Received: August 19, 2020. Revised: November 24, 2020.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

The Role of Machine Learning to Fight COVID-19

Shimaa Ouf[1]* **Noha Hamza[1]**

[1]*Department of Business Information Systems, Faculty of Commerce and Business Administration, Helwan University, Egypt*
* Corresponding author's Email: shimaaouf@yahoo.com

Abstract: Objectives: this paper aimed to first, exploring the relationship between multiple physical measurements. Second, predicting the treatment course of hospitalized patients with COVID-19 disease from physical measurements. Third, investigating the primary symptoms and the average duration of each symptom's disappearance. Fourth, provide the physicians with the prediction model to help them determine the best combination of drugs for Covid-19 patients'. Methods: this paper first, apply correlation analysis on the dataset to identify the relationships between the dataset attributes and selecting model features more efficiently to improve the accuracy results. Second, implements seven machine learning algorithms with four cross-validation techniques to predict the most appropriate treatment course for COVID-19 hospitalized patients. Two treatment courses were identified; course-LR was patients treated with Lopinavir-ritonavir and course LR + AR were patients treated with Lopinavir-ritonavir combined with arbidol for antiviral treatment. Results: by applying correlation analysis between dataset attributes, we found that there is no relationship between the presence of chronic diseases or the patient' age and the Covid-19 clinical classification. The prediction model results show that 10 fold cross-validation with Naïve Bayes and neural network achieving the highest accuracy of 85.71%. Conclusion: this paper has exploited correlation analysis and machine learning based approaches to identify relevant attributes in the COVID-19 patients' dataset and predicting the most appropriate treatment course.

Keywords: Covid-19, Machine learning, Artificial intelligence, Software architecture, Prediction.

1. Introduction

The Covid-19 pandemic has not only caused infections and deaths, but it has also a negative effect on the global economy on a scale not seen since at least the Great Depression. Covid-19 could devastate individual livelihoods, industries, businesses, and entire economies. The recognition of these impacts and analysing their effects on the different industries and economics considers a crucial task for research [1, 2].

In December 2019, pneumonia associated with Covid-19 arised in Wuhan, Hubei Province, China and has caused severe respiratory disease and death in humans [3, 4]. The number of infected cases due to Covid-19 is growing exponentially around the world.

The countries most affected by the Coronavirus are Italy, the USA, Spain, U.K, and France has already exceeded China. Until June 1, 2020, Coronavirus 2019 in China has achieved 83,017 infection cases and 4,634 deaths. It is rapidly increasing and has affected 196 countries and achieved 6000015 infection cases and 372000 deaths outside China. The only effective way until now for prevention has been social distancing and a lockdown of countries in the world. This will lead to badly damage on the global economy [5].

Coronavirus 2019 disseminated through close contact with the infected ones, the surface of objects, airborne, and especially respiratory droplets [6]. The clinical features of Coronavirus contain fatigue, fever, sore throat, cough, headache, shortness of breath, and muscle pain. According to the WHO, for the diagnostic of Covid-19, the most popular test technique used is a real-time reverse transcription-

polymerase chain reaction (RT-PCR). Chest radiological imaging such as X-ray and computed tomography (CT) has a great impact on early detection and treatment of this virus [7, 8].

Richard Baldwin, a Professor of International Economics at the Graduate Institute in Geneva, tells that the Covid-19 is as economically infectious as it is medically infectious [9, 10]. As a result, integrating new technologies like Data Science, Artificial Intelligence, Machine learning, which showed its advanced performance in healthcare sector, is very important to control and prevent this pandemic by providing their technical views and possible solutions to diagnostic the cases accurately and fast to prevent this natural pandemic [11, 12]. The need for creative solutions became very important after Coronavirus to manage and analyse big data on the growing network of infected countries, patient details, community movements, and clinical trials [13, 14]. The automatic tracking of the travel history of infected cases is very important to study epidemiological correlations with the spread of the virus. These technologies used to predict when and where, the virus is likely to spread, and tell those regions to cover the required arrangements [14, 15].

The research papers focused only on introducing the benefits of applying machine learning in the covid-19 pandemic. Authors focused on applying machine learning techniques on the collected data from a social media to quantify Coronavirus content among online contender of establishment health guidance, in particular vaccinations. They used machine learning techniques to detect Coronavirus patient chest X-rays. The authors in this research papers ignored to apply the machine learning techniques to predict the best combination of drugs for covid-19 patients.

This paper focuses on applying the correlation analysis on the hospitalized Covid-19 dataset to select the most important features that maximize the prediction accuracy as shown in Table 2. Then, the most popular data mining classification techniques were applied with the four types of cross validation techniques to determine the best technique which achieve the best accuracy of predicting the best treatment course for each Covid-19 patient. Our proposed prediction system will be used by physicians to help them determining the best and suitable combination of drugs according to each case of covid-19 patients.

This paper discusses the following sections: Section two shows the Research Methodology; Section 3 introduces the Literature Review. Section 4 describes the characteristics of the dataset. Section 5 illustrates the analysis of classification performance. Section 6 represents the correlation analysis method. Section 7 introduces the data analysis and methods to build the prediction models. Section 8 introduces the results. Section 9 addresses the conclusion.

2. Research methodology

The goal of this paper is to explain how data mining techniques are used to fight Covid-19. The most popular scientific databases like Science Direct, Springer, and IEEE were used to find the relevant literature that fulfils the research goals. These research papers form the foundation of the literature review. Only peer-reviewed research papers, conference papers, written in the English language will be selected. The ambition is to use literature published between 2019 and 2020. The research focuses on data mining and its crucial role to fight COV19 and eliminate its spreading across the world. The search terms' purpose is to find a set of results in the domain of data mining in the time of Covid-19, the search terms used are "data mining and Covid-19", "Artificial Intelligence and Covid-19", and "the impact of machine learning on Coronavirus". The papers that were not truly related to data mining and its role to fight the Coronavirus deleted and only the papers that are related to our research are going to be employed. The results of the search process were 548 research papers. These papers published between 2019 and 2020 and then classified by the scientific databases where they appeared.

The choice of the retrieved research papers was determined independently by the authors based a set

Table 1 Inclusion and elimination criteria.

Selection criteria	Scientific database
Inclusion	-Peer-reviewed research articles, conference proceedings papers, book chapters, systematic review papers. -Due to Title: research papers related to the Covid-19, Coronavirus, Artificial Intelligence, Machine Learning, Data Mining technologies -Due to Abstract and conclusion: the research papers related to Covid-19 and Artificial Intelligence, Machine Learning, Data Mining technologies -Due to Full Text: research papers introducing technical aspects of new technologies on Covid 19.
Elimination	Languages: non English research papers were eliminated

of predefined inclusion and elimination criteria as shown in Table 1. The elimination characteristics contain language and the subject area. Firstly, the relevant titles of the retrieved research papers were taken into consideration. Secondly, abstracts of all research papers and conclusion sections were assessed. Research papers, meeting one of the elimination criteria were removed. Subsequently, a full-text review also assessed. Overall, many studies were included because they were focused primarily on the technical aspects of using Artificial Intelligence, Machine Learning, and Data Mining technologies to fight Covid-19. Finally, the research papers which included in our research are 39, Science Direct introduced (17 out of 39 research papers), Springer (14 out of 39 research papers), and IEEE (8 out of 39 research papers).

3. Literature review

In this paper, the authors propose an approach that applies a machine-learning algorithm, the Latent Dirichlet Allocation (LDA algorithm) to handle large quantities of data on Facebook as a social media. This approach identifies special topics within collections of posts from online communities surrounding the vaccine and Coronavirus debate. The collected Facebook public post data cover the period from 1/17/2020 to 2/28/2020. A large quantity of potentially dangerous Coronavirus misinformation exists online. The machine learning techniques were used to quantify Coronavirus content among online contender of establishment health guidance, in particular vaccinations ("anti-VAX"). The authors found that the anti-VAX society is developing a less focused debate around Coronavirus than its counterpart, the pro-vaccination ("pro-VAX") counterpart. As a result, the anti-VAX counterpart displays a range of "flavors" of Coronavirus topics; therefore, they can entreaty a broader cross-section of people seeking Coronavirus guidance online, like people wary of a mandatory fast-tracked Coronavirus vaccine or those seeking alternative remedies. The authors describe a mechanistic model that summarizes these results and could ease in evaluating the likely effectiveness of intervention strategies. The proposed approach is scalable and faces the urgent problem of defacing social media applications of having to analyze large quantities of online health misinformation and disinformation [10].

A positive chest X-ray of infected patients is a pivotal step to fight Coronavirus. The early finding shows that abnormalities exist in chest X-rays of patients who have of Coronavirus. There is a lot of research has confirmed that the accuracy of Coronavirus patient detection by using chest X-rays is strongly optimistic. There is a need for a substantial amount of training data to work on through deep learning networks as convolutional neural networks (CNNs). Because the Covid-19 outbreak is recent, it is complicated to collect many radiographic images in such a short time. The authors develop a model called CovidGAN an Auxiliary Classifier Generative Adversarial Network (ACGAN) to create a synthetic chest X-ray (CXR) images to enlarge the dataset and to enhance the performance of CNN in Coronavirus detection. The result showed that the synthetic images generated from CovidGAN could be used to improve the performance of CNN for Covid-19 detection. The accuracy of classifying images from CNN achieved 85%. By adding synthetic images generated by CovidGAN, the accuracy increased to 95% [16].

The Covid-19 pandemic is provisioning all over the globe. Medical imaging like computed tomography (CT) and X-ray do a crucial role in the global fight against Coronavirus. Integrating artificial intelligence (AI) technologies maximize the power of the imaging tools and support medical specialists. This paper displays a review, which reflects the impact of applying AI to fight Covid-19. AI-empowered image acquisition can obviously reshape the workflow with minimal contact with patients; automate the scanning procedure and introducing the best protection to the imaging technicians. AI can enhance work effectiveness by the precise delineation of infections in CT images and X-ray, simplifying subsequent quantification. AI support radiologists to make clinical decisions for the diagnosis of the disease. This paper analyzes many research papers that present the medical imaging and analysis techniques (image acquisition, segmentation, diagnosis), involved with COVID-19 to provide efficient and accurate imaging solutions in Coronavirus [17].

This paper reflects the impact of applying new technologies to fight COVID-19 crisis. The authors discuss the importance of integrating wearable devices for controlling the people in quarantine and those at risk for checking the health status of management personnel and caregivers, and for simplifying the processes of triage for admission to hospitals. It shows the importance of using unobtrusive sensing systems to discover the disease and observe patients with light symptoms whose clinical situation could unexpectedly aggravate in improvised hospitals [18].

The authors discuss the impact of applying the new technologies such as the blockchain, Artificial

Intelligence (AI), Internet of Things (IoT), 5G, and Unmanned Aerial Vehicles (UAVs) to fight COVID-19 outbreak. This paper covers a comprehensive review of the Coronavirus 2019 to display its clinical features, transmission mechanism, and diagnosis rules [19].

Artificial intelligence (AI) applications have been used to solve different medical problems. The authors preview a systematic review showing the great impact of using AI applications based on machine learning (ML) and data mining algorithms for diagnosing, detecting, and fighting COVID-19. This paper preview this critical virus, discuss the challenges of applying data mining and ML algorithms and present the benefits of this technique to the health sector. Five databases, namely, Science Direct, IEEE Xplore, PubMed, Web of Science, and Scopus were used to achieve the paper's goal. The authors confirm the crucial role of applying ML and data mining techniques in the health sector especially with the COVID-19. AI technologies are used to develop different applications to remotely support analyzing the extracted features and classes of infected cases with COVID-19. The goal of this paper is to help researchers to conduct new researches that can lead communities and governments to early control the spreading of COVID-19 by using the features and classes collected in the literature [20].

The Coronavirus 2019 appears in late December 2019 in China, but it is spreading rapidly all over the world. It is important to make prediction researches on the speed of an epidemic spreading all over the world. Traditional models of epidemic deal with all people with Coronavirus as having the same rate of infection and as a result, they incapable of representing the evolution trend of an epidemic. In this paper, a hybrid artificial intelligence (AI) model was introduced for Coronavirus 2019 to predict the new daily confirmed infected cases at different time intervals. The authors introduce a grouped multi-parameter strategy that puts the rates of the confirmed infected cases in the past into different groups by time. Then uses Natural Language Processing (NLP) technology to analyse and extract new knowledge like control measures of the epidemic and people's awareness of epidemic prevention that are then represented into semantic features. The prediction results of a hybrid artificial-intelligence (AI) model is extremely consistent with real infected cases, which confirm that the introduced architecture analyses the transmission law and development trend of the virus compared with the previous architecture [21].

In this paper, the automatic classification of pulmonary disease problem, containing the newly emerged Coronavirus 2019, from X-ray images, was addressed. Deep Learning is the best method to discover large, high-dimensional features from medical images. The Convolutional Neural Network (CNN) called Mobile Net is used and trained from scratch to prove the significance of the extracted features for the classification task. The dataset that contains 3905 X-ray images are used for training MobileNet v2 that has been achieved excellent results in related tasks. The training results of CNNs from scratch classifying the X-rays between the seven classes and between Covid-19 and non-COVID-19. The accuracy of classification using this method achieved 99.18% accuracy, 99.42% Specificity, and 97.36% Sensitivity in the detection of COVID-19. Besides, rapid, low-cost, and automatic detection of the Coronavirus 2019 disease was accomplished, using a large-scale dataset of pulmonary infections. The detection of Coronavirus 2019 from medical images leads to reducing exposure of nursing and medical staff to the outbreak [22].

AI technology is very important to categorize the disease into different categories of severities. This report introduces Thoracic VCAR software (GE Healthcare, Italy) as an example of a good tool for the radiologist in the COVID-19 diagnosis. Thoracic VCAR presents quantitative measurements of lung involvement. It can create a fast, clear, and summarized report that connects important medical information to referring physicians [23].

Radiographies patterns on CT chest scans have achieved a higher specificity and sensitivity to detect COVID-19 compared to RT-PCR that on the World Healthcare Organization's report and have a small positive detection rate in the early stages. This paper reflects a technical review that analyzed different convolutional neural network (CNN) models to categorize CT samples with Coronavirus 2019 into two categories Influenza viral pneumonia or no-infection. The authors compare the mentioned study with one that is used 2D and 3D deep learning architectures, integrating them with the new clinical understanding, and achieved an Area under the ROC Curve (AUC) of 0.996 (95% CI: 0.989–1.00) for Coronavirus vs. Non-coronavirus cases per thoracic CT studies. They achieved a specificity of 92.2% and a sensitivity of 98.2% [24].

The infected cases of Coronavirus 2019 are increasing rapidly all over the world; more than 1.2 billion infected cases and around 65,000 have died of this disease, until today. This increasing number of infected and died cases and their medical data have generated important sources of information and

knowledge. In this paper, the authors show the importance of storing such a large amount of data, using different data storage technologies. These gathered data represent a crucial role to conduct scientific research and development about the Coronavirus 2019, pandemic, and guidelines to fight this virus and its after-effects. Big data are an innovative technology and solution that can save a large amount of data on these infected cases in a digital format. It supports computational analysis to extract patterns, trends, associations, and differences. The big data can be used gainfully to cut the risk of spreading Coronavirus. It can also help to discover the spreading rate of this virus and take control with detailed data capturing ability [25].

Data mining techniques were applied to discover new patterns from a large amount of infectious disease data. In this paper, the author analyzes the impact of using data mining techniques to categorize real groups of infected cases of COVID-19 data set of different states in India and union territories (UTs) by their high similarity to each other. The data mining techniques' result extracts a set of clusters of affected Indian states. These clusters optimize monitoring techniques in affected states in India, which will be very valuable to the doctors and government to understand the spreading of Coronavirus 2019. This will help government decisions and medical facilities and reduce the number of infected cases. The hierarchical analysis of clusters was performed to display relations using the observations gathered from the three types of cases of Coronavirus 2019 of Indian states and UTs. The cluster analysis categorized 27 states and 5 UTs into six clusters (I–VI) of affected cases with COVID-19 [26].

This paper evaluates the impact of Artificial Intelligence (AI) against COVID-19 to discover new drugs. It is important to gather diagnostic data about the infected cases. This will be important to train AI, save lives, and decrease economic damages. AI can be used to decide the right remediation and vaccines of Coronavirus 2019. AI has a crucial role to manage the pandemic, according to thermal imaging to scan public spaces for infected cases and by enforcing lockdown measures and social distancing [27].

The authors propose a model for COVID-19 disease to categorize the infected cases from chest CT images. The dataset, which stores data about the chest CT of COVID-19-infected cases are divided into training and testing sets. The training set was used for creating the COVID-19 disease classification model. The 20-fold cross-validation technique is used to split the dataset and prevent overfitting. A comparison between the proposed model and the other competitive models prove its effectiveness is done. The experimental results show that the proposed model performs better than competitive models like ANFIS, ANN, and CNN models in terms of accuracy, specificity, sensitivity, F-measure, and Kappa statistics by 1.9789%, 1.6827%, 1.8262%, 2.0928%, and 1.9276% respectively [28].

The authors propose and develop a model-based deep learning technology to enhance the accuracy of documented cases and to predict accurately the disease from the chest X-ray scans. The proposed model was based on convolutional neural networks (CNNs) to discover structural abnormalities and disease categorization that are important to extract the hidden patterns and eliminates manual interaction dependent on radiologists. The authors test their model by using a public dataset including 181 COVID-19 patients from different countries worldwide, which contains the infected cases (115 used for training and 66 for testing) to arrive at a conclusion regarding the accuracy of the proposed method. The results show very high accuracy (96.3%) and the proposed model proves its effectiveness to represent a simpler, faster, and more accurate method for Coronavirus 2019 detection [29].

Artificial intelligence technology plays a helpful role for successful analyses of COVID-19. In this paper, COVID-19 is discovered using deep learning technology. The model was verified using a dataset that includes three classes, namely: coronavirus, pneumonia, and normal X-ray imagery. The Fuzzy Color technique is applied to the dataset as a preprocessing step. The stacked dataset is trained using deep learning techniques (SqueezeNet, MobileNetV2) and the Social Mimic optimization method is used to process the feature sets obtained by the techniques. Then, efficient features are integrated and classified using Support Vector Machines (SVM). The proposed approach achieves 99.27% of the classification rate, which means the efficient contribution to the detection of coronavirus 2019 disease. The infected cases of COVID-19 were suffering from permanent damage in the lungs, which can later cause death. The proposed approach aims to recognize people with damaged lungs because of COVID-19 from pneumonia or normal people (not have COVID-19) [30].

Drug discovery is not an easy task and Artificial Intelligence technology, especially deep learning can help speed this process by facilitating predicts which existing drugs or brand-new drug-like molecules could handle Coronavirus 2019. Deep learning can help distribute important information across the

world and cut the spread of false information about Coronavirus 2019. The chatbots can be widely distributed to educate people across the world to reduce panic and cut the spread of false information about COVID-19. These chatbots have been created in platforms of popular messaging that young people already prefer, like Facebook Messenger. The people who extradited false information on these platforms can be redirected to an approved chatbot to extradite valid and effective information. AI could introduce unique aid and solutions to this deadly pandemic [7].

The spread of Coronavirus disease 2019 (COVID-19) all over the world has threatened humanity, industry, and the global economy. This paper discusses a systematic literature review, which explains the impact of the fourth industrial revolution or Industry 4.0 to fulfil customized requirements during the Coronavirus 2019 crisis. The search processes were applied to the most popular scientific databases like PubMed and SCOPUS. Industry 4.0 could help the diagnosis and detection of coronavirus 2019. Industry 4.0 can accomplished the requirements of customized gloves, face masks, and gather information for healthcare systems, which control and treat of COVID-19 patients. Industry 4.0 is helpful to achieve day-to-day updates of infected cases, age-wise, area-wise, and state-wise with appropriate surveillance systems. It also helps to achieve an innovative way for the appropriate isolation of the infected cases to speeding up drug manufacturing; minimize the high risk of mortality, and treatment process and care. It develops a virtual clinic during the telemedicine consultation application and helps to reduce the physical crowding of the infected cases in the clinics and hospitals [1].

From reviewing the research papers of COVID-19 and data mining, we find that there is lacks in the area of developing prediction models that can help physicians determine the most appropriate course of treatment. In addition to none of these studies used correlation analysis in feature extraction and selection that improve the prediction models' accuracy.

4. Dataset

A data set of hospitalized patients with COVID-19 was collected from 21 February until 18 March 2020 at Lishui Central Hospital, Zhejiang, China and Wuhan 4th hospital, Hubei, China, which is a designated hospital for COVID-19. All patients were diagnosed and treated using the diagnostic and treatment protocol provided by the General Office of the National Health Commission for novel Coronavirus pneumonia.

Data set information recorded included demographic data, clinical diagnosis, medical history, laboratory values (i.e., hematological parameters, biochemical parameters), virus time negative; chest computed tomographic (CT) scans, and hospital stay care steps (i.e., antiviral therapy, corticosteroid therapy, immunosuppressive therapy). The findings of the laboratory tests were divided by day 0-2, 3-5, 6-9, 10-14, and > 15 admission times.

The primary outcome was the recovery rate, as the discharge criteria were met by hospitalised patients with COVID-19. The conditions for clearance is as follows: (1) the body temperature is back to normal for over 3 days; (2) evident improvement of breathing symptoms; (3) strong absorption of inflammation by pulmonary imagery; (4) negative nuclei-acid examination, two consecutive periods, of the respiratory tract, such as sputum and nasopharyngeal swabs (at least 24 hours).

The secondary main outcome were hospital stay, the rate of COVID-19 RNA clearance, the time of nucleic acid turning negative and symptoms disappeared upon entry. The outcome and duration of the SARS-COV-2 nucleic acid test for respiratory samples was the subject of the final test. If COVID-19 nucleic acid was negative for two consecutive tests, the first time the nucleic acid turned negative was used for the test.

5. Methods

Fig. 1 provides the steps accomplished to build the prediction models on the COVID – 19 dataset to predict the best course of treatment from physical measurements that make patient's health better. These steps have included, data pre-processing, model training and validation, and classification.

Step 1: Data pre-processing

In this paper, the predictive models built using python. Fig. 1 shows the flow of applied experimental methods. In the experiment, the dataset is previewed on the Spyder platform.

Step 2: Correlation Analysis Methodology

Correlation analysis is an extensively used technique that identifies interesting relationships in data. This paper has exploited correlation analysis in the process of feature extraction and selection for the prediction model to identify relevant and most important attributes in the dataset in order to improve the accuracy of the model and make the model learn based on relevant features. These relationships help us in predicting the treatment course from physical measurements which have a significant impact on

Table 2. Model features description

Attribute name	Description
Group	Treatment course. Two courses are available: **1-** Group **LR+*Ar*** patients were treated with ***lopinavir-ritonavir*** combined with **arbidol** for antiviral therapy. **2-** Group ***Lr*** patients were treated with ***lopinavir-ritonavir*** only.
Course of lopinavir-ritonavir (LR, 400 mg and 100mg)	Dosage of lopinavir-ritonavir
Course of arbidol (Ar, 200mg)	Dosage of arbidol
Gender	Sex (1 = male; 0 = female)
Age Year	Age by years
Clinical Classification	Patient clinical classification(0 = ordinary, 1 = heavy)
Hospital admission days	Days from onset of symptoms to hospital admission
No comorbidity	1 = no comorbitdity, 0= comorbitdity
cardiovascular/cerebrovas	1= yes, 0 = No
Endocrine system disease	1= yes, 0 = No
Malignant tumor	1= yes, 0 = No
Respiratory system disease	1= yes, 0 = No
Digestive system disease	1= yes, 0 = No
Renal disease	1= yes, 0 = No
Liver disease	1= yes, 0 = No
Fever Symptom	1= yes, 0 = No
Fever_Disappearance time	Number of day patient has fever symptom
Cough	1= yes, 0 = No
Cough_Disappearance time	Number of day patient has Cough symptom
Chest tightness	1= yes, 0 = No
Chest_Disappearance time	Number of day patient has Chest tightness symptom
Fatigue	1= yes, 0 = No
Fatigue_Disappearance time	Number of day patient has Fatigue symptom
Diarrha	1= yes, 0 = No
Diarrha_Disappearance time	Number of day patient has Diarrha symptom
Others	1= yes, 0 = No
others_Disappearance time	Number of day patient has Others symptoms
Clinical outcome	Aggravated,turn to ICU, Improved and discharge, Died
Hospital stay	Number of days the patient stayed in hospital
COVID-19 RNA clearance	Clearance of COVID-19 RNA (1= yes, 0 = No)
The time of nucleic acid turning negative	Number of days nucleic acid turning to negative
Chest CT findings at discharge	Advances, Absorption, No change
Days after hospitalization: The laboratory test results were grouped by admission time on the day 0-2, 3-5, 6-9, 10-14, and >15. • White Blood Cell Count • Neutrophil count • Lymphocyte count • Monocyte count • CRP: **C-Reactive Protein** Test • PCT: Procalcitonin Test	

classifying a COVID-19 patient's treatment group. In mental health situations, correlation analysis has been performed in Python, which involves a dataset of COVID-19 patients' demographic data, signs and symptoms, medical history, laboratory values, time of virus-negative, anti-virus treatment and chest computed tomographic (CT) scans. Pearson's product-moment correlation and other different classification algorithms have been utilized for this analysis.

Correlation analysis determines the strength of a relationship between two item sets, which can be a dependent and an independent variable or even two independent variables. Numerically, this relationship is usually determined by a decimal value, known as the correlation coefficient.

The correlation coefficient determined under coefficient determined under a certain predefined range (depending on the algorithm) [34]. Based on the value of the coefficient in the given range, its strength and direction can be determined. The coefficient has a positive sign indicates that the two variables are positively correlated, because negative

sign indicates a negative correlation. The higher number of coefficients indicates that the two variables have a strong correlation and the lower value indicates otherwise. This relationship helps to identify which independent variables can have the strongest impacts on the dependent variables, and therefore leads to predict the outcome of a dependent variable more efficiently. Having a strong relationship, the independent variable can be considered a strong predictor of the dependent variable. Fig. 2 shows the correlation matrix for all dataset attributes. Fig. 3 shows the most 15 attributes correlated with the treatment course dependent variable that are represent the model features described in details in Table 2.

The product moment correlation was first proposed by Francis Galton in 1880s, then later modified by Karl Pearson in 1896, and has since been known as the Pearson product-moment correlation coefficient. It is a statistical analysis of the collinear relationship between two variables. It involves the ratio of covariance and the standard deviation of the data values between two given variables as shown in table 3. Consider two variables A and B. Then the Pearson's correlation coefficient can be calculated using the following formula: [34].

$$C_{A,B} = \frac{\text{Covariance (A,B)}}{\sigma_A \sigma_B} \quad (1)$$

Where $C_{A,B}$ is the correlation coefficient, Covariance (A, B) is the covariance, and σA a nd σB are the standard deviations of A and B, respectively. In the case of a dataset involving two set, {a1, a2,.., an} and {b1, b2, …, bn}, the correlation coefficient can be calculated as: [35]

$$c = \frac{\sum_{i=1}^{n}(a_i-\bar{a})\,(b_i-\bar{b})}{\sqrt{\sum_{i=1}^{n}(a_i-\bar{a})^2}\sqrt{\sum_{i=1}^{n}(b_i-\bar{b})^2}} \quad (2)$$

where n is the sample size, ai and bi are the ith data values, and a, b are the mean values. The value of the coefficient (C) ranges between −1 and +1. Values close to +1 shows a strong positive correlation, those close to −1 show strong negative correlation, and those closest to 0 shows no relation.

Table 3. Shows the pearson's correlation coefficient value for a set of variables

Variable 1	**Variable 2**	**Pearson's correlation coefficient value**	**Comment**
No comorbidity	Clinical classification	0.22	There is no relationship between the presence of chronic diseases and the Covid-19 clinical classification.
Age year	Clinical classification	0.37	There is no strong relationship between the patient' age and the Covid-19 clinical classification.
Chest tightness	Clinical classification	0.50	There is positive a relationship between the chest tightness 'CT scan' result and the Covid-19 clinical classification.
Age year	The time of nucleic acid turning negative	0.036	There is no relationship between patient' age and the recovery time.
Age year	COVID-19 RNA clearance	-0.1	
Course of lopinavir-ritonavir	The time of nucleic acid turning negative	0.6	There is a positive relationship between the treatment course of lopinavir-ritonavir and the time of nucleic acid turning negative

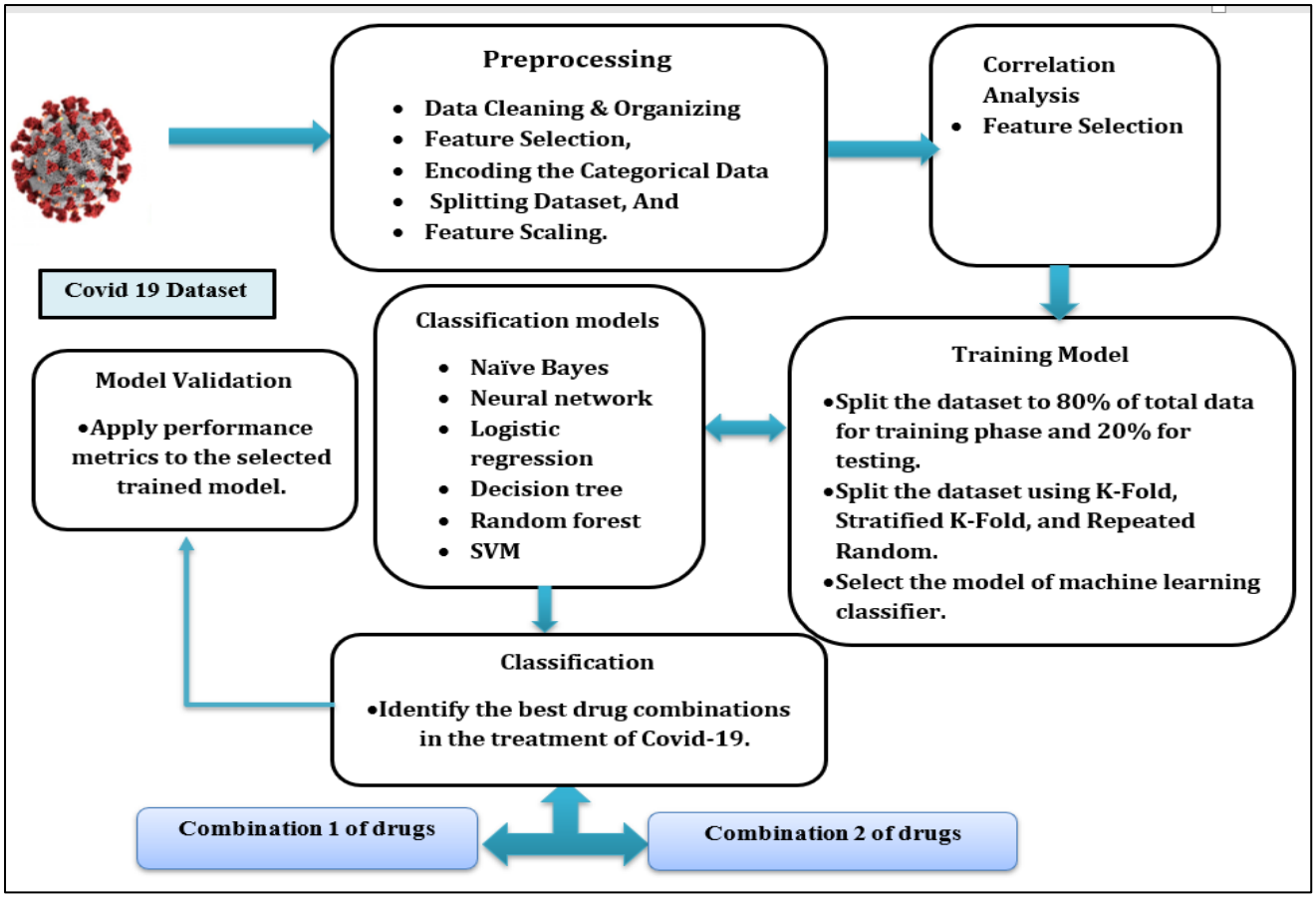

Figure. 1 Workflow to classify the best combination of drugs in the treatment of COVID-19

Step 3: Model training and validation

The four types of cross-validation techniques (hold out, k-fold, stratified k-fold cross-validation, and repeated random) that divide a dataset into training and test set used. The COVID-19 dataset is separated into two sections, section one represents a training set and section two represents the test set. Choosing the ratio of data for the training and testing section is very important to run cross-validation techniques. Holdouts cross-validation: the COVID-19 dataset is separated into two sections; training and test set. We train the first section of 80% of the data in the dataset and test the other 20% of the data [30]. The K-fold Cross-validation: the COVID-19 dataset is separated into k-folds. We train the model on k-1 sections and make testing on the other section. The processing still works, we make testing of all the k parts; k times were repeated while changing the test section one-by-one. In this research, we assign 10 values for K. Each model is trained and tested 10 times. The major advantage of this method is minimizing the bias associated with random sampling [21]. Stratified k-fold: the data is separated into folds to guarantee that each fold owns the same proportion of observations with a given categorical value [31]. A hybrid technique which combines holdout validation and k-fold cross-validation is called repeated k-fold cross-validation. This technique works as follows: it splits the data randomly to create the training-test set and then the process of splitting is repeated and the results were evaluated each time [32].

Step 4: Classification

The seven classification techniques for predicting the best drug combinations in the treatment of Covid-19 were applied. These techniques include Logistic regression, Support Vector Model, KNN, Decision Tree, Naïve Bayes, Random Forest, and Neural Network. The evaluation of running each classification technique, using the selected cross-validation technique is recorded.

6. Classification performance analysis

Table 4. Confusion matrix

	Predicted Positive	Predicted Positive
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

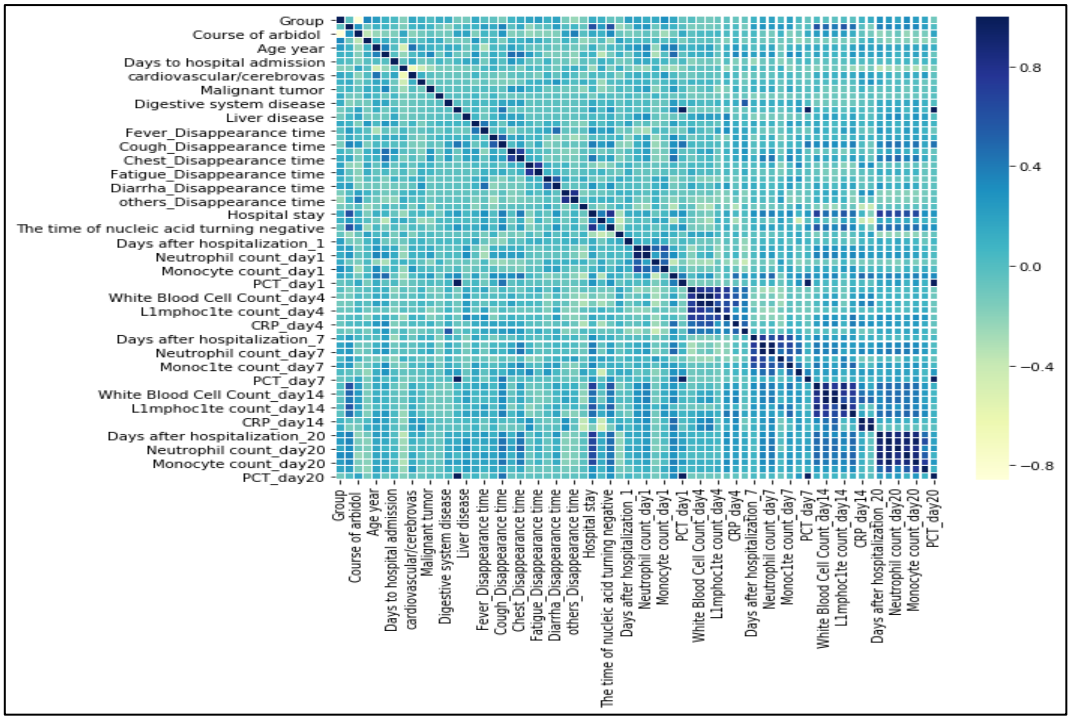

Figure. 2 Dataset grid correlation matrix

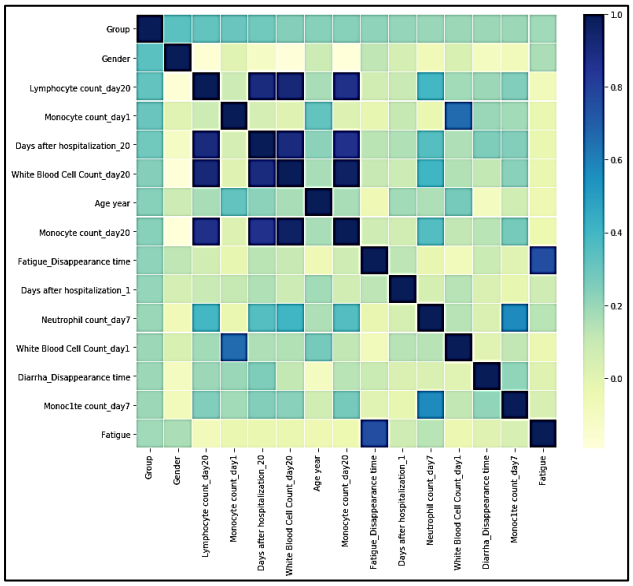

Figure. 3 Correlation for treatment group

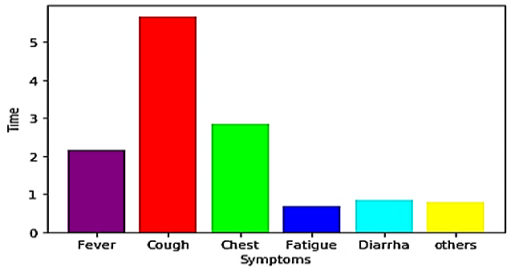

Figure. 4 COVID-19 symptoms disappearance

There are different metrics were applied to evaluate the performance of each classification model. The confusion matrix was used as an important metric, it represents four expected values: True Positive (TP) was used to find the right diagnosis. True Negative was used to find the right diagnosis. True Negative (TN) was used to represent the incorrect number of regular records. False Positive (FP) was used to introduce a set of regular records that are classified as an anomaly diagnosis. False Negative (FN) is a collection of abnormal observed as an ordinary diagnosis as shown in Table 4 [33].

In the confusion matrix, after the values of possible outcomes calculated, the following evaluation metrics are calculated.

Accuracy: it is the most popular and important metric for the results of the classification techniques. It is calculated as the addition of true positives and true negatives divided by the total values of confusion matrix components. The highest accuracy model is the best. It is represented in (3)

$$Accuracy(\%) = \frac{TP+TN}{TP+FP+FN+TN}\ 100\% \quad (3)$$

Precision: give the relationship between the predicted values of true positive and the predicted values of full positive. It is represented in (4)

$$Percision = \frac{TP}{TP+FP} \quad (4)$$

Sensitivity or Recall: represents the ratio between the prediction true positive values and the addition of true positive values of prediction and predicted false

negative values. It is represented in (5)

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

F1-Measure: it is used to integrate precision and recall for measuring of the accuracy of models. F1-measure is the double of the division of the multiplication to the summation of precision and recall metrics. It is represented in (6)

$$F1 - score = 2\frac{Precision*Recall}{Precision+Recall} \quad (6)$$

7. COVID-19 symptoms analysis

Since the outbreak of (COVID-19) disease in December 2019, different digestive symptoms have been commonly reported in patients infected with this disease. In this section we investigate the digestive symptoms of COVID-19 patients and the disappearance period of each symptom. As recognizing the symptoms and period of each one helps in understanding how this virus spreads and helps to bring awareness to most common symptoms to enable earlier COVID-19 diagnosis and thus provide earlier treatment before mild disease progresses to severe disease. The primary symptoms of COVID-19 infected patients are fever, cough, chest tightness, fatigue, and diarrhea. The average time from infection to onset of symptoms is five days. Fig. 4 shows the average duration of symptom disappearance. Cough is the last symptom disappears, then the chest tightness and the fever. Fatigue and diarrhea are the fastest symptoms disappear.

8. Results

Of those seventy-three hospitalized sufferers with Coronavirus 2019 who was discharged from February eight to March thirteen, 2020. The biggest wide variety of sufferers (60) was admitted to Wuhan fourth hospital. The relaxation of sufferers has been from Lishui Central Hospital. The suggest age of group LR+Ar range from twenty-one to eighty-one years, and 33.3% had been women. The suggest age of group LR range from 30 to 87 years, and 67.6% had been women. In the group LR+Ar, the ratio of normal and heavy kind of Coronavirus 2019 became 29.2% and 71.8%, respectively. The ratio of normal and heavy kind of Coronavirus 2019 became 38.2% and 61.8% in the group LR. The variations of sex ratio and ages between group LR+Ar and group LR sufferers had statistically significant. Number of sufferers in all stage of ages has no statistical difference. Some sufferers had one or more coexisting medical conditions. Hypertension, cardiovascular and diabetes had been the popular comorbidities. 38.5% sufferers had comorbidities in the group LR+Ar, cardiovascular and cerebrovascular disease 20.5% and endocrine system disease 10.3%. 55.9% sufferers had comorbidities in the group-LR, cardiovascular and cerebrovascular diseases 35.3% and endocrine system diseases 17.6%. The most popular symptoms and signs had been coughed 69.2%, fever 76.9%, and chest tightnes 25.6% in group LR+Ar. The most popular symptoms and signs had been fevered 82.4%, cough 73.5%, and chest tightness 32.4% in group LR+Ar. Less popular signs had been fatigue, diarrhea, headache, and sore throat [36].

Table 5. Naïve bayes with four cross validation techniques

Technique / Naïve Bayes	**Holdout**	**10 Fold-cross Validation**	**Stratified 10-fold Cross-Validation**	**Repeated Random**
Average macro Precision	54 %	98 %	88 %	58 %
Average macro Recall	53 %	86 %	83 %	50 %
Average macro F-Measure	53.50 %	91.61 %	85.43%	53.70 %
Average macro Accuracy	46.67 %	85.71%	83.33%	50%

As mentioned in Table 2, we determine the most important features which affect the results of prediction and maximize the accuracy. The seven data mining, classification techniques with the four types of cross-validation were applied on the COVID – 19 dataset. The results, approving that the best and ideal technique, which has a great performance, is **K Fold-cross Validation with neural network and naïve Bayes.** The precision, recall, f-measure, and accuracy measures used to evaluate the performance of each technique as shown in Tables 5-11. The Physician can use this prediction system to determine the best combinations of drugs for each patient.

The Naive Bayes data mining technique with the four cross validation techniques for classifying the best combinations of drugs for the covid-19 patients were applied on the COVID-19 hospitalized patients' dataset. The k- fold cross-validation, and stratified **K Fold-cross Validation** with Naïve Bayes technique achieve the highest accuracy of **85.71**%, and 83.33%

Table 6. neural network with the four cross validation techniques.

Technique /Neural network	**Holdout**	**10Fold-cross Validation**	**Stratified 10-fold Cross-Validation**	**Repeated Random**
Average macro Precision	%75	98%	88 %	74%
Average macro Recall	%72	86 %	83 %	73%
Average macro F-Measure	73.47 %	91.61%	85.43%	73.50%
Average macro Accuracy	73.33 %	85.71%	83.33%	%72. 73

Table 7. Logistic regression with the four cross validation techniques

Technique / Logistic regression	**Holdout**	**10 Fold-cross Validation**	**Stratified 10-fold Cross-Validation**	**Repeated Random**
Average macro Precision	61 %	98 %	88 %	52 %
Average macro Recall	60 %	71 %	83 %	50 %
Average macroF-Measure	60.50 %	82.34 %	85.43 %	50.98%
Average macro Accuracy	60%	71.43%	83.33%	50%

as shown in Table 5.

The neural network data mining technique with the four cross validation techniques for classifying the best combinations of drugs for the covid-19 patients are applied on the COVID-19 hospitalized patients' dataset. The k- fold cross validation and stratified k-fold with neural network technique achieve the highest accuracy of **85.71**% and 83.33% as shown in Table 6.

The **Logistic regression** data mining technique with the four cross validation techniques for classifying the best combinations of drugs for the covid-19 patients were applied on the COVID-19 hospitalized patients dataset. The stratified k- fold and k- fold cross validation achieve the highest accuracy of 83.33% and 71.43% as shown in Table 7.

Table 8. SVM with the four cross validation techniques

Technique / SVM	**Holdout**	**10 Fold-cross Validation**	**Stratified 10-fold Cross-Validation**	**Repeated Random**
Average macro Precision	72 %	89 %	88 %	74 %
Average macro Recall	67 %	71 %	83 %	73 %
Average macroF-Measure	69.41 %	78.99 %	85.43 %	73.97 %
Average macro Accuracy	66.67 %	71.43 %	83.3 %	72.73%

Table 9. KNN with the four cross validation techniques

Technique / KNN	**Holdout**	**10 Fold-cross Validation**	**Stratified 10-fold Cross-Validation**	**Repeated Random**
Average macro Precision	52%	73 %	58%	64%
Average macro Recall	47%	57 %	61%	75%
Average macroF-Measure	49.37 %	64.02 %	59.46%	69.06%
Average macro Accuracy	46.67 %	57.14 %	58.33%	63.64%

The SVM data mining technique with the four cross validation techniques for classifying the best combinations of drugs for the covid-19 patients were applied on the COVID-19 dataset. The stratified k-fold and repeated random achieve the highest accuracy of 83.3% and 72.73% as shown in Table 8.

The KNN data mining technique with the four cross validation techniques for classifying the best combinations of drugs for the covid-19 patients were applied on the COVID-19 dataset. The repeated random and stratified k- fold achieve the highest accuracy of 63.64% and 58.33% as shown in Table 9.

The random forest data mining technique with the four cross validation techniques for classifying the best combinations of drugs for the covid-19 patients were applied on the COVID-19 hospitalized patients' dataset. The stratified k- fold and holdout achieve the highest accuracy of 83.34% and 66.66% as shown in

Table 10. random forest with the four cross validation techniques

Technique / Random forest	**Holdout**	**10 Fold-cross Validation**	**Stratified 10-fold Cross-Validation**	**Repeated Random**
Average macro Precision	67 %	60%	88 %	59%
Average macro Recall	67 %	43%	83 %	59%
Average macro F-Measure	67%	50.1%	83 %	59%
Average Accuracy	66.66 %	42.86%	83.34 %	59%

Table 11. Decision tree with the four cross validation techniques

Technique / Decision tree	**Holdout**	**10 Fold-cross Validation**	**Stratified 10-fold Cross-Validation**	**Repeated Random**
Average macro Precision	68 %	80 %	88 %	64 %
Average macro Recall	67 %	57 %	83 %	59 %
Average macro F-Measure	67.50 %	66.57%	85.43%	61.40 %
Average macro Accuracy	66.67 %	57.14%	83.33%	59.09%

Table 10.

The **Decision tree** data mining technique with the four cross validation techniques for classifying the best combinations of drugs for the covid-19 patients were applied on the COVID-19 hospitalized patients' dataset. The stratified k- fold and holdout achieve the highest accuracy of 83.33% and 66.67% as shown in Table 11.

9. Conclusion

This paper has exploited correlation analysis in the process of feature extraction and selection for the prediction model to identify relevant and most important attributes in the dataset in order to improve the accuracy of the machine learning models. In this paper, we have been applying the four cross-validation techniques with the most popular classification techniques to predict the best treatment for the patients of COVID-19 based on the symptoms of each case. This section epitomizes the results of evaluating the generated prediction models. The results show that 10 fold cross-validation with Naïve Bayes achieving the highest accuracy and neural network of 85.71%.

The correlation analysis shows first, that there is no relationship between the presence of chronic diseases or the patient' age and the Covid-19 clinical classification. Second, there is positive a relationship between the chest tightness 'CT scan' result and the Covid-19 clinical classification. Third, there is a positive relationship between the treatment course of lopinavir-ritonavir and the time of nucleic acid turning negative. Fourth, There is no relationship between patient' age and the recovery time. The dataset set shows that the primary symptoms of COVID-19 infected patients are fever, cough, chest tightness, fatigue, and diarrhea. The average time from infection to onset of symptoms is five days. Cough is the last symptom disappears, then the chest tightness and the fever. Fatigue and diarrhea are the fastest symptoms disappear. There are multiple relationships in the dataset need to be analyzed. Future work will focus in analyzing the relationship between the COVID-19 symptoms and the chronic diseases in addition to studying the laboratory test results grouped by admission time and chronic diseases.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Abstract, introduction, research methodology, literature review, methods, Classification Performance Analysis, results, and conclusions was made by Dr. Shimaa Ouf. Dataset and Correlation Analysis Methodology was made by Noha Hamza.

Acknowledgments

This work was not supported by any organization.

References

[1] M. Javaid, A. Haleem, R. Vaishya, S. Bahl, R. Suman, and A. Vaish, "Industry 4.0 technologies and their applications in fighting COVID-19 pandemic", *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, Vol. 14, No. 4, pp. 419-422, 2020.

[2] T. Laing, "The economic impact of the Coronavirus 2019 (Covid-2019): Implications for the mining industry". *The Extractive Industries and Society*, Vol. 7, No. 2, pp. 580-582, 2020.

[3] T. Chakraborty and I. Ghosh, "Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis", *Chaos, Solitons & Fractals*, Vol. 135, 109850, 2020.

[4] S. Hanumanthu, "Role of Intelligent Computing in COVID-19 Prognosis: A State-of-the-Art Review", *Chaos, Solitons & Fractals*, Vol. 138, 109947, 2020.

[5] H. Panwar, P. Gupta, M. Siddiqui, R. Morales-Menendez, and V. Singh, "Application of Deep Learning for Fast Detection of COVID-19 in X-Rays using nCOVnet", *Chaos, Solitons & Fractals*, Vol., 109944, 2020.

[6] M. Haghani, M. Bliemer, F. Goerlandt, and J. Li, "The scientific literature on Coronaviruses, COVID-19 and its associated safety-related research dimensions: A scientometric analysis and scoping review", *Safety Science*, Vol. 129, 104806, 2020.

[7] A. Ahuja, V. Reddy, and O. Marques, "Artificial Intelligence and COVID-19: A Multidisciplinary Approach", *Integrative Medicine Research*, Vol. 9, No. 3, 2020.

[8] L. Bai, D.Yang, X. Wang, L. Tong, X. Zhu, C. Bai, and C. Powell, "Chinese experts' consensus on the Internet of Things-aided diagnosis and treatment of coronavirus disease 2019", *Clinical eHealth*, Vol. 3, pp. 7-15, 2020.

[9] A. Ardakani, A. Kanafi, U. Acharya, N. Khadem, and A. Mohammadi, "Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: Results of 10 convolutional neural networks", Computers in biology and medicine, Vol. 3, pp. 7-15, 2020.

[10] D. Dong, Z. Tang, S. Wang, H. Hui, L. Gong, Y. Lu, F. Yang, "The role of imaging in the detection and management of COVID-19: a review", *IEEE Reviews in Biomedical Engineering*, doi:10.1109/RBME.2020.2990959, In Press.

[11] Y. Ke, T. Peng, T. Yeh, W. Huang, S. Chang, S. Wu, J. Song, "Artificial intelligence approach fighting COVID-19 with repurposing drugs", *Biomedical Journal, In Press*, 2020.

[12] I. Apostolopoulos, D. Ioannis, I. Sokratis, I. Aznaouridis, and MA. Tzani. "Extracting possibly representative COVID-19 Biomarkers from X-Ray images with Deep Learning approach and image data related to Pulmonary Diseases", *Journal of Medical and Biological Engineering*, Vol. 40, pp. 462–469, 2020.

[13] T. Ozturk, M. Talo, E. Yildirim, U. Baloglu, O. Yildirim, and U. Acharya, "Automated detection of COVID-19 cases using deep neural networks with X-ray images", *Computers in Biology and Medicine*, Vol. 121, 2020.

[14] X. Wu, H. Hui, M. Niu, L. Li, L. Wang, B. He, and J. Tian, "Deep learning-based multi-view fusion model for screening 2019 novel coronavirus pneumonia: a multicentre study", *European Journal of Radiology*, Vol. 128, 2020.

[15] S. Tuli, S. Tuli, R. Tuli, and Gill, S. S. (2020). "Predicting the Growth and Trend of COVID-19 Pandemic using Machine Learning and Cloud Computing". *Internet of Things, Volume 11*, 2020, 100222.

[16] A. Waheed, M. Goyal, D. Gupta, A. Khanna, F. Al-Turjman, and P. Pinheiro, "Covidgan: Data augmentation using auxiliary classifier GAN for improved Covid-19 detection", *IEEE Access*, Vol. 8, pp. 91916-91923, 2020.

[17] F. Shi, J. Wang, J. Shi, Z. Wu, Q. Wang, Z. Tang, and D. Shen, "Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for Covid-19", *IEEE Reviews in Biomedical Engineering*, pp.1-1, 2987975, 2020.

[18] X. Ding, D. Clifton, j. Nan, N. Lovell, P. Bonato, W. Chen, and X. Long, "Wearable Sensing and Telehealth Technology with Potential Applications in the Coronavirus Pandemic". *IEEE Reviews in Biomedical Engineering*, pp. 1-1, 2992838, 2020.

[19] V. Chamola, V. Hassija, V. Gupta, and M. Guizani, "A Comprehensive Review of the COVID-19 Pandemic and the Role of IoT, Drones, AI, Blockchain, and 5G in Managing its Impact", *IEEE Access*, Vol. 8, 90225-90265, 2020.

[20] A. Albahri, and R. Hamid, "Role of biological Data Mining and Machine Learning Techniques in Detecting and Diagnosing the Novel Coronavirus (COVID-19): A Systematic Review", *Journal of Medical Systems*, Vol. 44, No. 7, 2020.

[21] N. Zheng, S. Du, J. Wang, H. Zhang, W. Cui, Z. Kang, and H. Long, "Predicting Covid-19 in china using hybrid AI model". *IEEE Transactions on Cybernetics*. Vol. 50, No. 7, 2020.

[22] I. Apostolopoulos, S. Aznaouridis, and M. Tzani, "Extracting possibly representative COVID-19 Biomarkers from X-Ray images with Deep

Learning approach and image data related to Pulmonary Diseases". *Journal of Medical and Biological Engineering*, Vol. 40, pp. 462–469, 2020.

[23] M. Belfiore, F. Urraro, R. Grassi, G. Giacobbe, Patelli, S. Cappabianca, and A. Reginelli, "Artificial intelligence to codify lung CT in Covid-19 patients", *La radiologia medica*, Vol. 125 pp. 500–504, 2020.

[24] C. Butt, J. Gill, D. Chun, and B. Babu, "Deep learning system to screen coronavirus disease 2019 pneumonia", *Applied Intelligence*, 2020.

[25] A. Haleem, M. Javaid, I. Khan, and R. Vaishya, "Significant Applications of Big Data in COVID-19 Pandemic". *Indian Journal of Orthopaedics*, Vol. 54, pp. 526-528, 2020.

[26] S. Kumar, "Monitoring Novel Corona Virus (COVID-19) Infections in India by Cluster Analysis". *Annals of Data Science, Annals of Data Science*, Vol. 7, No. 3, pp. 417–425, 2020.

[27] W. Naudé, "Artificial intelligence vs COVID-19: limitations, constraints and pitfalls", *AI & Society*, Vol. 35, pp. 761–765, 2020.

[28] D. Singh, V. Kumar, and M. Kaur, "Classification of COVID-19 patients from chest CT images using multi-objective differential evolution–based convolutional neural networks", *European Journal of Clinical Microbiology & Infectious Diseases*, Vol. 35, pp.761–765, 2020.

[29] S. Vaid, R. Kalantar, and M. Bhandari, "Deep learning COVID-19 detection bias: accuracy through artificial intelligence", *International Orthopaedics*, Vol. 44, pp.1539–154, 2020.

[30] R. Gupta, A. Ghosh, A. Singh, and A. Misra, "Clinical considerations for patients with diabetes in times of COVID-19 epidemic", *Diabetes & metabolic syndrome*, Vol. 14, No. 3, pp. 211–212, 2020.

[31] M. Lyons, D. Keith, S. Phinn, T. Mason, and J. Elith, "A comparison of resampling methods for remote sensing classification and accuracy assessment", *Remote Sensing of Environment*, Vol. 208, pp. 145-153, 2020.

[32] T. Wong, and P. Yeh, "Reliable accuracy estimates from k-fold cross validation", *IEEE Transactions on Knowledge and Data Engineering*, 2019.

[33] C. Oliveira, P. Niccolai, A. Ortiz, S. Sheth, E. Shapiro, L. Niccolai, and C. Brandt, "Development and Validation of a Natural Language Processing Algorithm for Surveillance of Cervical and Anal Cancer and Precancer: A Split-validation Study", *JMIR Medical Informatics*, 32469840, 2020.

[34] V. Levkivskyi, N. Lobanchykova, and D. Marchuk, "Research of algorithms of Data Mining", *E3SWC*, Vol. 166, pp. 05007, 2020.

[35] J. Ahn, M. Park, H. Lee, S. Ahn, S. Ji, K. Song, and B. Son, "Covariance effect analysis of similarity measurement methods for early construction cost estimation using case-based reasoning. Automation in Construction", *Automation in Construction*, Vol. 81, pp. 254-266, 2020.

[36] X. Lan, C. Shao, X. Zeng, Z. Wu, and Y. Xu, "Lopinavir-ritonavir alone or combined with arbidol in the treatment of 73 hospitalized patients with COVID-19: a pilot retrospective study", *medRxiv*, 2020.