
Table 4
Classification Accuracy of Multiple classifiers (With attributes reduction)

| Data analysis model                              | TP Rate | FP Rate | Precision | Recall | F-Measure |
|--------------------------------------------------|---------|---------|-----------|--------|-----------|
| Deep Learning                                    | 0.992   | 0.993   | 0.813     | 0.992  | 0.894     |
| Multilayer Perceptron classifier (one hidden layer) | 0.968   | 0.757   | 0.887     | 0.968  | 0.913     |
| Multilayer Perceptron (22 hidden layer)          | 0.998   | 0.977   | 0.818     | 0.998  | 0.899     |
| Bayesian auto Regression                         | 1.000   | 0.922   | 0.834     | 1.000  | 0.913     |
| KNN                                              | 0.960   | 0.886   | 0.895     | 0.960  | 0.926     |
| FURIA                                            | 0.980   | 0.882   | 0.832     | 0.980  | 0.900     |
| VFDT                                             | 1.000   | 1.000   | 0.823     | 1.000  | 0.897     |
| LMT                                              | 0.984   | 0.757   | 0.869     | 0.984  | 0.911     |

The amount of improvement in the True Positive (TP) Rate, False Positive (FP) Rate, Precision, Recall, and F-Measure are as Table 5.

As can be seen in Table 5, the proposed method has improved the classification metrics in all cases examined. This improvement is more pronounced in some methods and less in others, but in all of them, the proposed method offers acceptable improvement that indicating the classification error is reduced. Therefore, the proposed method can be useful for increasing classification accuracy.
Table 5
Shows the amount of improvement in the True Positive (TP) Rate, False Positive (FP) Rate, Precision, Recall, and F-Measure (percentage)

| Data analysis model                                      | TP Rate | FP Rate | Precision | Recall | F-Measure |
|----------------------------------------------------------|---------|---------|-----------|--------|-----------|
| Deep Learning                                            | 0.2     | 0.2     | 0.1       | 0.2    | 0.2       |
| Multilayer Perceptron classifier (one hidden layer)      | 2.98    | 1.2     | 0.9       | 2.98   | 0.5       |
| Multilayer Perceptron (22 hidden layer)                  | 0.6     | 0.3     | 0.4       | 0.6    | 0.4       |
| Bayesian auto Regression                                 | 0.4     | 2.9     | 2.3       | 0.4    | 1         |
| KNN                                                      | 5.15    | 8.4     | 8.45      | 9.5    | 6.6       |
| FURIA                                                    | 2.5     | 0.5     | 1         | 2.5    | 0.6       |
| VFDT                                                     | 0       | 0       | 2.6       | 0      | 0.2       |
| LMT                                                      | 2.93    | 8.76    | 1.52      | 2.93   | 0.9       |

Figures 3 to 10 shows the number of classification error of each model separately which Figs a is in no attribute reduction and Figs b is with attribute reduction conditions.

For example, in Fig. 3a, it can be seen that the number of classification errors is 6 whereas number of classification errors in Fig. 3.b is 5, and also for other Figs. So, this indicates that the proposed method reduces the number of classification errors.

8. Conclusion

In this paper, the analysis on the Covid-19 data set was performed in three steps. The first stage includes a selection of a classification model with no reduction attributes. Tested classification algorithms include Deep learning, Multilayer Perceptron, KNN, Bayesian Auto Regression, Logistic Model Trees (LMT), Hoeffding tree (VFDT), and Fuzzy Unordered Rule Induction Algorithm. In the second stage, a rough set reduction algorithm based on genetic algorithms is employed and finally, an optimization of the classification is conducted using the reduced attributes. The proposed model is evaluated on a global COVID-19 dataset. Experimental results demonstrate that the rough set theory based on genetic analysis has a great contribution in increasing the diagnostic accuracy of the COVID-19 disease behaviour.

Declarations

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.
Conflict of interest: All authors declare that they have no conflict of interest.

Funding details: No funds have been used.

Author contributions: All authors participated equally in the preparation of the article.

References

[1] World Health Organization. Coronavirus Disease 2019 (COVID-19) Situation Report – 28.Who.int.2020. Available from: https://www.who.int/docs/defaultsource/coronaviruse/situation-reports/20200217-sitrep-28-covid-19.pdf. [Accessed 18 February 2020].

[2] A. S. Albahri, Rula A. Hamid, Jwan k. Alwan, Z.T. Al-qays, A. A. Zaidan, B. B. Zaidan, A O. S. Albahri, A. H. AlAmoodi, Jamal Mawlood Khlaf, E. M. Almahdi, Eman Thabet, Suha M. Hadi, K I. Mohammed, M. A. Alsalem, Jameel R. Al-Obaidi, H.T. Madhloom, Role of biological Data Mining and Machine Learning Techniques in Detecting and Diagnosing the Novel Coronavirus (COVID-19): A Systematic Review, J Med Syst. 2020; 44(7): 122. Published online 2020 May 25. doi: 10.1007/s10916-020-01582-x

[3] Akhtar, M., Kraemer, M., and Gardner, L.: A Dynamic Neural Network Model for Predicting Risk of Zika in Real Time. BMC Medicine, 17(171) (2019). https://doi.org/10.1186/s12916-019-1389-3.

[4] L. Huang, R. Han, T. Ai, P. Yu, H. Kang, Q. Tao, et al., Serial quantitative chest CT assessment of COVID-19: Deep-Learning Approach, Radiology: Cardiothoracic Imaging, vol. 2, p. e200075, 2020.

[5] Xi. Yang, Xu. He, J. Zhao, Y. Zhang , Sh. Zhang, et al., COVID-CT-Dataset: a CT scan dataset about COVID-19, arXiv preprint arXiv:2003.13865, 2020.

[6] L. Zhang, Xu. Kong, Xi. Li, P. Yu, H. Kang, Q. Tao, et al., CT imaging features of 34 patients infected with COVID-19, Radiology: Cardiothoracic Imaging, vol. 2, p. e200075, 2020.

[7] Ying Song, Shuangjia Zheng, Liang Li, Xiang Zhang, Xiaodong Zhang, Ziwang Huang, Jianwen Chen, Huiying Zhao, Yusheng Jie, Ruixuan Wang, et al. Deep learning enables accurate diagnosis of novel coronavirus (covid-19) with ct images. medRxiv, 2020.

[8] Shuai Wang, Bo Kang, Jinlu Ma, Xianjun Zeng, Mingming Xiao, Jia Guo, Mengjiao Cai, Jingyi Yang, Yaodong Li, Xiangfei Meng, et al. A deep learning algorithm using ct images to screen for corona virus disease (covid-19). medRxiv, 2020.

[9] Feng Shi, Liming Xia, Fei Shan, Dijia Wu, Ying Wei, Huan Yuan, Huiting Jiang, Yaozong Gao, He Sui, and Dinggang Shen. Large-scale screening of covid-19 from community acquired pneumonia using infection size-aware classification. arXiv preprint arXiv:2003.09860, 2020.

[10] A. Narin, C. Kaya, and Z. Pamuk, "Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks," arXiv:2003.10849, 2020.
[11] X. Xu, X. Jiang, C. Ma, P. Du, X. Li, S. Lv, et al., "Deep learning system to screen Coronavirus disease 2019 pneumonia," arXiv:2002.09334, 2020.

[12] E. Hosseini, K. Z. Ghafoor, A. S. Sadiq, M. Guizani and A. Emrouznejad, "COVID-19 Optimizer Algorithm, Modeling and Controlling of Coronavirus Distribution Process," in IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 10, pp. 2765-2775, Oct. 2020, doi: 10.1109/JBHI.2020.3012487.

[13] S. Wang et al., "A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19)," medRxiv, 2020.

[14] Fang Y, Zhang H, Xie J, Lin M, Ying L, Pang P, et al. Sensitivity of chest CT for COVID-19: comparison to RT-PCR. Radiology 2020:200432.

[15] Ai T, Yang Z, Hou H, Zhan C, Chen C, Lv W, Tao Q, Sun Z, Xia L. Correlation of Chest CT and RT-PCR Testing in Coronavirus Disease 2019 (COVID-19) in China: A Report of 1014 Cases. Radiology (2020).

[16] M.Y. Ng, E.Y.P. Lee, J. Yang, F. Yang, X. Li, H. Wang, et al. Imaging profile of the COVID-19 infection: radiologic findings and literature review. Radiol Cardiothorac Imaging, 2020.

[17] Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 395 (10223):497e506, 2020.

[18] Huang P, Liu T, Huang L, Liu H, Lei M, Xu W, et al. Use of chest CT in combination with negative RT-PCR assay for the 2019 novel coronavirus but high clinical suspicion. Radiology 295(1):22e3, 2020.

[19] Chung M, Bernheim A, Mei X, Zhang N, Huang M, Zeng X, et al. CT imaging features of 2019 novel coronavirus (2019-nCoV). Radiology 295(1):202e7,2020.

[20] Bernheim A, Mei X, Huang M, Yang Y, Fayad ZA, Zhang N, et al. ChestCT findings in coronavirus disease-19 (COVID-19): relationship to duration of infection. Radiology 200463, 2020.

[21] E. Frank, M. A. Hall, I. H. Witten, The WEKA workbench, Online Appendix for “Data Mining: Practical Machine Learning Tools and Techniques”, Morgan Kaufmann, Fourth Edition (2016).

[22] Bishop CM. Neural Networks for Pattern Recognition. Oxford. New York, NY, 1995.

[23] Ricco G. Bayesian Vector Autoregressions, with Oxford Research Encyclopedia, Oxford University Press, 2018.

[24] Hühn, J., Hüllemeyer, E.: Furia: an algorithm for unordered fuzzy rule induction. Data Mining and Knowledge Discovery 19(3), 293–319, 2009.

[25] Geoff Hulten, Laurie Spencer, Pedro Domingos: Mining time-changing data streams. In: ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, 97-106, 2001.
[26] Niels Landwehr, Mark Hall, and Eibe Frank. Logistic model trees (PDF). ECML PKDD, 2003.

[27] Altman, Naomi S. "An introduction to kernel and nearest-neighbour nonparametric regression" (PDF). The American Statistician. 46 (3): 175–185, 1992.

[28] Ziarko, W."The Discovery, Analysis and Representation of Data Dependencies in Databases", Knowledge Discovery in Databases, AAAI MIT Press, Cambridge, MA, pp.213-228, 1993.

[29] Mirzae E., Esmaeilpour M. A New Hybrid Method to Increase the Prediction in Data Reduced Using Rough Set and Swarm Intelligence Model. JSDP. 14 (3):51-64, 2017.

[30] Chouchoulas and Q. Shen. Rough set-aided keyword reduction for text categorisation. Applied Artificial Intelligence,15(9):843-873, 2001.

[31] Grefenstette, J. J. (1994). Genetic algorithms for machine learning. Boston, MA: Kluwer.

[32] Ren, Y.-G., Wang, Y.: Rough Set Attribute Reduction Algorithm Based on GA. Mini-Micro Systems 27(5), 862–865 (2005)

[33] https://github.com/ieee8023/covid-chestxray-dataset. Available 18. June 2020.

**Figures**
The MSD-Covid19 Algorithm

Input: D, Dataset and A, The attributes of dataset and M, Selected methods of data mining;
Output: AR, Minimal attributes set as reduction subset;

Stage1:
1. For i=1 to M:
2. Run Mi on the D(A).
   End For

Stage2:
Initialization:
3. Let A be a set of all attributes and let (a₁ ... aₙ) be an ordered list of attributes and the fitness function method is FN(x) = K. (Eq.6)
   //The new values FN(x) uses as a probability distribution, and the new population randomly chooses, using "Roulette Wheel" distribution.

Start
4. For i=1 to A:
5. AR := A - aₖ
6. If FN(AR) does maximized Return AR.
   End For

Stage3:
7. For i=1 to M:
8. Run Mi on the D(AR).
   End For

End

Figure 1

The Multi-stage-Covid19 Diagnosis Algorithm
Figure 2

COVID-19 public dataset collection [33]
Figure 3

The classification error of the Deep Learning model
Figure 4

Shows the classification error of the Multilayer Perceptron model
Figure 5

Shows the classification error of the Multilayer Perceptron model (22 hidden layer)
Figure 6

Shows the classification error of the Bayesian auto Regression model.
Figure 7

Shows the classification error of the KNN model before
Figure 8

Shows the classification error of the FURIA model
Figure 9

Shows the classification error of the VFDT model
Figure 10

Shows the classification error of the LMT model