Mansoura Engineering Journal

Volume 47 | Issue 2 Article 5

4-27-2022

Covid-19 Patients Diagnosis (CPD) Strategy Using Data Mining Techniques

Alaa Mohamed
Master Degree Researcher of Electronics and Communication Department, Faculty of Engineering, Mansoura University works at Delta Higher Institute for Engineering and Technology, alaa.mostafa545454@gmail.com

Ahmed Saleh
Professor at the Computers and Control Department, Faculty of Engineering, Mansoura University, Egypt., aisaleh@yahoo.com

Doaa A. Altantawy
Assistant Professor at the Electronics and Communication Department, Faculty of Engineering, Mansoura University, Egypt, doaa1adel@mans.edu.eg

Mohy Eldin Abo-Elsoud
Professor at the Electronics and Communication Department, Faculty of Engineering, Mansoura University, Egypt, mohyldin@gmail.com

Recommended Citation

Mohamed, Alaa; Saleh, Ahmed; A. Altantawy, Doaa; and Abo-Elsoud, Mohy Eldin (2022) "Covid-19 Patients Diagnosis (CPD) Strategy Using Data Mining Techniques," *Mansoura Engineering Journal*: Vol. 47 : Iss. 2 , Article 5.
Available at: https://doi.org/10.21608/bfemu.2022.233811

Mansoura University
Faculty of Engineering
Mansoura Engineering Journal

Covid-19 Patients Diagnosis (CPD) Strategy Using Data Mining Techniques

Alaa M. Mohamed*, Ahmed I. Saleh, Doaa A. Altantawy and M.A. Abo-Elsoud

KEYWORDS:
Covid-19 diagnosis, data mining, IKNN, feature selection

***Abstract*— Covid-19, the world continues to live in anxiety and instability despite efforts to find a vaccine and emerge from this crisis. Especially after the emergence of a new mutated Corona virus called Omicron. This mutated sparked a state of controversy about the extent of its impact and its ability to spread among people. The Covid-19 epidemic has thrown the world economy into disarray. It also resulted in a widespread suspension of work and output throughout society, hurting economic and society. This paper introduces a Covid-19 Patients Diagnosis (CPD) strategy that works to find a fast and highly effective prognosis for diagnosing Covid-19 patients. The proposed strategy has two main stages named Feature Selection Stage (FSS) and Covid-19 Diagnosis Stage (CDS). The FSS has main objective to select the powerful features for the diagnosis stage. The features are selected in the FSS by using Chi-Square Feature Selection (CSFS) method. In fact, CSFS is a filter feature selection technique that has the ability to quickly choose the most effective subset of features. Then, quick and accurate diagnosis is provided by using Improved K-Nearest Neighbors (IKNN). The main idea in the proposed IKNN is that a circle with a radius value that equals the average distance of K of the closet items will be constructed and then the nearest M of items will be determined to classify the patient to the correct class "Covid" or "Non-Covid". The results explain that the proposed strategy called CPD gives an accuracy of up to 96.36%.**

I. INTRODUCTION

EVERY day that passes in the world, the number of infections and deaths increases as a result of Covid-19 this infectious epidemic. The world has lost millions of people due to this epidemic so far, as the number of cases reached 271,963,258 cases, and deaths reached 5,331,019 cases, according to the latest statistics issued by the World Health Organization (WHO) at the time of writing this paper [1]. Symptoms of Covid-19 disease are fever, fatigue, cough, shortness of breath, and headache, where the severity of symptoms varies based on immunity and health history of the patient, and it takes about 4-6 days to appear [2]. This epidemic, which continues despite its appearance nearly two years ago, and with the continuous efforts of the medical staff and researchers to find a solution to this crisis, rapid and effective diagnosis is the best solution to reduce the risks of this disease and the possibility of its occurrence, as well as to provide appropriate medical services for the affected person.

Real-Time Reverse Transcription Polymerase Chain Reaction (RT-PCR) has been implemented for Covid-19 diagnosis [2]. But it takes 2–3 hours to provide findings and

Received: (09 January, 2022) - Revised: (20 February, 2022) - Accepted: (24 February, 2022)

***Corresponding Author,** Alaa M. Mohamed, at Delta Higher Institute for Engineering and Technology, Egypt, Master Degree Researcher of Electronics and Communication Department, Faculty of Engineering, Mansoura University. (e-mail: Alaa.mostafa545454@gmail.com)*

Ahmed I. Saleh, Professor at the Computers and Control Department, Faculty of Engineering, Mansoura University, Egypt. (e-mail: aisaleh@yahoo.com)

Doaa A. Altantawy, Assistant Professor at the Electronics and Communication Department, Faculty of Engineering, Mansoura University, Egypt. (e-mail: doaa1adel@mans.edu.eg)

M.A. Abo-Elsoud, Professor at the Electronics and Communication Department, Faculty of Engineering, Mansoura University, Egypt (e-mail: mohyldin@gmail.com)

necessitates the use of recognized laboratories, expensive equipment, and well-trained employees [3]. However, numerous recent investigations have revealed up to 20% false-negative outcomes [4]. Covid-19 diagnosis has also been done using Computed Tomography (CT) images. However, the vast majority of Covid-19 patients have a natural chest CT scan. Furthermore, physicians and other patients are at risk when using imaging equipment for Covid-19 patients. There is a high probability that the virus will survive on the scanning room's surface even with careful cleaning [2]. Because of these drawbacks, RRT-PCR and CT are inappropriate for large-scale screening aimed at a speedy diagnosis of patients.

Blood testing is another method was used in many studies for detected covid-19 patient because Covid-19 patient has different characteristics is linked with significantly decreased lymphocyte stiffness, increased monocyte cell size, the appearance of smaller and less deformable erythrocytes, and the presence of large, deformable, activated neutrophils [5]. As a result, blood analysis employed in this study to reduce the risks associated with the use of other methods, maintain medical staff, and reduce the possibility of Covid-19 spreading as a result of direct contact between patients and medical personnel.

Data mining (DM) is an artificial intelligence method for effective knowledge, and discovering new in a dataset. It's also been used to diagnose and predict a variety of disorders, including coronavirus and coronavirus Middle East respiratory syndrome [6]. The massive global dataset linked to the Covid-19 pandemic on a daily basis is a significant resource that should be collected and evaluated for important, valid, and pattern analysis to help stop the sawing of the Covid-19 virus. In the healthcare area, DM is most used in a variety of applications. It is frequently employed in the healthcare profession for a variety of purposes, including forecasting patient performance, modeling health outcomes, monitoring treatment and infection efficacy, and hospital rating [2].

This paper provides a new strategy called Covid-19 Patients Diagnosis (CPD) strategy that is divided into two stages called Feature Selection Stage (FSS) and Covid-19 Diagnosis Stage (CDS). The FSS aims to select the meaningful features for the next CDS. In the FSS, the most effective subset of features is selected by using Chi-Square Feature Selection (CSFS) method as a filter selection method that can select the most effective features quickly. Then, the accurate and quick diagnosis is provided by using Improved K-Nearest Neighbors (IKNN) that depends on using the average distance of the nearest k items that is used as a radius of a constructed circle to determine the nearest M items inside the circle based on its center.

This paper is structured as follows; section 2 summarizes previous work for Covid-19 diagnosis methods. Section 3 discusses the proposed Covid-19 patient diagnosis approach. Section 4 displays the experimental results. Section 5 summarizes the findings and recommendations for future research.

II. RELATED WORK

In this section, the prior studies efforts on Covid-19 diagnosis techniques will be discussed. In [7], Distance Biased Naïve Bayes (DBNB) was introduced to diagnose Covid-19 cases using laboratory tests. The DBNB goes through two basic stages to be able to classify Covid-19. The first step included Advanced Particle Swarm Optimization (APSO) which combines filter and wrapper methods to choose more useful features from the dataset. Then, these features were used to classify Covid-19 patients in the second step using DBNB which is used to overcome the drawback of classical NB. DBNB has two modules in which the first module was used to weight the selected features while the second was used to take a final decision depending on the distance between the center of the target class and the input patient that needs to be classified.

As presented in [8], a support system for Covid-19 diagnosis using blood tests named Heg.IA. In preprocessing step, the dataset has a numeric form. Then, Particle Swarm Optimization (PSO) and Evolutionary Search (ES) were used as a feature selection technique. To detect Coivd-19 patients using several classification techniques, but Bayes net method achieves high accuracy equal 95%.

In [9], a Covid-19 Diagnostic Technique (CDT) for Covid-19 detection based on a blood sample was introduced. CDT has applied a Genetic Algorithms (GA) and relief algorithm to select the most accurate feature from the dataset, which consists of 100 features. Then, a random forest classifier was used to classify the dataset .The result shows that the model has good accuracy. As introduced in [10], a new Corona Patients Detection Strategy (CPDS) was represented to check Covid-19 patients. CPDS was implemented by applying two steps. The first is a Hybrid Feature Selection Methodology (HFSM) for extracting the best feature from the dataset. The second stage is an Enhanced K-Nearest Neighbor (EKNN) implemented as a classifier for detecting Covid-19 patients.

As presented in [11], a novel Handcrafted Feature Generation Technique and a Hybrid Feature Selector (HFGT-HFS) were introduced as an automatic COVID-19 detection method. In HFGT-HFS feature generation was implemented for the select statistical and textural features. Then classification was performed using two classifiers called (i) Artificial Neural Networks (ANN) as well as (ii) Deep Neural Network (DNN). The experimental results in [11] explain that the ANN and DNN models provided classification accuracy of 94.10% and 95.84% respectively.

III. THE PROPOSED COVID-19 PATIENTS DIAGNOSIS (CPD) STRATEGY

Automated medical diagnosis is becoming increasingly relevant, particularly when making quick decisions about dangerous viral disorders like Covid-19 [12, 2,4]. Covid-19 patients must be diagnosed as soon as possible because their own interaction with others is increasing unsurvivors number on a daily basis. As a result, direct interaction with Covid-19 patients may endanger the lives of medical staff, putting them at risk of death. This paper introduces a

Covid-19 Patients Diagnosis (CPD) technique to deliver more accurate and timely diagnosis findings in order to address this worldwide and dangerous dilemma. As represented in Fig.1, the proposed CPD strategy consists of two main stages, which are; (i) Feature Selection Stage (FSS) and (ii) Covid-19 Diagnosis Stage (CDS). While FSS aims to extract the powerful subset of features, CDS try to provide accurate diagnosis depended on the extracted features from FSS. The details of FSS and CDS stages will be described in the next subsections.

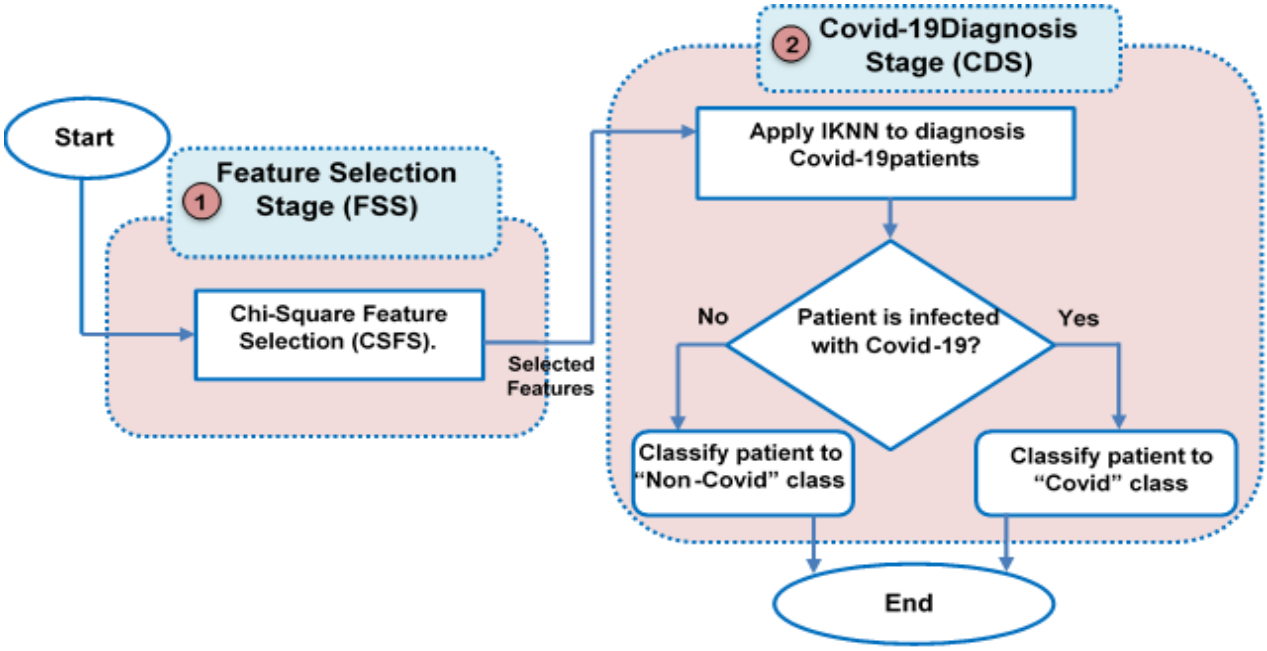


Fig. 1. The Covid-19 Patient Diagnosis (CPD) strategy.

A. *Feature Selection Stage (FSS)*

In the FSS, Chi-Squared Feature Selection (CSFS) method is used for measuring the goodness of a feature [13,14]. The CSFS calculates the degree of independence between a two of variables belonging to the same class category using (1)

$$CSFS = \sum_{i=1}^{n}\sum_{a=1}^{2}\frac{(O_{ia}-R_{ia})^2}{R_{ia}} \quad (1)$$

Where $R_{ia} = \frac{O_a * O_i}{N}$

Where O_{ia} is the number of samples with i^{th} feature value in a^{th} class, O_i is the number of samples with i^{th} feature value. O_a is a number of samples in class a, and N no of samples.

B. *Covid-19 Diagnosis Stage (CDS)*

In the CDS, Improved K-Nearest Neighbors (IKNN) is designed as a proposed classification method. It is used to precisely classify Covid-19 patients. IKNN is a new instance of KNN classifier that can enhance the efficiency of the classification in which it can be used in various applications including Covid-19 patient classification. Classical KNN is considered as one of the simplest and straightforward classification methods that can be easily understood [15]. It can handle multi-class problems and can give accurate classifications when the neighbors are correctly identified [16]. Hence, KNN can make a decision based on the training objects in the training phase in which it measures the distance between the testing object (e.g., Covid-19 patient) and every object in the training dataset. Then, it identifies the K nearest objects to classify the testing object (unclassified object) and assigns it to the majority vote class of those *K* objects.

Consequently, there are two main operators to execute KNN which are; (i) the value of K that represents the neighbors count to be considered and (ii) the distance measure. Euclidean distance is a popular method used to calculate the distance between a new object and the training objects [16]. Despite KNN's simplicity, it is a lazy learner in which its calculations depend on the value of K to classify the testing object according to the majority vote of those K objects. It is noted that, in the case of the neighbors are incorrectly identified, the testing object (e.g., patient) will be incorrectly classified [17]. Moreover, depending only on the nearest K of training objects which close to the testing object after calculating the distance between them may give misclassification. Hence, KNN does not have the ability to give accurate classifications in real-time applications such as Covid-19 patients classification [15].

In this paper, IKNN is introduced to enhance the KNN's performance by deducing a circle around the testing object located at the center of this circle. The circle has a radius equal to the average distance of the nearest K of training objects which is close to this testing object. According to a set of training objects inside the deduced circle, a new number of the nearest neighbors to the testing object (center) of this circle donated by *M* symbol will be determined. Finally, the majority vote of those *M* objects will be implemented inside the circle to accurately classify the testing object to its corresponding class.

To implement IKNN classifier, suppose that there are '*n*' dimensional Feature Space; $FS=\{f_1, f_{2,\ldots,}f_n\}$. Additionally, an input training object *T* can be represented to FS as; $T(f_1, f_{2,\ldots,} f_n)=\{f_{1T}, f_{2T,\ldots,}f_{nT}\}$ and Covid-19 patient in testing dataset *E* as $E(f_1, f_{2,\ldots,} f_n)=\{f_{1E}, f_{2E,\ldots,}f_{nE}\}$. Finally, there are '*c*' classes denoted as; $C_{categories}=\{Covid, Non\text{-}Covid\}$. To clarify the idea, suppose that there are 2-dimensional feature space $F=\{f_1, f_2\}$. Fig.2 illustrates the sequential steps of IKNN method to classify the patient to *"Covid"* or *"Non-Covid"* class according to f_1 and f_2 features. According to the covid-19 dataset, it is assumed that the input training data of *'h'* objects (patients) is represented by $A=\{T_1, T_{2,\ldots,} T_h\}$. Also, the testing data of *'q'* patients is represented by $Q=\{E_1, E_{2,\ldots,}E_q\}$. Hence, each data object of $T_i \in A$ and $E_j \in Q$ is expressed as an ordered set of *'n'* features in *'n'* dimensional space. Accordingly, the expression is as following; $T_i(f_1, f_{2,\ldots,} f_n)=\{f_{1i}, f_{2i,\ldots,}f_{ni}\}$ and $E_j(f_1, f_{2,\ldots,} f_n)=\{f_{1j}, f_{2j,\ldots,} f_{nj}\}$. The implementation of IKNN is as same as KNN in which two techniques require to extract appropriate distance measurement methods. Hence, this paper uses Euclidean distance as it is the most common method [17].

There are many steps to implement IKNN that start from the calculation of distance between the testing object's data $E_j(f_1, f_{2,\ldots,} f_n)= \{f_{1j}, f_{2j,\ldots,}f_{nj}\}$ and the training object's data $T_i(f_1, f_{2,\ldots,}f_n)= \{f_{1i}, f_{2i,\ldots,}f_{ni}\}$ in the *'n'* dimensional space by using (2).

$$D(E_j, T_i) = \sqrt{\sum_{x=1}^{n} (f_{xj} - f_{xi})^2} \qquad (2)$$

Where $D(E_j, T_i)$ denotes to the Euclidean distance between two objects E_j and T_i at features (1,2,…,n). fx_j is the data value of j^{th} testing object (E_j) at x^{th} feature, fx_i is the data value of i^{th} training object (T_i) at x^{th} feature, and *n* refers to the total features number. Then, it determines the nearest K of the input training objects which close to E_j testing object according to the smallest distances. Secondly, the average distance (r) for the nearest K of training objects which close to E_j testing object is calculated by using (3).

$$r = \frac{1}{K}\sum_{l=1}^{K} D_l \qquad (3)$$

Where *r* is the average distance of the nearest *K* training objects that represents a radius of the circle containing the testing object E_j in its center. *K* denotes the number of training objects which close to the testing object E_j. D_l indicates to the distance of l^{th} training object that is closed to the testing object. Depending on the radius (r), a circle should be created based on the testing object E_j in the center. Finally, the testing object E_j can be classified into its relevant class category depending on the maximum votes of the nearest *M* of the training objects inside the circle. It is noted that, *M* must be an odd value and does not exceed the number of training objects inside the circle. The major steps of the proposed IKNN are explained in algorithm 1.

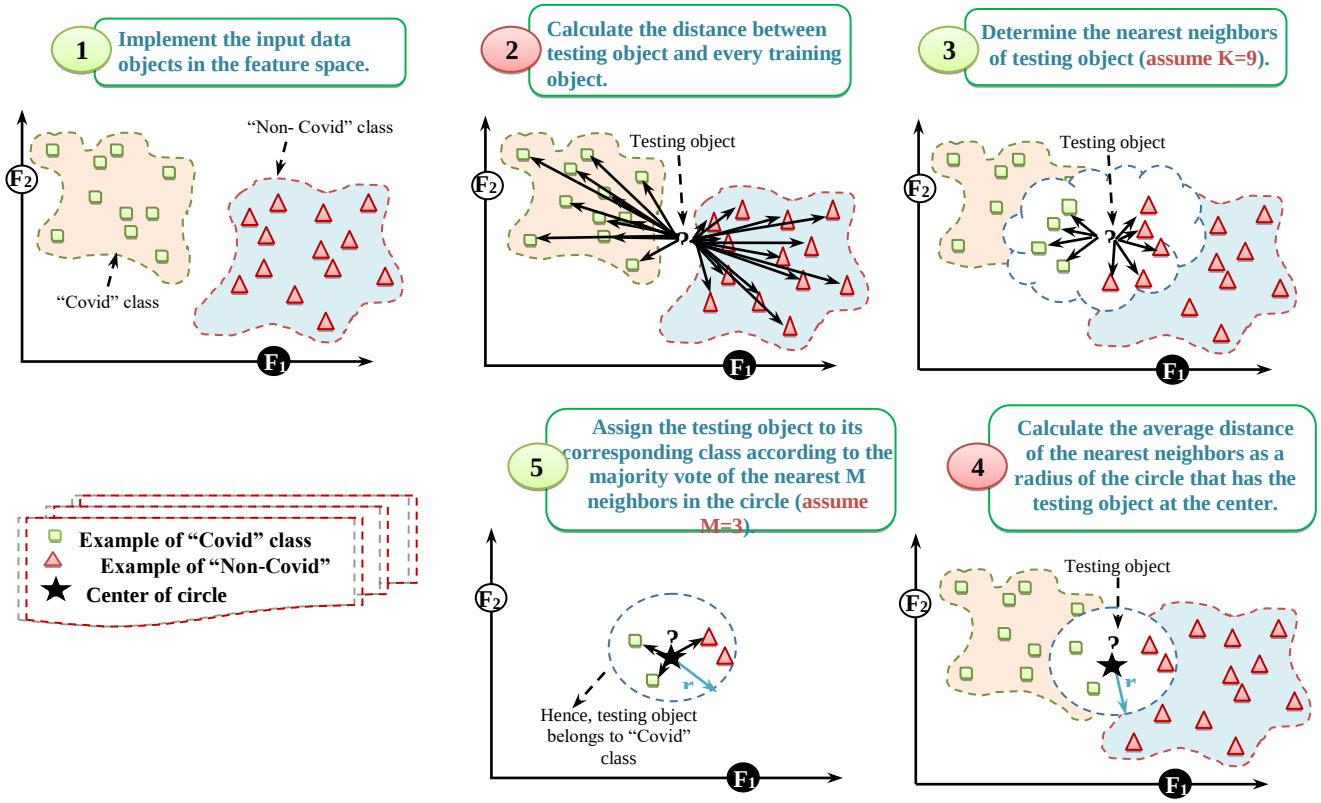


Fig. 2. Steps of the Improved KNN using two features (F1, F2) and two classes ("Covid", "Non-Covid").

IV. EXPERIMENTAL RESULTS

The major goal of this paper is to use data mining methods to improve classification performance and diagnostic accuracy. So, a new diagnostic method called Covid-19 Patients Diagnosis (CPD) strategy has been provided. CPD has two main stages; Feature Selection Stage (FSS) and Covid-19 Diagnosis Stage (CDS). In FSS, the most effective features will be selected. Then, these features will be used in the next stage called CDS. In CDS, IKNN classifier will be used for providing accurate classification. The performance of the proposed CDS strategy will be measured against other recent strategies based on two different datasets using Confusion Matrix metrics [2]. One of these datasets is blood tests called "Data1" and another data is CT images called "Data 2". Also, the speedup of the proposed CDS strategy against other recent strategies will be calculated.

A. The Used Datasets Explanation

In this paper, two datasets which are called Data1 and Data 2 are used to test the performance of the proposed CDS. Data1 contains a group of "Covid" and "Non-Covid" patient laboratory data. The results given in this paper will depend on a Data1 that consists of 1624 patients and 34 features [18]. After the CSFS method was implemented in the feature selection stage, the number of most important features became 19 features. Training and testing datasets have been separated. There are 1300 cases used as training patients (80% of total datasets) and 324 testing patients (20% of total datasets).

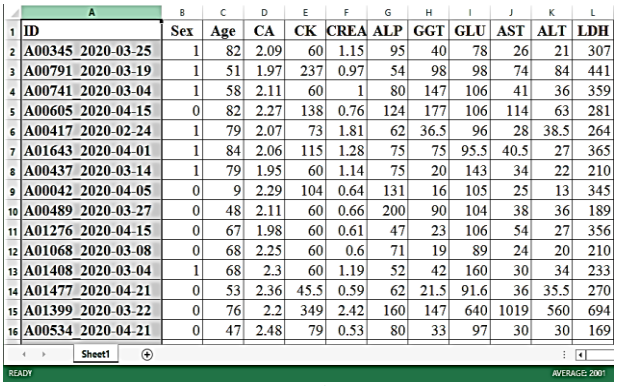
Data 2 [19, 20] is the second dataset used in this paper. Data 2 consists of 812 CT images for patients. Training and testing datasets have been separated. There are 650 training CT images (80% of total datasets) and 162 testing CT images (20% of total datasets). Snapshots of the used datasets as examples of them are presented in Fig.3 and Fig.4 respectively. Additionally, more details about Data1 and Data 2 are presented in Table 1 respectively. These details include number of infection patient, and gender of cases.

B. The Used Evaluation Performance Metrics

The performance of the CDS strategy will be tested using five metrics in the following section. The metrics are accuracy, error, recall, precision, and run time. These parameters are calculated using confusion matrix during the implementation of CDS strategy. Fig.5 and Fig.6 show the results of confusion matrix for the CDS strategy based on Data1 and Data 2 respectively.

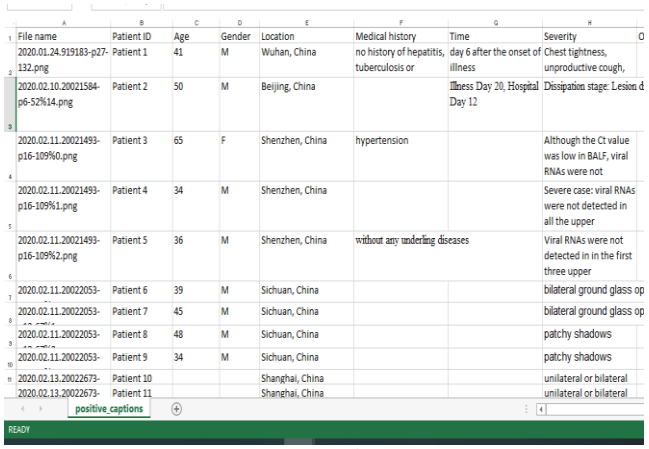
TABLE 1
DATA DESCRIPTION.

Criteria		value	
		Covid-19	**Non-Covid-19**
sick cases	*Data1*	786	838
	Data 2	349	463
		Female	**Male**
Cases gender	*Data1*	39.03%	60.9%
	Data 2	22.17%	77.82%



ID	Sex	Age	CA	CK	CREA	ALP	GGT	GLU	AST	ALT	LDH
A00345_2020-03-25	1	82	2.09	60	1.15	95	40	78	26	21	307
A00791_2020-03-19	1	51	1.97	237	0.97	54	98	98	74	84	441
A00741_2020-03-04	1	58	2.11	60	1	80	147	106	41	36	359
A00605_2020-04-15	0	82	2.27	138	0.76	124	177	106	114	63	281
A00417_2020-02-24	1	79	2.07	73	1.81	62	36.5	96	28	38.5	264
A01643_2020-04-01	1	84	2.06	115	1.28	75	75	95.5	40.5	27	365
A00437_2020-03-14	1	79	1.95	60	1.14	75	20	143	34	22	210
A00042_2020-04-05	0	9	2.29	104	0.64	131	16	105	25	13	345
A00489_2020-03-27	0	48	2.11	60	0.66	200	90	104	38	36	189
A01276_2020-04-15	0	67	1.98	60	0.61	47	23	106	54	27	356
A01068_2020-03-08	0	68	2.25	60	0.6	71	19	89	24	20	210
A01408_2020-03-04	1	68	2.3	60	1.19	52	42	160	30	34	233
A01477_2020-04-21	0	53	2.36	45.5	0.59	62	21.5	91.6	36	35.5	270
A01399_2020-03-22	0	76	2.2	349	2.42	160	147	640	1019	560	694
A00534_2020-04-21	0	47	2.48	79	0.53	80	33	97	30	30	169

Fig.3. A snapshot from the Data1.



File name	Patient ID	Age	Gender	Location	Medical history	Time	Severity
2020.01.24.919183-p27-132.png	Patient 1	41	M	Wuhan, China	no history of hepatitis, tuberculosis or	day 6 after the onset of illness	Chest tightness, unproductive cough,
2020.02.10.20021584-p6-52%14.png	Patient 2	50	M	Beijing, China		Illness Day 20, Hospital Day 12	Dissipation stage: Lesion d
2020.02.11.20021493-p16-109%0.png	Patient 3	65	F	Shenzhen, China	hypertension		Although the Ct value was low in BALF, viral RNAs were not
2020.02.11.20021493-p16-109%1.png	Patient 4	34	M	Shenzhen, China			Severe case: viral RNAs were not detected in all the upper
2020.02.11.20021493-p16-109%2.png	Patient 5	36	M	Shenzhen, China	without any underling diseases		Viral RNAs were not detected in in the first three upper
2020.02.11.20022053-	Patient 6	39	M	Sichuan, China			bilateral ground glass op
2020.02.11.20022053-	Patient 7	45	M	Sichuan, China			bilateral ground glass op
2020.02.11.20022053-	Patient 8	48	M	Sichuan, China			patchy shadows
2020.02.11.20022053-	Patient 9	34	M	Sichuan, China			patchy shadows
2020.02.13.20022673-	Patient 10			Shanghai, China			unilateral or bilateral
2020.02.13.20022673-	Patient 11			Shanghai, China			unilateral or bilateral

Fig.4. A snapshot from the Data 2.

C. Experiment the Proposed Covid-19 Patient Diagnosis (CPD) Strategy

In this section, the IKNN is compared to the classical KNN using feature selection method called CSFS and also without using CSFS method. The IKNN in both cases; using CSFS and without using CSFS outperforms the classical KNN in terms of accuracy error, sensitivity, precision, and run time as presented in Table 2 and Table 3 according to Data1 and Data 2 datasets respectively. In Table 2 and Table 3, the IKNN with CSFS method gives the best results because it can provide the maximum accuracy, precision, and recall values but it can provide the minimum error and execution time. On the other hand, the classical KNN without using feature selection technique provides the worst results according to both datasets. It is noted that the results of the compared methods in Table 2 is better than the results in Table 3. Thus, blood tests dataset

(Data1) can provide more accurate results than CT image dataset (Data 2).

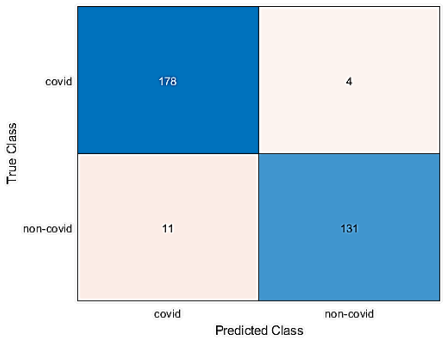


Fig.5. The result of confusion matrix for the CDS strategy based on the Data1.

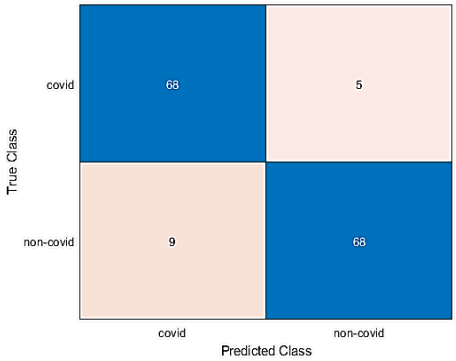


Fig.6. The result of confusion matrix for the CDS strategy based on the Data 2.

Improved KNN Algorithm

- ***Inputs:***
  - *TDS= (D, F); Training dataset.*
  - *LTRO=$L_{tr1};\ldots;L_{trH}$; Labels of training objects*
  - *H=|LTDO| or |D|; The number of objects or labels in training data set.*
  - *TED=(Q,F); Testing dataset (COVID-19+ cases) .*
  - *LTEO= $L_{te1};\ldots;L_{teS}$; Labels of testing objects.*
  - *S=|LTEO| or |Q|; The number of objects or labels in testing data set.*
  - *A=|F|; No. of features in training and testing data set.*
  - *TC=$c_1;\ldots;c_T$; Target classes of medical classification.*
  - *T =| T_C |; No. of classes in the system.*
  - *K= No. of nearest neighbors*
- ***Output:***
  - *Cr=$Cr_1;\ldots\ldots Cr_S$; Classification result for each test object.*
- ***Steps:***

/*Calculate Euclidean distance between each testing object in Q and every training object in D */

1: For every $E_t \in Q$
2: For every $I_i \in D$
3: Calculate $E_{Dist}(E_t, I_i) = \sqrt{\sum_{x=1}^{W}(f_{xt} - f_{xj})^2}$.
4: End For
5: End For

/*Determined K- nearest neighbor of training objects for each testing object */

6: For every $E_t \in Q$
7: For every $I_i \in D$
8: Calculate Neighbors (E_t)=K of training objects with smallest distance $D_{Dist}(E_t, I_i)$.
9: End For
10: End For

/*Calculate average value of K- nearest neighbor of training objects */

11: For every $E_t \in Q$
12: Calculate $= \frac{\sum_{d=1}^{K} D_i}{K}$.
13: End For
14: Draw a circle with a radius "r" and its center " testing object".

Algorithm Parameters	
TDS	Training data set, TRD= (D, F).
D	Training objects.
F	The features of training or testing objects, F= $f_1;\ldots;f_x$.
LTRO	Labels of training objects, LTRI=$L_{tr1};\ldots;L_{trH}$.
H	The number of objects or labels in training data set, H=|LTRO| or |D|.
TED	Testing data set, TED= (Q, F).
LTEO	Labels of testing objects, LTEO=$L_{te1};\ldots;L_{trS}$.
Q	Testing objects.
S	No. of objects or labels in testing data set, S=|LTEO| or |Q|.
A	The number of features in training and testing data set, A=|F|.
TC	Classes of medical classification; $T_C=c_1;c_2;\ldots;c_T$.
T	No. of classes in the system, T =|TC |;
E_t	Test object belong to Q; $E_t \in Q$
I_i	Training object belong to D; $I_i \in D$
f_j	Feature of used object.
f_{xtc}	The values of the feature x at E_t
f_{xjc}	The values of the feature x at I_j
$D_{Dist}(E_t, I_i)$	Distances between testing object E_t and training object I_i.
K	No. of nearest neighbors.
r	Radius of the circle that has test object on its center
Z	No. of object inside the circle.
M	Odd No. as a nearest neighbors inside a circle.
Cr	Classification results.

/* classify the testing object according to the majority of the votes of the training objects inside the circle*/.

15: For every $E_t \in Q$
16: For every M $\in$ Z
17: If (Z == even number)
18: set (M < Z) & M: odd number.
19: End If
20: Cr =Max. votes (M)
21: End for
22: End for
23: Measure an accuracy of the proposed classifier.

Algorithm 1: Covid-19 patients' diagnosis using the Improved KNN algorithm.

TABLE 2
A COMPARISON BETWEEN IKKN AND NORMAL KNN USING DATA1.

Techniques	IKNN+CSFS	IKNN without CSFS	KNN+CSFS	KNN without CSFS
Accuracy	96.36%	94.54%	90.90%	72.59 %
Error	3.63%	5.46%	9.1%	27.41%
Sensitivity	100%	88.88%	82.75%	83.50%
Precision	92.30%	96.12%	92.21%	80.41%
Run Time	22.62 sec.	30.20 sec.	27.14sec.	40.84 sec.

TABLE 3
A COMPARISON BETWEEN IKKN AND NORMAL KNN USING DATA 2.

Techniques	IKNN+ CSFS	IKNN without CSFS	KNN+CSFS	KNN without CSFS
Accuracy	90.07%	86.41%	80.13%	61.18%
Error	9.93%	13.59%	19.37%	38.82%
Sensitivity	97.80%	93.17%	88.42%	62.96%
Precision	94.18%	91.71%	85.18%	60.71%
Run Time	96.45 sec.	110.20 sec.	54.63 sec.	78.59 sec.

D.Experiment the Proposed Covid-19 Patient Diagnosis (CPD) Strategy against other recent strategies

The CPD strategy will be compared to many of the most commonly utilized Covid-19 diagnosis methodologies to ensure that it is effective in term of accuracy, error, precision, sensitivity, and run time. Those methods are Heg.IA [8], CPDS [10], and HFGT-HFS [11]. As shown in Table 4, the proposed CPD strategy reaches to the highest accuracy while it reaches to the lowest error and run-time values according to Data1 and Data 2. This demonstrates the effectiveness of CPD as its phases can effectively collaborate. The results based on Data1 that contains blood tests dataset are shown in Figs.(7→11) and the results based on Data 2 that contains CT images dataset are shown in Figs.(12→16).

Figs.(7→11) show that Heg-IA, CPDS HFGT-HFS, and CPD provide accuracy values which are 0.93, 0.89, 0.85, and 0.963 respectively. Accordingly, Heg-IA, CPDS, HFGT-HFS, and CPD techniques have error values which are 0.7, 0.11, 0.15, and 0.036 respectively. The Sensitivity of CPD is 1, while Heg-IA, CPDS, and HFGT-HFS values are 0.92, 0.90, and 0.87 respectively. CPD has precision value 0.92 while Heg-IA, CPDS, and HFGT-HFS give 0.90, 0.89, and 0.85 respectively. Table 4 shows a comparison between CPD and the common diagnostic strategies at 1300 patients. The proposed CPD strategy outperformed the other compared diagnostic strategies, as it gives a high accuracy value and a small error rate. As show in Table 4, HFG-HFS gives poor accuracy with high error.

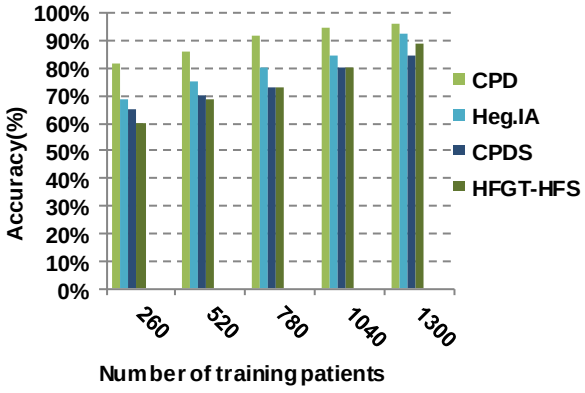


Fig.7. Accuracy of several diagnostic strategies.

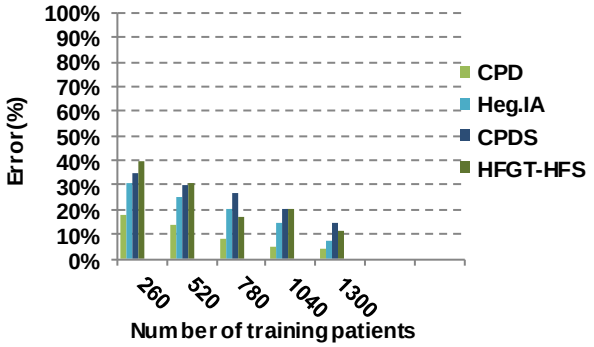


Fig.8. Error of several diagnostic strategies.

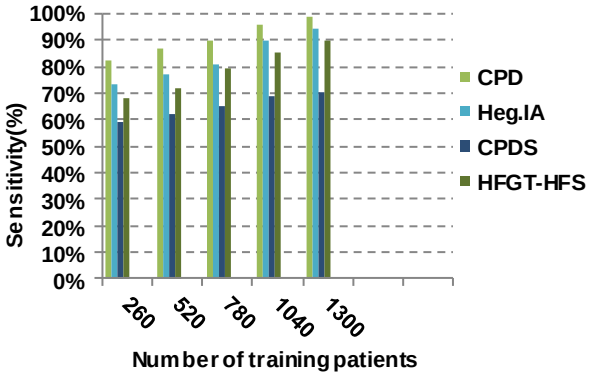


Fig.9. Sensitivity of several diagnostic strategies

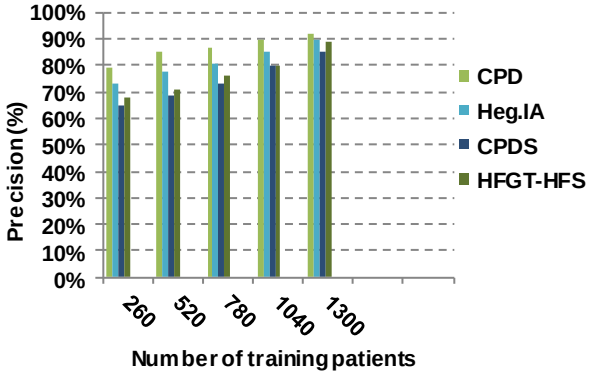


Fig.10. Precision of several diagnostic strategies

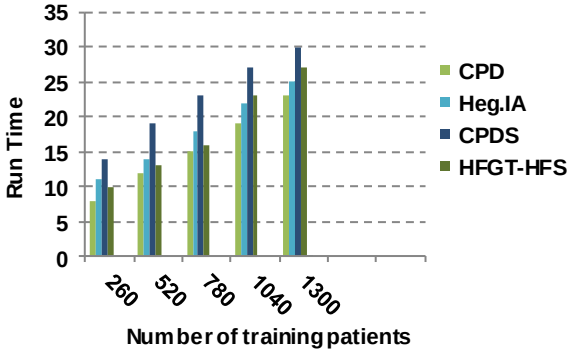


Fig.11. Rum time of several diagnostic strategies

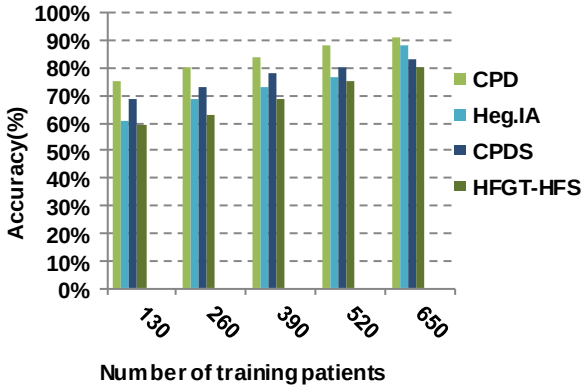


Fig.12. Accuracy of several diagnostic strategies

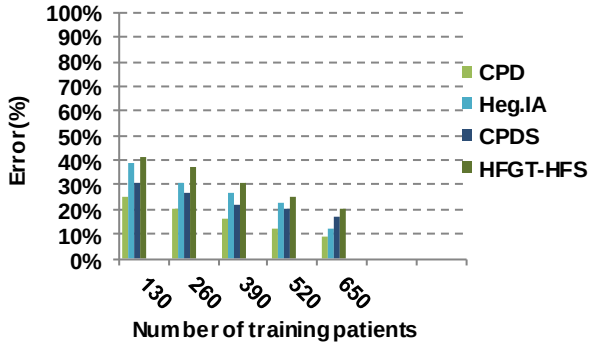


Fig.13. Error of several diagnostic strategies.

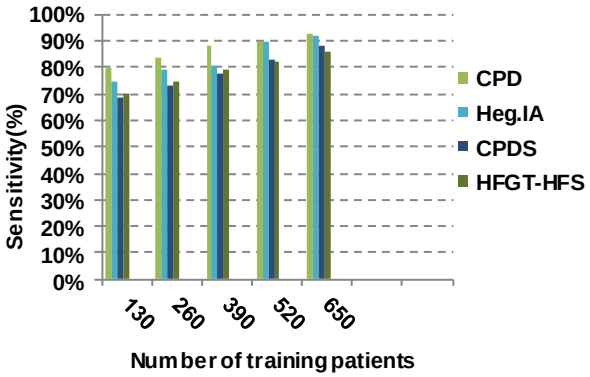


Fig.14. Sensitivity of several diagnostic strategies

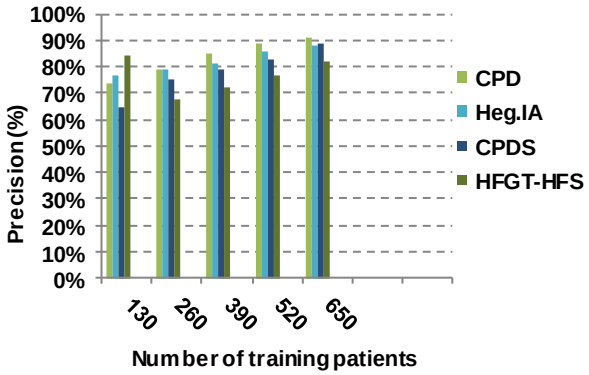


Fig.15. Precision of several diagnostic strategies

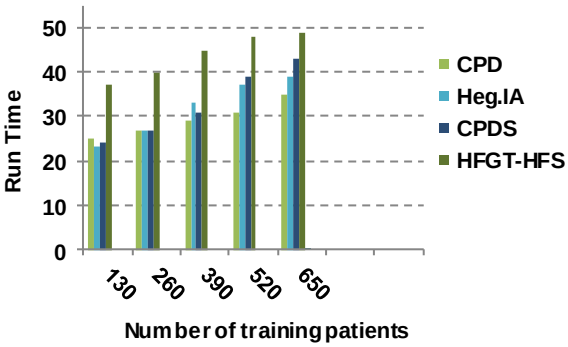


Fig.16. Rum time of several diagnostic strategies.

Figs.(12→16) show that Heg-IA, CPDS, HFGT-HFS, and CPD provide accuracy values which are 0.88, 0.80, 0.83, and 0.90 respectively. Accordingly, Heg-IA, CPDS, HFGT-HFS, and CPD techniques have error values which are 0.11, 0.19, 0.16, and 0.9 respectively. The Sensitivity of CPD is 0.93, while Heg-IA, CPDS, and HFGT-HFS values are 0.92, 0.86, and 0.88 respectively. CPD has precision value 0.91 while Heg-IA, CPDS, and HFGT-HFS give 0.88, 0.82, and 0.89 respectively. Table 4 shows a comparison between CPD and the common diagnostic strategies. The proposed CPD strategy outperformed the other diagnostic strategies, as it gives a high accuracy and a small error rate. As show in Table 4, CPDS gives poor accuracy with high error. According to Figs.(7→16) and Table 4, the proposed CPD strategy outperform other strategies based on using Data1 or Data 2. It is noted that the results of strategies based on Data1 is better than the results based on Data 2. Thus, blood tests dataset (Data1) can provide more accurate results than CT image dataset (Data 2). Thus, the proposed CPD outperforms other strategies based on using blood tests dataset (Data1).

TABLE 4:
SHOW A COMPARISON BETWEEN CPD AND THE COMMON DIAGNOSTIC STRATEGIES

Techniques	Data 1				Data 2			
	CPD	**Heg.IA**	**CPDS**	**HFG-HFS**	**CPD**	**Heg.IA**	**CPDS**	**HFG-HFS**
Accuracy	96.36%	92.87%	88.98%	84.93%	90.66%	88.35%	80.41%	83.21%
Error	3.63%	7.13%	11.02%	15.07%	9.34%	11.65%	19.59%	16.79%
Sensitivity	100%	92.30%	90.11%	87%	93.15%	92%	86.10%	88.07%
Precision	92.30%	90.05%	89.17%	85%	91%	88.31%	82.4%	89.5%
Run Time	22.62 sec.	23 sec.	27 sec.	30 sec.	35 sec.	39.12 sec.	49.2 sec.	43 sec.

V. CONLUSIONS AND FUTURE RESEARCH

Despite the efforts exerted to decrease the accelerating sawing of Covid-19 as a result of the great risks that this disease causes in all aspects of life, the rapid diagnosis of this disease is the most effective way to limit this spread and the possibility of providing healthcare to the injured and protecting healthy people. In this paper, a new CPD strategy that consists of two stages namely FSS and CDS was introduced. It is important to extract the meaningful features from the used dataset. In the FSS, the most effective subset of features is selected by using CSFS method. Then, an accurate and quick diagnosis is provided by using IKNN. It is noted that, the proposed CPD strategy outperformed other strategies according to both datasets; Data1 and Data 2. The results of CPD based on Data1 (blood tests) is better than its results based on Data 2 (CT images). The evaluation results based on Data1 showed that the CPD strategy outperformed other strategies called Heg.lA, CPDS, and HFG-HFS in which the CPD provided the best accuracy, error, precision, sensitivity, and run time with values equal 0.963, 0.036, 0.93, 1, and 22.62s.

In the future work, the proposed CPD strategy will be developed by combining the CSFS method in the feature selection stage with a wrapper stage that contains the Gray Wolf Optimization (GWO) to select must accurate features. Additionally, the proposed CPD strategy will be tested on large datasets from different areas to measure the scalability of this strategy.

AUTHORS CONTRIBUTION

Alaa M Mohamed is responsible for conception of the work, software, and drafting the article

Ahmed I. Saleh is responsible for project administration and final approval of the version to be published

Doaa A. Altantawy is responsible for data collection, data analysis, and interpretation.

M.A. Abo-Elsoud is responsible for supervision and critical revision of the article.

FUNDING STATEMEN:

The author did not receive any financial support of the research authorship and publication of this article.

DECLARATION OF CONFLICTING INTERESTS STATEMENT:

The author declared that there are no potential conflicts of interest with respect to the research authorship or publication of this article.

REFERENCES

[1] S. Liu, M. Huang, Y. Xu, J. Kang, et. al., "CRISPR/Cas12a Technology Combined with RT-ERA for Rapid and PorTable SARS-CoV-2 Detection," Virologica Sinica, Springer, vol.36, no.5, pp.1087-1083, 2021, https://doi.org/10.1007/s12250-021-00406-7.

[2] N. Mansour, A. Saleh, M. Badawy, H. Ali, "Accurate detection of covid-19 patients based on feature correlated Naïve Bayes (FCNB) classification strategy," Journal of Ambient Intelligence and Humanized Computing, Springer, pp. 1–33, 2021 Jan 15, https://doi.org/10.1007/s12652-020-02883-2.

[3] Z. Li, Y. Yi, X. Luo, N. Xiong, et. al., "Development and clinical application of a rapid IgM-IgG combined antibody test for SARS-CoV-2 infection diagnosis," Journal of medical virology, vol.92, no.9, pp.1518-1524,2020, doi: 10.1002/jmv.25727. http://www.ncbi.nlm.nih.gov/pubmed/32104917.

[4] D. Ferrari, A. Motta, M. Strollo, G. Banfi, and M. Locatelli, "Routine blood tests as a potential diagnostic tool for COVID-19," Clin. Chem. Lab. Med., vol. 58, no. 7, pp. 1095–1099, 2020, https://doi.org/10.1515/cclm-2020-0398.

[5] M. Kubánková et al., "Physical phenotype of blood cells is altered in COVID-19," Biophys. J., vol. 120, no. 14, pp. 2838–2847, 2021.

[6] L. Muhammad, A. Haruna, I. Mohammed, M. Abubakar, B. Badamasi, et. al., "Performance evaluation of classification data mining algorithms on coronary artery disease dataset," 2019 9th International Conference on Computer and Knowledge Engineering (ICCKE), IEEE, PP. 1-5, 2019, https:// doi.org10.1109/ICCKE48569.2019.8964703.

[7] W. Shaban, A. Rabie, A. Saleh, and M. Abo-Elsoud, "Accurate Detection of COVID-19 Patients Based on Distance Biased Naıve Bayes (DBNB) Classification Strategy," Pattern Recognition, Elsevier, vol.119, pp.108110, 2021, https://doi.org/10.1016/j.patcog.2021.108110.

[8] V. A. de Freitas Barbosa et al., "Heg. IA: An intelligent system to support diagnosis of Covid-19 based on blood tests," Res. Biomed. Eng., pp. 1–18, 2021. https://doi.org/10.1007/s42600-020-00112-5.

[9] R. I. Doewes, R. Nair, and T. Sharma, "Diagnosis of COVID-19 through blood sample using ensemble genetic algorithms and machine learning classifier," World J. Eng., 2021.

[10] W. Shaban, A. Rabie, A. Saleh, et al., "A new COVID-19 Patients Detection Strategy (CPDS) based on hybrid feature selection and enhanced KNN classifier," Knowl. Based Syst. Elsevier, vol.205, pp.1–18, 2020, http://dx.doi.org/10.1016/j. knosys.2020.106270.

[11] F. Ozyurt, T. Tuncer, and A.Subasic, "An automated COVID-19 detection based on fused dynamic exemplar pyramid feature extraction and hybrid feature selection using deep learning," Computers in Biology and Medicine, Elsevier, Vol. 132, PP. 1-10, 2021, https://doi.org/10.1016/j.compbiomed.2021.104356.

[12] M. Allam, and M. Nandhini," A Study on Optimization Techniques in Feature Selection for Medical Image Analysis," International Journal on Computer Science and Engineering (IJCSE), Vol. 9, no.3, PP. 75-82, 2017.

[13] Y. Li, "Text feature selection algorithm based on Chi-square rank correlation factorization," Journal of Interdisciplinary Mathematics, Taylor& Francis Group, Vol. 20, no.1, PP.153-160 , 2017, https://doi.org/10.1080/09720502.2016.1259769.

[14] P. Samant and R. Agarwal, "Machine learning techniques for medical diagnosis of diabetes using iris images," Comput. Methods Programs Biomed., vol. 157, pp. 121–128, 2018, https://doi.org/10.1016/j.cmpb.2018.01.004.

[15] Ch.Wang, Y. Long, W. Li, W. Dai, Sh Xie, Y Liu, Y. Zhang, M. Liu, Y. Tian, Qiang , and Y. Duan, "Exploratory study on classification of lung cancer subtypes through a combined K-nearest neighbor classifier in breathomics," Scientific Reports, vol.10, no.1, PP. 1-12, 2020. https://doi.org/10.1038/s41598-020-62803-4

[16] M. jabbar, B. Deekshatulu, and P. Chandra," Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm," Procedia technology, Elsevier, Vol. 10, PP. 85–94, 2013, https://doi.org/10.1016/j.protcy.2013.12.340.

[17] S. Du, and J. Li, "Parallel Processing of Improved KNN Text Classification Algorithm Based on Hadoop, " Proceedings of the 7th International Conference on Information, Communication and Networks, IEEE, PP. 167–170, 2019.

[18] S. B. Rikan, A. S. Azar, A. Ghafari, J. B. Mohasefi, and H. Pirnejad, "COVID-19 diagnosis from routine blood tests using artificial intelligence techniques," Biomed. Signal Process. Control, vol. 72, p. 103263, 2022.

[19] https://github.com/UCSD-AI4H/COVID-CT.

[20] J. Zhao, Y. Zhang, X. He, P. Xie, COVID-CT-Dataset: A CT scan dataset about COVID-19, 2020, arXiv preprint, arXiv: 2003.13865v1.

Title in Arabic:

استرتيجية تشخيص مرضى الكوفيد-19 باستخدام طرق تعدين البيانات

Abstract in Arabic:

كوفيد-19, لا يزال العالم يعيش في حالة من القلق وعدم الاستقرار على الرغم من الجهود المبذولة لإيجاد لقاح والخروج من هذه الأزمة خاصة بظهور فيروس كورونا المتحور الجديد المسم اوميكرون ، والذي أثار حالة من الجدل حول مدى تأثيره و قدرتها على الانتشار بين الناس. لقد ألقى وباء كوفيد-19 بالاقتصاد العالمي في حالة من الفوضى. كما أدى إلى توقف واسع النطاق للعمل والإنتاج في جميع أنحاء المجتمع ، مما أضر بالاقتصاد والمجتمع. تقدم هذه الورقة إستراتيجية تشخيص مرضى كوفيد-19 (CPD) التي تعمل على إيجاد تشخيص سريع وفعال للغاية لتشخيص مرضى كوفيد-19 . تتكون الاستراتيجية المقترحة من مرحلتين رئيسيتين هما مرحلة اختيار الميزات (FSS) ومرحلة التشخيص (CDS) الهدف الرئيسي لـ FSS هو اختيار الميزات القوية لمرحلة التشخيص. يتم تحديد الميزات في FSS باستخدام طريقة Chi-Square Feature Selection (CSFS).في الواقع ، CSFSهي تقنية اختيار ميزة عامل التصفية التي لديها القدرة على اختيار مجموعة فرعية أكثر فعالية من الميزات بسرعة. بعد ذلك ، يتم توفير التشخيص السريع والدقيق باستخدام .(K-Nearest Neighbors (IKNN الفكرة الرئيسية في IKNN المقترحة هي أن دائرة بنصف قطر تساوي متوسط المسافة لعناصر الاقرب وعددهم K سيتم بناؤها ثم سيتم تحديد أقرب M من العناصر لتصنيف المريض إلى الفئة الصحيحة "Covid" أو ."Non-Covid" توضح النتائج أن الإستراتيجية المقترحة المسماة CPD تعطي دقة تصل إلى 96.32٪