Data Mining Approach to Analyze COVID-19 Dataset of Mexican Patients

Waheeda Almayyan
Computer Information Department
College of Business Studies, PAAET, Kuwait

ABSTRACT

The pandemic originated by coronavirus (COVID-19), force governments to choosing different health policies to stop the infection and inspire many research groups to work on patient's data to understand the virus behaviour. This research suggests a two-phase prediction system with several learning algorithms to explore the COVID-19 dataset, where Chi-square is employed at the first stage. Cuckoo search and Grey Wolf Optimiser approaches have been proposed in the second stage to inherit their advantages to select the most distinctive features. The proposed classification model is trained and tested with six machine learning algorithms. The proposed model resulted in 96.5% of Accuracy with samples of 95839 patients with several incomplete data.

Keywords

Data Mining; Chi-square feature selection; Grey Wolf Optimiser; Cuckoo search; COVID-19

1. INTRODUCTION

In January 2020, The World Health Organization (WHO) declared a state of emergency to address a virus outbreak in Wuhan, Hubei province in China [1]. Two months after this announcement, the organization started to alert that the world is facing a pandemic. In Mexico, the first cases of COVID-19 were reported by the Secretary of Health on February 28, 2020 [2]. More than 720 000 positive cases and more than 76 000 deaths have been reported in Mexico up to September 2020, and it continues to rise [3].

Applying Data mining techniques to discover relationships and usage patterns within data allows the acquisition of meaningful information from large-scale datasets [4]. Data mining techniques are successfully applied in many economic, industrial, scientific, and medical fields to handle massive datasets [5]. Therefore, data reduction techniques are mostly required for filtering, ranking priorities, and providing means to detect and isolate redundant features. These algorithms increase the quality of analyses and the success of recognition systems [4]. Medical Data Mining involves using algorithms and techniques to automate disease diagnosis and prediction. Numerous algorithms have been suggested in medical diagnosis literature and investigated on several benchmark datasets for cancer, heart disorders, and diabetes [6].

The growth in collecting medical data presents a new opportunity for physicians to improve patient diagnosis. In recent years, practitioners have increased their usage of computer technologies to improve decision-making support. In the health care industry, machine learning is becoming an essential solution to aid patients' diagnosis.

Machine learning is an analytical tool used when a task is large and complicated to program, transforming medical records knowledge, pandemic predictions, and genomic data analysis [7]. Data mining techniques may provide valuable input in this regard, particularly in making diagnoses based on clinical text, radiography Images, etc. According to Bullock et al. [8], Machine learning and deep learning can replace humans by giving an accurate diagnosis. The perfect diagnosis can save radiologists' time and can be cost-effective than standard tests for COVID-19.

The paper is organized as follows. Sections 2 and 3 describes the related work and research methodology. Section 4 presents experimental results and discussion, followed by the conclusion in Section 5.

2. RELATED WORK

Regarding COVID-19 prediction, from the literature, we can mention the following work. X-rays and computed tomography (CT) scans can be used for training the machine learning model. Several initiatives are underway in this regard. Wang and Wong [9] developed a deep convolutional neural network, which can diagnose COVID-19 from chest radiography images. They applied it over an open dataset comprising 13,975 CXR images across 13,870 patient cases. In Roda et al. [10], the authors discussed in detail why it is difficult to predict the COVID-19 epidemic accurately. In Roosa et al. [11], the authors described real-time forecasts of the COVID-19 epidemic in China from February 5, 2020, to February 24, 2020, with good results. In Ton et al. [14], the authors describe the rapid identification of potential inhibitors of SARS-CoV-2 main protease by deep model docking of 1.3 billion compounds. In Wang et al. [12], the authors describe a novel phase-adjusted estimation approach of the number of Coronavirus Disease cases in Wuhan, China.

In [13], Narin et al. employed five pre-trained convolutional neural network-based models (ResNet50, ResNet101, ResNet152, InceptionV3 and Inception-ResNetV2) for the detection of coronavirus pneumonia infected chest X-ray radiographs. by using 5-fold cross-validation. Results obtained showed that the pre-trained ResNet50 model provided the highest classification performance compared to other models. Barstugan et al. used several patches from CT images of COVID-19 patients to diagnose the virus in the early phase [14]. Five feature extraction methods have been used to find a feature set that separated contaminated stains with high Accuracy. This model classifier's best Accuracy was 99% with 10-fold cross-validation and Grey-Level Size Zone Matrix feature extraction method. Wiguna and Riana (2020) provided a classification model using the C4.5 algorithm that classifies three categories (supervised patients, suspected, and asymptomatic individuals and achieved an accuracy of 92.8% [15].

Muhammad et al. predicted the recovery of COVID-19 patients of the epidemiologic dataset of South Korean patients and algorithms such as support vector machine, naive Bayes, logistic regression, random forest, and K-nearest neighbor python language. Results showed that this model could predict patients' recovery with a possibility of 99.8% Accuracy [16]. Khanday et al. provided a model for textual clinical reports for detecting COVID-19 using basic and hybrid algorithms [8]. These reports have been categorized into four classes. Various features such as Term Frequency/ Inverse Document Frequency and Bags of Words have been extracted from these reports. Finally, the Bayesian classifier gives excellent results by having 94% precision, 96% recall, 95% f1-score, and Accuracy of 96.2%.

Yan et al. [17] suggested an XGBoost machine learning tool with three biomarkers to predict the survival of individual patients with more than 90% Accuracy: Lactic Dehydrogenase (LDH), lymphocyte, and high-sensitivity C-reactive protein (hs-CRP). [18] proposed a machine learning model that can predict the cases of COVID-19. The proposed model resulted in 70-80% of Accuracy with samples of 53 patients with some incomplete data that is restricted to two hospitals in Wenzhou, Zhejiang, China. The most predictive features were alanine aminotransferase (ALT), myalgias, and hemoglobin, in this order using Support Vector Machine and k-NN classifiers.

Not all COVID-19 positive patients will need rigorous attention. Being able to prognosis which will be affected more severely can help in directing assistance and planning medical resource allocation and utilization. Despite the amount of work as mentioned earlier, no prior work has conducted a systematic study of the impact of feature reduction techniques on the possibility of predicting the confirmed cases of COVID-19. We believe that our model will fill a current gap in existing research, which lacks studying the effect of feature selection on Coronavirus data.

This paper aims to identify the critical features required to construct the Covid-19 detection model, thereby achieving maximum performance. In this paper, Chi-square feature selection and a meta-heuristic algorithm are utilized to develop a two-phase classification model. The motivation for selecting Chi-square feature selection is that they rank the features based on the statistical significance test and consider only those dependent on the class label. Because of its advantages, we decided to use meta-heuristic algorithms to select the most distinctive features.

The classification learning models combined with dimensionality reduction seek to achieve three primary objectives: (i) to learn the best feature representation of the dataset used; (ii) to validate the performance of several meta-heuristic algorithms in conjunction with a feature selection technique; and (iii) to learn the classification model that computes the best performance. Six classifiers will validate the model: random forests (RandF), k-nearest neighbor (k-NN), Iterative Classifier Optimizer (ICO), JRip; PART, and Fisher's linear discriminant analysis (FLDA) classifiers.

3. MATERIALS AND METHODS

The proposed approach is applied to the raw and pre-processed datasets. We assessed the original datasets first with all six classifiers. In the second experiment, we applied the chi-square feature selection technique to obtain a unique and reduced set of ranking-based features and validate them with the classifiers. The third set of experiments used the reduced datasets obtained by chi-square and then applied GWO and CSA algorithms.

To compare the classification with PCA, the fourth set of experiments applied PCA directly from raw data. The final experiment starts with the reduced datasets obtained by chi-square and then applied PCA. The validation and analysis module used the performance metrics mentioned in Section 2.6, such as Accuracy, specificity, balanced Accuracy, and precision. The representation of this approach is illustrated in Figure 1.

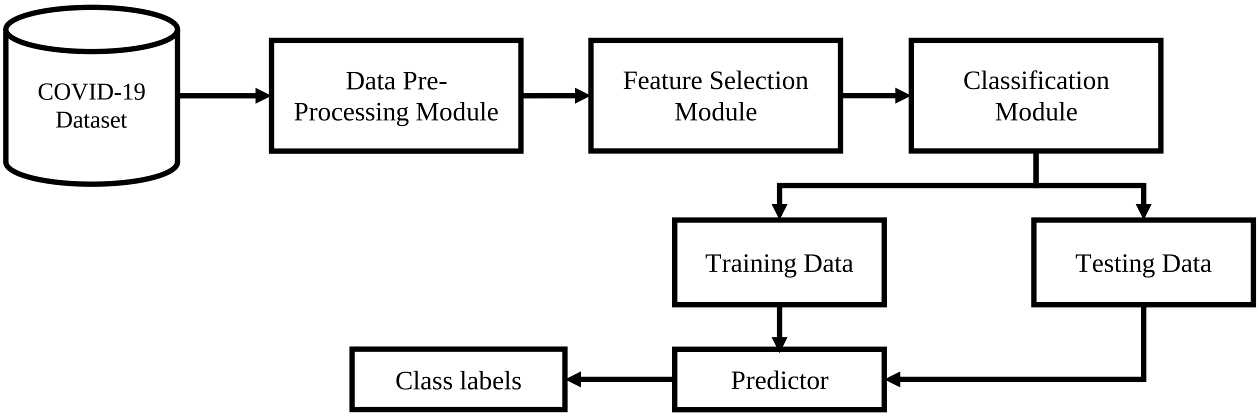

Figure 1. Proposed System

3.1 Description of the dataset

The dataset used in the research was the COVID-19 Patient Health open research dataset obtained from the Kaggle repository [19]. It was gathered individually from each medical unit in Mexico and standardized by federal government guidelines. This dataset consists of 95839 clinical information of patients with suspected or confirmed COVID-19 between January 15, 2020, and May 3, 2020. The dataset contains 19 features and one diagnostic class, as shown in Table 1; 16 of them contain missing feature values. The employed dataset in this research included 48720 Males and 47119 Females.

The dataset includes demographic features, clinical information, and essential medical conditions of suspected and confirmed COVID-19. Including gender, the type of care the Patient received in the unit, intubated state, pneumonia state, age, pregnancy, Diabetes state, Chronic obstructive pulmonary disease (COPD) stay, Asthma status, Immunosuppression status, Hypertension status, Cardiovascular status, Obesity status, Chronic kidney failure status, smoking history, Intensive care unit status, did the Patient had contact with another case diagnosed with SARS CoV-2, and death date if possible. Several features were not considered for the analysis, such as Nationality, Country of origin, Migrant, Entry date, Municipality, Medical Unit Id, National entity, Local entity, and Date of symptoms.

Table 1. The COVID-19 Mexico patient health dataset features details

No	Feature Name	Possible value and Description	Missing Values
1.	Gender	1=Women 2=Man	No Missing Values
2.	Patient_type	Type of care the patient received in the unit. 1= Outpatient 2= Hospitalized	No Missing Values
3.	Intubated	Patient required intubation 1 = YES 2 = NO 98/97 = NOT APPLICABLE 99 = NOT AVAILABLE	Number of Missing Values
4.	Pneumonia	Patient was diagnosed with pneumonia 1 = YES 2 = NO 98/97 = NOT APPLICABLE 99 = NOT AVAILABLE	Number of Missing Values
5.	Age	98/97 = NOT APPLICABLE 99 = NOT AVAILABLE	Number of Missing Values
6.	Pregnant	1 = YES 2 = NO 98/97 = NOT APPLICABLE	Number of Missing Values
7.	Diabetes	1 = YES 2 = NO 98/97 = NOT APPLICABLE	Number of Missing Values
8.	Copd	1 = YES 2 = NO 98/97 = NOT APPLICABLE	Number of Missing Values
9.	Asthma	1 = YES 2 = NO 98/97 = NOT APPLICABLE	Number of Missing Values
10.	Immunosuppression	1 = YES 2 = NO 98/97 = NOT APPLICABLE	Number of Missing Values
11.	Hypertension	1 = YES 2 = NO 98/97 = NOT APPLICABLE	Number of Missing Values
12.	Other_diseases	Another comorbidity 1 = YES 2 = NO 98/97 = NOT APPLICABLE	Number of Missing Values
13.	Cardiovascular	1 = YES 2 = NO 98/97 = NOT APPLICABLE	Number of Missing Values
14.	Obesity	1 = YES 2 = NO 98/97 = NOT APPLICABLE	Number of Missing Values
15.	Chronic_kidney_failure	1 = YES 2 = NO 98/97 = NOT APPLICABLE	Number of Missing Values
16.	Smoker	1 = YES 2 = NO 98/97 = NOT APPLICABLE	Number of Missing Values
17.	Another_case	Did the Patient had contact with another case diagnosed with SARS CoV-2 1 = YES 2 = NO 98/97 = NOT APPLICABLE	Number of Missing Values
18.	Outcome	1 = COVID-19 POSITIVE 2 = COVID-19 NEGATIVE	No Missing Values

		3 = Uncertain	
19.	ICU	Did the patient required admission to an Intensive Care Unit. 1 = YES 2 = NO 98/97 = NOT APPLICABLE 99 = NOT AVAILABLE	Number of Missing Values
20.	Death_date (class)	From 15-01-2020 to 03-05-2020 9999-99-99 = Live	No Missing Values

3.2 Dimensionality reduction

In most machine learning applications, the dataset's high dimensionality is considered a significant problem [20]. It obliges involving a large amount of memory, and it may lead to overfitting. So, improving the classification algorithm's performance usually starts with removing redundant data and reducing insignificant features [21]. Characterizing diseases, genome expression, medical images can help physicians to improve patient diagnosis. Dimensionality reduction [22] is the process of reducing the number of variables considered. It can extract latent features from raw datasets or reduce the data while maintaining the structure.

3.2.1. Chi-square

Chi-squared is a standard statistical technique that estimated deviation from the expected distribution if the feature incidence is not dependent on the class value [23]. The Chi-square feature selection process will contribute to giving a new dataset. The value of the Chi-square metric is calculated as [24]

Chi-square = t(tp,(tp+fp)Ppos) + t(fn,(fn+tn)Ppos) +t(fp,(tp+fp)Pneg)+t(tn,(fn+tn)Pneg) 1

In Equation 1, tp=true positives, fp=false positives, tn=true negatives, fn=false negatives, Ppos=probability of number of positive cases, Pneg=probability of number of negative cases and t(count, expect) = (count − expect)2/expect. The chi-square approach follows the next steps to deduct results.

1. Develop an analysis plan to specify the Significance rank and Test method through choosing significance level any value between 0 and 1 and then applying the chi-square test to test independence level to identify whether there is a significant relationship between two categorical attributes.

2. The sample data have to be analyzed to calculate the degrees of freedom, predictable frequencies, test value, and the P-value associated with the test. The degrees of freedom is computed as:

Degrees of freedom: DF= (r-1)*(c-1) 2

where r is the number of levels of one categorical variable and c is the number of levels for other categorical variable. Then, the Test Statistic is computed as

Test Statistic $: x^2(f,c) = \left[\frac{N*(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)}\right]$ 3

Where A = No. of times feature 't' and class label 'c' co-occurs.
B = No. of times 't' appears without 'c'.
C = No. of times 'c' appears without 't'.
D = No. of times neither 'c' nor 't' appears.
N = Total number of records.

3.2.2. Grey Wolf Optimizer

The GWO algorithm is a metaheuristic algorithm that was introduced by Mirjalili in 2014 [25]. The GWO algorithm is primarily motivated by the leadership and hunting behaviors of grey wolves in nature, mostly in packs consisting of 5-12 with a strict hierarchy, and each wolf in the group has a defined role in the hunting process [26]. Each herd has four leading ranks that are modeled as a pyramid [30]. The first level in the hierarchy of grey wolves is the pack leader called alpha (α). This level is responsible for decisions about all wolf behaviors. Beta (β) wolves are the second level and are responsible for helping in the packs' choices. The third level is the delta (δ) wolves and are responsible for watching the boundaries of the territory, protecting the group, hunting, etc. The last level in the hierarchal model is called omega (ω), which is the weakest of the wolves, as it should obey other individuals' orders [25].

The second inspiration of GWO is that wolves are known for their group-hunting strategy to catch prey. The hunting process starts with tracking and chasing the prey, then encircling and harassing until it finally stops moving before attacking in a manner that can be mathematically modeled as

$\vec{D} = |\vec{C}.\overrightarrow{X_p} - \vec{X}(t)|$, 4

$\vec{X}(t+1) = |\overrightarrow{X_p}(t) - \vec{A}.\vec{D}|$

where $\vec{X}$ and $\overrightarrow{X_p}$ represent the prey and grey wolf's positions at an iteration (t), respectively. $\vec{A}$ and $\vec{C}$ are coefficient vectors that can be formulated as follows:

$\vec{A} = |2a.\overrightarrow{rand_1} - \vec{a}|$, 5

$\vec{C} = 2.\overrightarrow{rand_2}$

where $\overrightarrow{rand_1}$and $\overrightarrow{rand_2}$are random vectors in [0,1]. $\vec{a}$ is linearly decreasing from 2 to zero over iterations as follows:

$a = 2 - t * \frac{2}{iterations}$ 6

The grey wolf (X,Y) can be updated according to the prey's position (X∗,Y∗). The vectors $\vec{A}$ and $\vec{C}$ are used to update the position of the best grey wolf. As the alpha, beta, and delta indicate the best three solutions, the rest of the wolves update their positions according to the best three solutions ($\overrightarrow{X_1}$, $\overrightarrow{X_2}$, and $\overrightarrow{X_3}$). It can be expressed as follows:

$\vec{X}(t+1) = (\overrightarrow{X_1} + \overrightarrow{X_2} + \overrightarrow{X_3})/3$, 7

$\overrightarrow{X_1} = \overrightarrow{X_\propto} - \overrightarrow{A_1}.(\overrightarrow{D_\propto}).\overrightarrow{D_\propto} = |\vec{C}.\overrightarrow{X_\propto} - \vec{X}|$,

$\overrightarrow{X_2} = \overrightarrow{X_\beta} - \overrightarrow{A_2}.(\overrightarrow{D_\beta}).\overrightarrow{D_\beta} = |\overrightarrow{C_2}.\overrightarrow{X_\beta} - \vec{X}|$,

$\overrightarrow{X_3} = \overrightarrow{X_\delta} - \overrightarrow{A_3}.(\overrightarrow{D_\delta}).\overrightarrow{D_\delta} = |\overrightarrow{C_3}.\overrightarrow{X_\delta} - \vec{X}|$

3.2.3. Cuckoo search

The Cuckoo search algorithm (CSA) is basically derived from the strange reproductive behavior of particular cuckoo species. These species choose to put eggs in randomly chosen nests of other host birds but have similar matching patterns of the hosts' eggs to reduce their probability of discovering them [28]. The cuckoos rely on these host birds to accommodate their eggs. Sometimes, when the host birds recognize unfamiliar eggs, it usually rejects it or abandons their nests. According to the CSA, each egg in the nest represents a possible solution, and the foreign cuckoo egg represents a new solution. The goal is to employ potentially better solutions (cuckoos) to replace the nests' solution [29]. CSA generates a new candidate solution (nest) xi(r+1) for a cuckoo n [30]

$$xi(r+1) = xi(r) + \alpha \otimes \text{Lévy}(\lambda) \quad 8$$

where s is the step size, and α >0 is the step size scaling, related to interest. In most cases, α is set to 1. The symbol ⨁ is an entry-wise multiplication that is similar to those used in the PSO algorithm [29].

The CSA is based on Lévy flights to avoid a local optimum [28]. The concept of Lévy flights explores the solution space (s) by providing a random walk with random steps drawn from a Lévy distribution for large steps, given by:

$$\text{Lévy}(\lambda) \sim u = s\text{-}\lambda,(1<\lambda\leq 3) \quad 9$$

3.2.4. Selected classification models

For this research, the classification models use the default value for most of the hyperparameters. The models were: (1) random forests (RandF); (2) k-nearest neighbor (k-NN); (3) Iterative Classifier Optimizer (ICO); (4) JRip; (5) PART and (6) Fisher's linear discriminant analysis (FLDA) classifiers. Table 2 describes the parameter settings for each classifier.

The reasoning behind these particular choices was to provide a realistic set of results and show the learners' different characteristics. The Random Forest method (RandF) received increased attention within several classification problems [31,32]. RandF is an ensemble machine learning technique that was developed by Breiman [33]. Random forest is an ensemble learning method for classification works by constructing a collection (“forest”) of (random) decision trees at training time and returning the class that is the mode of all of the classes of the individual trees. RandF classifiers attempt to mitigate the tendency of decision trees to overfit the training data set.

The k-NN is an easy algorithm to understand and implement and a powerful tool because it does not assume anything about the data other than a distance measure can be calculated consistently between two instances. k-NN is a type of instance-based learning or lazy learning where the function is only approximated locally, and all computation is deferred until classification. It is a non-parametric method used for classification or regression [34, 35]. Iterative classifier optimizer (ICO) uses cross-validation and optimizes the number of iterations for the given classifier; it is capable of handling missing, nominal, binary classes and attributes like numeric, nominal, binary, empty nominal [36].

JRip is a rule-based algorithm that learns propositional rules by repeatedly growing rules and pruning them. During the growth phase, antecedents are added greedily until a termination condition is satisfied. Antecedents are then pruned in the next phase, subject to a pruning metric. Once the ruleset is generated, further optimization is performed where rules are evaluated and deleted based on their performance on randomized data [37]. FLDA has no requirement on the distribution of data sets, which attracts the interest of many scholars. Hence, FLDA has developed numerous variations for different purposes since its first publication 80 years ago [38]. PART is a partial decision tree algorithm, the developed version of C4. 5 and RIPPER algorithms. The PART algorithm's main specialty is that it does not need to perform global optimization like C4. 5 and RIPPER to produce the appropriate rules [39].

Table 2 Parameters tuning for classifier in Weka

Classifier	Basic Parameters
RandF	maxDepth=0; numExecutionSlots=1; numFeatures=0; seed=1
k-NN	KNN=1; windowSiz= 0; nearestNeighbourSearchAlgorithm = “LinearNNSearch”
ICO	lookAheadIterations=50; classValueIndex=-1; evaluationMetric=“RMSE”
JRip	Folds=3; minNo=2.0; optimizations=2
PART	confidenceFactor =0.25; numFolds=3; minNumObj=2
FLDA	Ridge=1.0E-6

2.2.5. Evaluation process

In this article, training and evaluation of models were performed using 10-fold cross-validation, in which nine folds of data are used to train the classifier, and one-fold is used for testing the classifier. This is performed ten times, and accordingly, testing and training encompass all ten folds to prevent a model being overfitted to the dataset. When learning medical data, Accuracy and the error rate usually favor the majority class [40]. To evaluate our approach's feasibility in the medical domain, we applied the 10-fold cross-validation procedure through all the experiments. The Accuracy of the selected classifiers was calculated using the Accuracy, specificity, balanced Accuracy, and precision, which are more appropriate measures for imbalanced datasets [41]. According to the confusion matrix of a two-class problem, the main formulations are defined in Equations 10-13 according to the confusion matrix of a two-class problem.

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+FP+TN} \quad 10$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad 11$$

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{TP}{TP+NP} + \frac{TN}{TN+FP}\right) \quad 12$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad 13$$

4. EXPERIMENTAL RESULTS

We propose a model based on a two-phase prediction system with several learning algorithms to explore the Covid-19 dataset, where Chi-square is employed at the first stage. CSA and GWO approaches have been proposed in the second stage to inherit their advantages to select the most distinctive features. The proposed approach was applied to the raw data first with all the missing values and the treated dataset. We decided to treat missing attribute values as particular values. Rather than finding some known attribute value as its value, we treat the missing value as a new value for the attributes containing missing values and treat it in the same way as other values [42]. This approach assumes that we handle these values as they do not influence future analyses.

The first set of experiments will assess the raw and pre-processed dataset with all six classifiers. Afterward, we performed several types of experiments for analysis where we applied the feature selection techniques of Chi-square to obtain a unique and reduced set of ranking-based features with the diagnosis of the Covid-19 and validate them with the classifiers. The last test used the reduced datasets obtained by Chi-square and then applied a set of experiments on several metaheuristic algorithms. For comparing the classification results with PCA, we conclude with experiments that apply PCA directly from raw data and the reduced datasets obtained by Chi-square.

Our procedure starts with first training the model using the unprocessed dataset. Once it is built, the next step is to use it for testing the suggested model. Table 3 shows the first set of experimental results of all classifiers with the full set of features. It presents the classification performance of random forests, k-NN, ICO, JRip, PART, and FLDA classifiers over the selected dataset. It shows that the ICO classifier achieved superior results in all terms as it achieved 96.5%, 99.7%, 99.7%, and 55.4% in Accuracy, Specificity, Balance Accuracy, and Precision, respectively. Followed by the JRip, which showed the second highest rates of Accuracy, then the others. The dataset with the missed values assured the leading of the ICO in all the evaluation metrics. It recorded 96.5%, 99.7%, 99.7%, and 54.8% in Accuracy, Specificity, Balance Accuracy, and Precision. ICO classifier presented better results using raw datasets.

Table 3. Full dataset-Comparative study of the unprocessed dataset

Raw Data (without missing)						
	RF	kNN	ICO	JRip	PART	FLDA
Accuracy	0.958	0.954	**0.965**	0.964	0.961	0.812
Specificity	0.987	0.984	**0.997**	0.995	0.991	0.809
B.Acc	0.993	0.991	**0.997**	**0.997**	0.995	0.904
Precision	0.340	0.270	**0.554**	0.513	0.382	0.145
Raw Data (with missing)						
	RF	kNN	ICO	JRip	PART	FLDA
Accuracy	0.959	0.954	**0.965**	0.964	0.963	0.885
Specificity	0.989	0.983	**0.997**	0.996	0.994	0.890
B.Acc	0.994	0.991	**0.997**	**0.997**	0.996	0.945
Precision	0.321	0.278	**0.548**	0.507	0.458	0.202

In the first feature selection phase, we start applying the standard Chi-square feature selection algorithm to select 17 features among 19 features of the COVID-19 dataset. Table 4 lists the selected features according to the average rank obtained by standard Chi-square search algorithms for the original dataset and dataset with missed values. From the results, we decided to exclude the least frequent features in both datasets in the analysis, i.e., asthma and smoker features.

Table 4. Features selected by Chi-squared technique

Raw Data (without missing)		
Average Merit	**Average Rank**	**Feature**
10681.785 +- 115.458	1 +- 0	intubated
7987.026 +-74.364	2 +- 0	icu
6861.725 +-30.559	3 +- 0	patient_type
6217.721 +-56.082	4 +- 0	pneumonia
3873.114 +-47.049	5 +- 0	Age
2535.106 +-43.582	6 +- 0	outcome
1983.434 +-38.307	7 +- 0	diabetes
1465.166 +-34.277	8 +- 0	hypertension
731.578 +-13.412	9 +- 0	another_case
652.984 +-32.175	10.1 +- 0.3	chronic_kidney_failure
598.566 +-30.985	10.9 +- 0.3	copd
384.695 +-20.792	12 +- 0	cardiovascular
250.84 +- 5.945	13.6 +- 0.49	pregnant
249.152 +-14.392	13.8 +- 0.98	obesity
242.072 +- 5.717	14.6 +- 0.49	sex
183.503 +- 9.063	16 +- 0	immunosuppression
104.925 +- 8.361	17 +- 0	other_diseases
38.139 +- 4.58	18 +- 0	asthma
28.112 +- 6.86	19 +- 0	Smoker
Raw Data (Missing)		

Average Merit	Average Rank	Feature
6861.725 +-30.559	1 +- 0	patient_type
6216.965 +-56.081	2 +- 0	pneumonia
3877.145 +-52.696	3 +- 0	age
2535.106 +-43.582	4 +- 0	outcome
1956.449 +-40.046	5 +- 0	diabetes
1445.032 +-34.846	6 +- 0	hypertension
1015.759 +-29.547	7 +- 0	intubated
638.437 +-32.188	8.1 +- 0.3	chronic_kidney_failure
578.258 +-29.742	8.9 +- 0.3	copd
357.454 +-20.584	10 +- 0	cardiovascular
297.333 +-17.452	11 +- 0	icu
243.01 +-15.523	12.5 +- 0.67	obesity
242.072 +- 5.717	12.6 +- 0.49	sex
215.603 +- 4.549	13.9 +- 0.3	another_case
153.622 +-11.018	15 +- 0	immunosuppression
81.493 +- 9.202	16 +- 0	other_diseases
20.738 +- 2.372	17 +- 0	asthma
4.423 +- 0.357	18.1 +- 0.3	pregnant
0.469 +- 1.408	18.9 +- 0.3	smoker

We tested the results obtained after constructing feature sets using Chi-square in Table 5. For both datasets, results show that the highest Specificity, Accuracy, and Balanced Accuracy readings were scored with the ICO algorithm, where it scored 99.7%, 96.5%, and 99.7%, respectively, with 17 features. Worth noting, the classification Chi-square-based feature selection approach's performance succeeds in selecting a smaller number of features and achieving a similar classification performance than using all features. Overall, the ICO algorithm obtained the best results yet, JRip and PART presented better results using raw data. Compared to the raw data, Chi-square improved in the computations of k-NN.

Table 5. Performance metrics of the model after the first step of feature selection

Raw Data (without missing)						
	RF	kNN	ICO	JRip	PART	FLDA
Accuracy	0.958	0.955	**0.965**	0.964	0.961	0.811
Specificity	0.987	0.985	**0.997**	0.996	0.992	0.809
B.Acc	0.993	0.992	**0.997**	**0.997**	0.995	0.904
Precision	0.336	0.284	**0.554**	0.507	0.376	0.145
Raw Data (with missing)						
	RF	kNN	ICO	JRip	PART	FLDA
Accuracy	0.959	0.955	**0.965**	0.964	0.963	0.885
Specificity	0.989	0.984	**0.997**	0.995	0.994	0.890
B.Acc	0.993	0.991	**0.997**	0.996	0.996	0.945
Precision	0.328	0.290	**0.548**	0.504	0.446	0.202

In the second stage, we apply two meta-heuristic algorithms over the reduced set of features as they have gained popularity as tools for solving a wide array of optimization problems in many different areas of application. For this purpose, we have selected GWO and CSA algorithms.

Table 6 lists the obtained features using GWO and CSA. Worth noting that the obtained features were standard except for the "icu" feature. This step has reduced the feature size from 16 to 5 to 6 features only.

Table 6. Results of the second step of feature selection

Technique	No. of Features	Selected Features
CSA	6	patient_type intubated pneumonia diabetes outcome icu
GWO	5	patient_type intubated pneumonia diabetes outcome

Table 7 presents the classification performance of the chosen classifiers over the obtained dataset from GWO and CSA. Most of the classifiers, except the FLDA, performed exceptionally well in terms of Accuracy, Specificity, Balanced Accuracy, and Precision, where it scored an average of 96.5%, 99.7%, 99.7%, and 54.3%, respectively, with six features extracted using GWO technique with the raw dataset. CSA technique performed was promising using five features. As it scored 96.5%, 99.7%, 99.8%, and 55.1% for Accuracy, Specificity, Balanced Accuracy, and Precision, respectively. The dataset with the missed values, an average of 96.5%, 99.7%, 99.7%, and 64.3%, was recorded for Accuracy, Specificity, Balanced Accuracy, Sensitivity, and Precision, respectively, using the GWO technique. CSA and GWO search helped in improving the classification performance with a limited number of features. According to the results generated, the five common features were the patient_type, intubated, pneumonia, diabetes, and outcome. This helped in reducing dimensionality by 70%. The FLDA classifier suffers the largest Accuracy drops, while the other classifiers using both raw and pre-processed datasets maintain Accuracy above 96%.

Table 7. Performance metrics of the model after the second step of feature selection

Raw Data (without missing)							
Technique		**RF**	**kNN**	**ICO**	**JRip**	**PART**	**FLDA**
CSA	Accuracy	**0.965**	**0.965**	0.964	**0.965**	**0.965**	0.794
	Specificity	0.997	0.997	**0.998**	0.996	0.997	0.791
	B.Acc	**0.997**	**0.997**	**0.997**	**0.997**	**0.997**	0.895
	Precision	0.544	**0.551**	0.538	0.542	0.543	0.135
GWO	Accuracy	**0.965**	**0.965**	**0.965**	**0.965**	**0.965**	0.790
	Specificity	0.995	0.995	**0.998**	0.995	0.995	0.787
	B.Acc	**0.997**	**0.997**	0.996	**0.997**	**0.997**	0.893
	Precision	0.522	0.522	0.552	0.521	**0.524**	0.133
Raw Data (with missing)							
		RF	kNN	ICO	JRip	PART	FLDA
CS	Accuracy	**0.965**	**0.965**	0.964	**0.965**	**0.965**	0.897
	Specificity	**0.997**	**0.997**	0.998	0.996	**0.997**	0.906
	B.Acc	**0.997**	**0.997**	0.997	**0.997**	**0.997**	0.953
	Precision	0.550	**0.556**	0.546	0.537	0.643	0.208
GWO	Accuracy	**0.965**	**0.965**	0.965	**0.965**	**0.965**	0.886
	Specificity	0.995	0.995	0.998	**0.995**	**0.995**	0.893
	B.Acc	**0.997**	**0.997**	0.996	**0.997**	**0.997**	0.946
	Precision	0.524	0.524	0.552	0.520	**0.525**	0.192

Principal component analysis (PCA) is a widespread statistical data transformation and dimensionality reduction method that uses an orthogonal transformation method to transform correlated features into a set of linearly uncorrelated features, i.e., principal components [43]. The number of principal components will be fewer or equal to the original dataset's number of features.

In the rest of the experiments, we will compare the raw data's best results with PCA. The first set of experiments uses PCA to transform the original dataset into a new set of features, also known as principal components. As shown in the first part of Table 8, the best result was 96.4% Accuracy, 100.0% Specificity, 99.4% Balanced Accuracy and 50% Precision. The results obtained using the missing dataset scored were 96.4%, 100%, 99.7%, and 51.7% for Accuracy, Specificity, Balanced Accuracy, and Precision, respectively. Overall, the ICO algorithm obtained the best results with the original dataset yet, JRip and PART presented better results using raw data.

In the second phase, we take Chi-square as input and transform them into a new set of PCA features. Results shows that the ICO classifier achieved superior result with the original dataset in most metrics as it achieved 99.6%, 100% and 99.6% in Accuracy, Specificity, and Precision respectively. Followed by the JRip, which showed the highest rate of Balanced Accuracy, then the others. The dataset with the missed values assured the leading of the ICO in all the evaluation metrics. It recorded 99.6%, 100%, and 99.6% in Accuracy, Specificity, and Precision. ICO classifier presented better results using raw datasets. From the comparison results, it has been observed that the ICO classifier performs much better than others in most cases.

Results showed that the second phase of feature selection using CSA and GWO techniques shows competitive results by selecting six and five features, respectively. The PCA algorithm shows worse results with a maximum number of 11 features from 19 features. At the same time, the suggested technique shows promising results with a smaller number of features.

The suggested technique in the second phase exhibits better classifier results compared to other methods. It is also discovered that Chi-sequare+CSA and Chi-sequare+GWO achieve maximum classification accuracy with an average value of 96.5%. The Chi-sequare+CSA and Chi-sequare+GWO exhibit almost the same performance level, as they achieved 99.7%, 99.7%, and 54.2% in Specificity, Balanced Accuracy, and Precision, respectively. Our results suggest that the imputation methods evaluated have a minor impact on the classification analyses. In this matter, we agree with researchers in [44] in saying that simple methods such as replacing the missing values by mean or the median values performed as well as complex strategies.

Table 8. Performance metrics of the model with PCA

Full dataset with PCA						
Raw Data (without missing)						
	RF	kNN	ICO	JRip	PART	FLDA
Accuracy	0.959	0.952	**0.964**	**0.964**	**0.964**	0.808
Specificity	0.990	0.980	**1.000**	0.999	**1.000**	0.806
B.Acc	**0.994**	0.989	0.956	0.993	0.946	0.903
Precision	0.325	0.251	**0.500**	0.441	0.400	0.144
Raw Data (with missing)						
	RF	kNN	ICO	JRip	PART	FLDA
Accuracy	0.958	0.951	0.964	0.964	0.964	0.877
Specificity	0.988	0.979	1.000	0.996	0.997	0.882
B.Acc	0.993	0.989	0.968	0.997	0.997	0.941
Precision	0.326	0.260	0.311	0.517	0.463	0.191
Full dataset with Chi-square then PCA						
Raw Data (without missing)						
	RF	kNN	ICO	JRip	PART	FLDA
Accuracy	0.957	0.952	0.996	0.964	0.964	0.808
Specificity	0.988	0.981	1.000	0.999	1.000	0.806
B.Acc	0.993	0.990	0.971	0.995	0.982	0.903
Precision	0.293	0.257	0.996	0.467	0.361	0.144
Raw Data (with missing)						
	RF	kNN	ICO	JRip	PART	FLDA
Accuracy	0.958	0.953	0.964	0.964	0.964	0.883
Specificity	0.989	0.982	0.999	0.999	1.000	0.889
B.Acc	0.993	0.990	0.995	0.994	0.965	0.944
Precision	0.310	0.262	0.479	0.431	0.333	0.196

5. CONCLUSION

In this research, we suggested a layered feature selection to enhance the decision about Covid19 of Mexican patients. We reduced the dataset's dimensions using a two-stage feature selection technique based on Chi-square and Cuckoo and GWO search approaches. From the comparison results, it has been observed that the ICO classifier performs much better than others in most cases. Results showed that the second phase of feature selection using CSA and GWO techniques shows competitive results by selecting six and five features, respectively. This helped in reducing dimensionality by 70% and maintain an accuracy above 96%.

6. REFERENCES

[1] Centers for Disease Control and Prevention. Coronavirus Disease 2019 (COVID-19) Symptoms; U.S. Department of Health & Human Services: Atlanta, GA, USA, 2020. Available online: https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html (accessed on 1 March 2021).

[2] Méndez-Arriaga, F. (2020). The temperature and regional climate effects on communitarian COVID-19 contagion in Mexico throughout phase 1. Science of the Total Environment, 735, 139560.

[3] en México, R. H. D. N. (2003). Dirección General de Epidemiología, Secretaría de Salud. México.

[4] Gupta, M. K., & Chandra, P. (2020). A comprehensive survey of data mining. International Journal of Information Technology, 1-15.

[5] Ratner, B. (2017). Statistical and Machine-Learning Data Mining:: Techniques for Better Predictive Modeling and Analysis of Big Data. CRC Press.

[6] Srinivas, K., Rani, B. K., & Govrdhan, A. (2010). Applications of data mining techniques in healthcare and prediction of heart attacks. International Journal on Computer Science and Engineering (IJCSE), 2(02), 250-255.

[7] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

[8] Khanday, A. M. U. D., Rabani, S. T., Khan, Q. R., Rouf, N., & Din, M. M. U. (2020). Machine learning based

approaches for detecting COVID-19 using clinical text data. International Journal of Information Technology, 12(3), 731-739.

[9] Wang L, Wong A (2020) COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 Cases from chest radiography images. https://arxiv.org/abs/2003.09871

[10] Roda, W.C.; Varughese, M.B.; Han, D.; Li, M.Y. Why is it difficult to accurately predict the COVID-19 epidemic? Infect. Dis. Model. 2020, 5, 271–281.

[11] Roosa, K.; Lee, Y.; Luo, R.; Kirpich, A.; Rothenberg, R.; Hyman, J.M.; Yan, P.; Chowell, G. Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th, 2020. Infect. Dis. Model. 2020, 5, 256–263.

[12] Wang, H.; Wang, Z.; Dong, Y.; Chang, R.; Xu, C.; Yu, X.; Zhang, S.; Tsamlag, L.; Shang, M.; Huang, J.; et al. Phase-adjusted estimation of the number of Coronavirus Disease 2019 cases in Wuhan, China. Cell Discov. 2020.

[13] Narin A, Kaya C, Pamuk Z. Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks. arXiv preprint arXiv:2003.10849. 2020 Mar 24.

[14] Barstugan, M., Ozkaya, U., & Ozturk, S. (2020). Coronavirus (covid-19) classification using ct images by machine learning methods. arXiv preprint arXiv:2003.09424.

[15] Wiguna W, Riana D. Diagnosis of Coronavirus disease 2019 (Covid-19) surveillance using C4. 5 Algorithm. Jurnal Pilar Nusa Mandiri. 2020;16(1):71-80.

[16] Muhammad LJ, Islam MM, Sharif US, Ayon SI. Predictive Data Mining Models for Novel Coronavirus (COVID-19) Infected Patients Recovery.

[17] Yan L, Zhang H-T, Xiao Y, Wang M, Sun C, Liang J, Li S, Zhang M, Guo Y, Xiao Y, Tang X, Cao H, Tan X, Huang N, Amd A, Luo BJ, Cao Z, Xu H, Yuan Y (2020) Prediction of criticality in patients with severe covid-19 Infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan. medRxiv. https://doi.org/10.1101/2020.02.27. 20028027

[18] Jiang X, Coffee M, Bari A, Wang J, Jiang X, Huang J, Shi J, Dai J, Cai J, Zhang T, Wu Z, He G, Huang Y (2020) Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. Compu Mater Contin 63(1):537–551.

[19] https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge , (accessed on 1 March 2021).

[20] Gao, L., Song, J., Liu, X., Shao, J., Liu, J., & Shao, J. (2017). Learning in high-dimensional multimedia data: the state of the art. Multimedia Systems, 23(3), 303-313.

[21] Sorzano, C. O. S., Vargas, J., & Montano, A. P. (2014). A survey of dimensionality reduction techniques. arXiv preprint arXiv:1403.2877.

[22] Chauhan, D., & Mathews, R. (2019, December). Review on Dimensionality Reduction Techniques. In International conference on Computer Networks, Big data and IoT (pp. 356-362). Springer, Cham.

[23] Thaseen, I. S., & Kumar, C. A. (2017). Intrusion detection model using fusion of chi-square feature selection and multi class SVM. Journal of King Saud University-Computer and Information Sciences, 29(4), 462-472.

[24] Thaseen, I. S., & Kumar, C. A. (2017). Intrusion detection model using fusion of chi-square feature selection and multi class SVM. Journal of King Saud University-Computer and Information Sciences, 29(4), 462-472.

[25] Mirjalili, S., Saremi, S., Mirjalili, S. M., & Coelho, L. D. S. (2016). Multi-objective grey wolf optimizer: a novel algorithm for multi-criterion optimization. Expert Systems with Applications, 47, 106-119.

[26] Saremi, S., Mirjalili, S. Z., & Mirjalili, S. M. (2015). Evolutionary population dynamics and grey wolf optimizer. Neural Computing and Applications, 26(5), 1257-1263.

[27] Mirjalili, S., Mirjalili, S. M., & Lewis, A. (2014). Grey wolf optimizer. Advances in engineering software, 69, 46-61.

[28] Yang, X. S., & Deb, S. (2009, December). Cuckoo search via Lévy flights. In 2009 World congress on nature & biologically inspired computing (NaBIC) (pp. 210-214). Ieee.

[29] Yang, X. S., & Deb, S. (2010). Engineering optimisation by cuckoo search. International Journal of Mathematical Modelling and Numerical Optimisation, 1(4), 330-343.

[30] Gandomi, A. H., Yang, X. S., & Alavi, A. H. (2013). Cuckoo search algorithm: a metaheuristic approach to solve structural optimization problems. Engineering with computers, 29(1), 17-35.

[31] Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. ISPRS journal of photogrammetry and remote sensing, 114, 24-31.

[32] Fawagreh, K., Gaber, M. M., & Elyan, E. (2014). Random forests: from early developments to recent advancements. Systems Science & Control Engineering: An Open Access Journal, 2(1), 602-609.

[33] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[34] García, V., Mollineda, R. A., & Sánchez, J. S. (2008). On the k-NN performance in a challenging scenario of imbalance and overlapping. Pattern Analysis and Applications, 11(3), 269-280.

[35] Rodrigues, É. O. (2018). Combining Minkowski and Cheyshev: New distance proposal and survey of distance metrics using k-nearest neighbours classifier. Pattern Recognition Letters, 110, 66-71.

[36] Meshram, S. G., Safari, M. J. S., Khosravi, K., & Meshram, C. (2020). Iterative classifier optimizer-based pace regression and random forest hybrid models for suspended sediment load prediction. Environmental Science and Pollution Research, 1-13.

[37] Williams, M. L., Mac Parthaláin, N., Brewer, P., James, W. P. J., & Rose, M. T. (2016). A novel behavioral model of the pasture-based dairy cow from GPS data using data mining and machine learning

techniques. Journal of dairy science, 99(3), 2063-2075.

[38] Venkatesh, N., & Jayaraman, S. (2010, August). Human electrocardiogram for biometrics using DTW and FLDA. In 2010 20th International Conference on Pattern Recognition (pp. 3838-3841). IEEE.

[39] Ali, S., & Smith, K. A. (2006). On learning algorithm selection for classification. Applied Soft Computing, 6(2), 119-138.

[40] Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD explorations newsletter, 6(1), 20-29.

[41] Bekkar, M., Djemaa, H. K., & Alitouche, T. A. (2013). Evaluation measures for models assessment over imbalanced data sets. J Inf Eng Appl, 3(10).

[42] Grzymala-Busse, J. W., & Hu, M. (2000, October). A comparison of several approaches to missing attribute values in data mining. In International Conference on Rough Sets and Current Trends in Computing (pp. 378-385). Springer, Berlin, Heidelberg.

[43] Jolliffe, I. T. (1986). Principal components in regression analysis. In Principal component analysis (pp. 129-155). Springer, New York, NY.

[44] De Souto, M. C., Jaskowiak, P. A., & Costa, I. G. (2015). Impact of missing data imputation methods on gene expression clustering and classification. BMC bioinformatics, 16(1), 1-9.