Global Journal on Application of Data Science and Internet of Things
Vol. 4, Issue 2 – 2020
ISSN: 2581-4370

Role of Machine Learning in Pattern Evaluation of COVID-19 Pandemic: A Study for Attribute Explorations and Correlations Discovery among Variables

Swarn Avinash Kumar[1], Harsh Kumar[2], Vishal Dutt[3], Himanshu Swarnkar[4]

[1]*IIIT Allahabad, UP, India.*
[2]*Peoples'Friendship University of Russia, Moscow, Russia.*
[3]*Department of Computer Science, Arybhatta College, Ajmer, India.*
[4]*Department of Computer Science Engineering, Engineering College, Banswara, India.*

Abstract

The Severe Acute Respiratory Syndrome (SARS-CoV-2) has challenged the highly developed health and care systems in countries around the world. This epidemic has spread its feet across the world in such a way that the medical system of developed countries has also collapsed. Even after consuming all the available medical resources, it has become impossible to save human life from this serious disease. Data of COVID-19 has been produced at a rapid pace from around the world using key traditional diagnostic techniques such as CT Scans, X-Ray scans, to identify the disease and this data proved to help machine learning algorithms prove its role in the field of health care. Assuming the base of this data, it is easy to predict the risk of severity with the help of machine learning algorithms. In this study, we analyse the dataset of the 1000 patients and applied the XG Boost classification algorithm for the evaluation of symptoms of COVID-19. Besides, the data was pre-processed for the better accuracy of the work and to find the correlations between the variable for getting the more prominent factors for the analysis of risk factors for disease.

Keywords: Boosting, Bagging, Learning Parameters, Regression, Dimensionality of data, Neural Network(s), Discrete, Hyper-parameters.

Introduction

COVID-19 is a severe respiratory illness caused by the virus SARS-CoV-2. The scientific community has focused on this disease with near-unprecedented intensity. However, the majority of primary studies published on COVID-19 suffered from small sample sizes [1-6]. Asymptomatic people who are infected with COVID-19 exhibit, by definition, no discernible physical symptoms of the disease. They are thus less likely to seek out testing for the virus, and could unknowingly spread the infection to others [7].

Artificial Intelligence (AI) has played a very important role in struggling with this crisis of COVID-19. The techniques of AI were used in the work of prognosis of the disease, care of patients, tracing of people in social distancing[8], finding a suitable treatment for the disease, etc. Being a subset of Artificial Intelligence, Machine Learning is capable to make machines learn like human intelligence that can learn from outside world problems, can learn from previous mistakes and experiences, that can take decisions logically and make predictions [9].

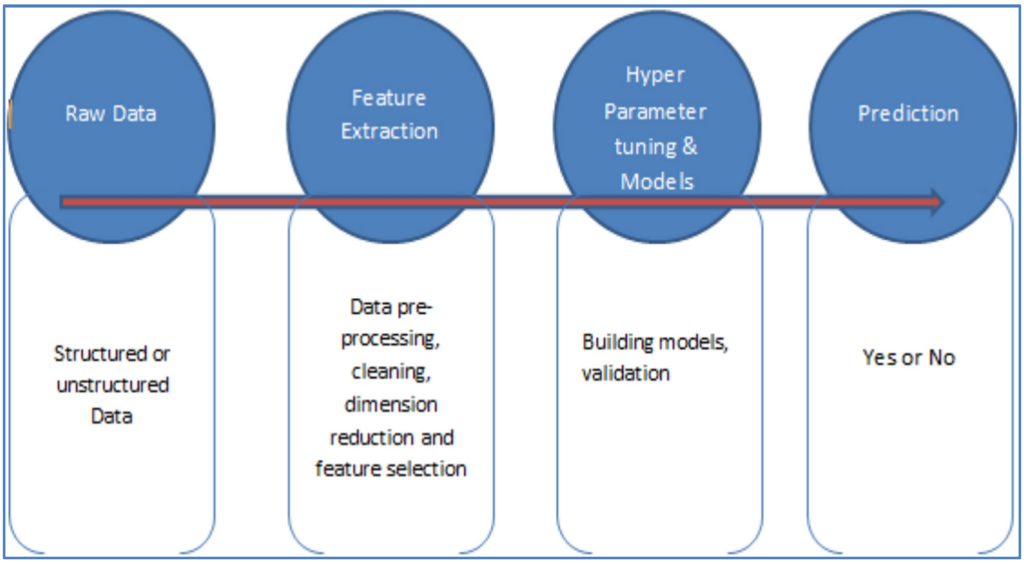
Raw Data
Feature Extraction
Hyper Parameter tuning & Models
Prediction
Structured or unstructured Data
Data pre-processing, cleaning, dimension reduction and feature selection
Building models, validation
Yes or No

Figure 1.Process of Learning for the evolution of Predictive model

Intelligent Predictive models are developed using supervised techniques of machine learning to predict COVID-19 infection. These supervised learning models of machine learning were used to detect negative and positive cases of COVID-19 in Mexico [10]. Many useful algorithms of machine learning were used to prepare such supervised predictive models such as Decision Trees, SVM (Support Vector Machine), LSTM (Long-Short Term Memory), SVR (Support Vector Regression), etc. [11].

Techniques of Machine Learning for COVID-19 Prediction

There are some trending Machine Learning Techniques widely used in the prediction of COVID-19

a) **SVR:** SVR is another form of SVM for the implementation of effective time series. And both SVM and SVR techniques are used to minimize the error in the margins and it employs the kernel function for both non-separable classes [12]. To improve the results achieved by these two, it is necessary to adapt their parameters by which its results can be improved and techniques like Heuristic search and regard grid can be utilized for achieving the best parameters. SVR can be written mathematically as follows for any multidimensional data:

$$A = f(x) = \sum_{i=1}^{M} W_i X_i + b \tag{1}$$

In the above expression, the Input feature values are represented by $X_{i,}$ and Input weights are represented by *Wi*, *b* represents the bias and *A* denotes the actual values [13].

b) LSTM: The LSTM was foreseen by Hochreiter and Schmidhuber [14]. An improved version of the RNN network is described as LSTM. In this model, RNN's deficiencies have been overcome by LSTM to overcome the drawbacks of RNA using hidden layers, known as memory cells. Self-connections mechanisms are used in memory cells to store the temporal state of the network. This state is controlled by the Input, Output, and forget gates. Input gate and output gate are used to control the flow of input and output in a memory cell [15]. Through the forget gate, the output information is passed from the previous neuron to the next neuron in the memory cell and is passed with high weights. Whether a piece of information will be stored in memory or not depends on the high activation of the result, i.e. the input unit whose activation is high will be stored in the information memory cell but if the activation of an input unit is low then its information is not stored in the memory cell And if the activation of any output unit is high, then the information stored in it and will be passed on to the next neuron [16].

$$Fg = sigmoid(W_{fg} X_t + W_{hfg} h_{t-1} + b_{fg}) \tag{2}$$

$$Ig = sigmoid(W_{ig} X_t + W_{hig} h_{t-1} + b_{ig}) \tag{3}$$

$$Og = sigmoid(W_{og} X_t + W_{hog} h_{t-1} + b_{og}) \tag{4}$$

Literature Review

Ardakani, A. An et al ., Authors used Clinical mammographic data in their study. The dataset contained the data of 1026 patients having 86 different attributes. They used the Hold out the method for the prediction. In the dataset, there were1020 CT Scan images were present. The CT is confirmed by the laboratory. And 86 out of them were surviving with viral and atypical pneumonia. The accuracy of their experiment was 99.51% with a specificity of 99.02% [17].

Rachna Murthy et al., (2020), Authors also worked in forecasting the new cases in the United States. They used the Spacio-temporal graph neural Network method in their study. The dataset they used was NYT COVID-19, the dataset had the new case records of the positive patients of COVID-19. They achieved a correlation of 0.998 in their proposed model [18].

Shahid et al.,(2020), Authors predicted the cases of COVID-19 with the specificity of confirmed, deaths and recovery cases. The data considered for the experiment was from January 2020 to June 2020 have all confirmed, death and recovery cases. The Models used in their study were ARIMA, LSTM, and SVR. These models were applied to the dataset of the patients of the 10 different countries. On this data, the models were compared and they found LSTM better in forecasting. They found 2.0463 and 0.0095 MAE for the confirmed and deaths respectively in their study [19].

Rustam et al.,(2020), Authors predicted the infection, deaths, and recovery cases from the data of the previous 56 days for the next 10 days. For this, they used Linear Regression, LASSO, SVM, and EM models. The data collected for prediction was from 22 June 2020 to 27 March 2020. And they achieved the best prediction results in the death rate prediction of R^2=0.98 [20].

Direkoglu et.al., (2020), the Authors forecasted the new cases in the next 10 days based onthe last 3 days data. They utilized the LSTM model to get the next 10 days' prediction of the new cases. The data used in their study was collected from WHO, CCDCP, and Worldometer. The records of the patients were before 10 April 2020 from worldwide. LSTM model was trained in their study and they achieved the RMSE of 1.5%[21].

Implementation

A) **XGBoost:** It is a most trending machine learning technique known for superior performance as compared to other machine learning algorithms. The abbreviation of XGBoost is eXtreme Gradient Boosting [22]. It is well suited for regression and classification predictive modeling problems. The XGBoost offers a wide range of hyper-parameters to control the procedures of training of the model. It provides fine-grained control. It is widely used for specifically imbalanced datasets. As the COVID-19 dataset is imbalanced due to the time series data, the XGBoost is well suited for modeling and training the dataset for the prediction [23].

B) **Pre-processing:** In this phase, the dataset is pre-processed. To pre-processing, the min, max, average, std. deviation and missing values were calculated. The attributes having missing and null values were removed before processing.

C) **Correlations:** The correlation is found between the various variables of the patients' dataset to explore the technical insights into the clinical characteristics of the COVID-19 dataset. The Spearman correlation test is applied for 2 continuous variables whereas, for one continuous and one categorical variable, the Kruskal-Wallis test was applied. If the p-value is below 0.6 then the outcome of the test is considered.

D) **Dataset Training & Testing:** To the implementation, the 1000 records of the patients were used and the records were split into two categories i.e. Training and Testing. For training, we used 70% of the data and the remaining 30% recodes are used for testing.In modeling the data, the 5-fold cross-validation with 70 boosting iterations is performed. Then to calculate the best hyper-parameters, we fed the iterations into the BOF (Bayesian Optimization function) [24]. Some parameters like max depth, learning rate, n_estimator, and gamma are included to tune the hyper-parameters.

Table 1.Dataset Variables Continues and Discrete

Variable Types	Number of Attributes
Discrete Variables	18
Continues Variables	21

E) **Workflow:** In the implementation of the XGBoost [25], some essential steps are required to accomplish the desired objective. The implementation is done in python, so the working is segmented into few steps like loading the essential libraries such as xgboost, readr, stringr, etc[26]. and after that the dataset is loaded to the algorithm, then the data cleaning processed and feature extraction is performed over the data and relevant variables, and at the end, the model is tuned for the better outcome. In tuning the model, the Bagging and boosting are performed to get the better outcomes for imbalanced dataset.

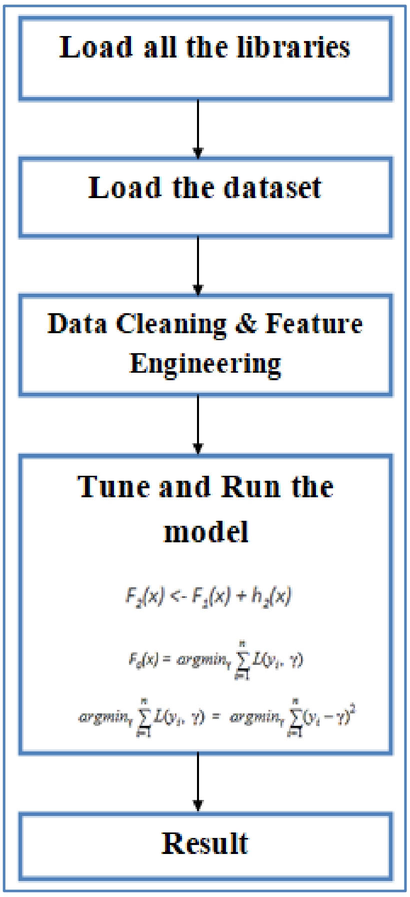
Load all the libraries
Load the dataset
Data Cleaning & Feature Engineering
Tune and Run the model
$F_2(x) <- F_1(x) + h_2(x)$
$F_0(x) = argmin_\gamma \sum_{i=1}^{n} L(y_i, \gamma)$
$argmin_\gamma \sum_{i=1}^{n} L(y_i, \gamma) = argmin_\gamma \sum_{i=1}^{n} (y_i - \gamma)^2$
Result

Figure 2.The workflow of the execution

Result and Analysis

A) **Attribute analysis:** In this part of the complete process, the attributes were analyzed in the context of the values contained. This study tries to get a better understanding of the attributes and their values through visualization.

i. **Statistical analysis of the Continuous attributes:** Some dominant statistics like Mean, Median, variance, etc. are calculated to get a better understanding of the attribute data.

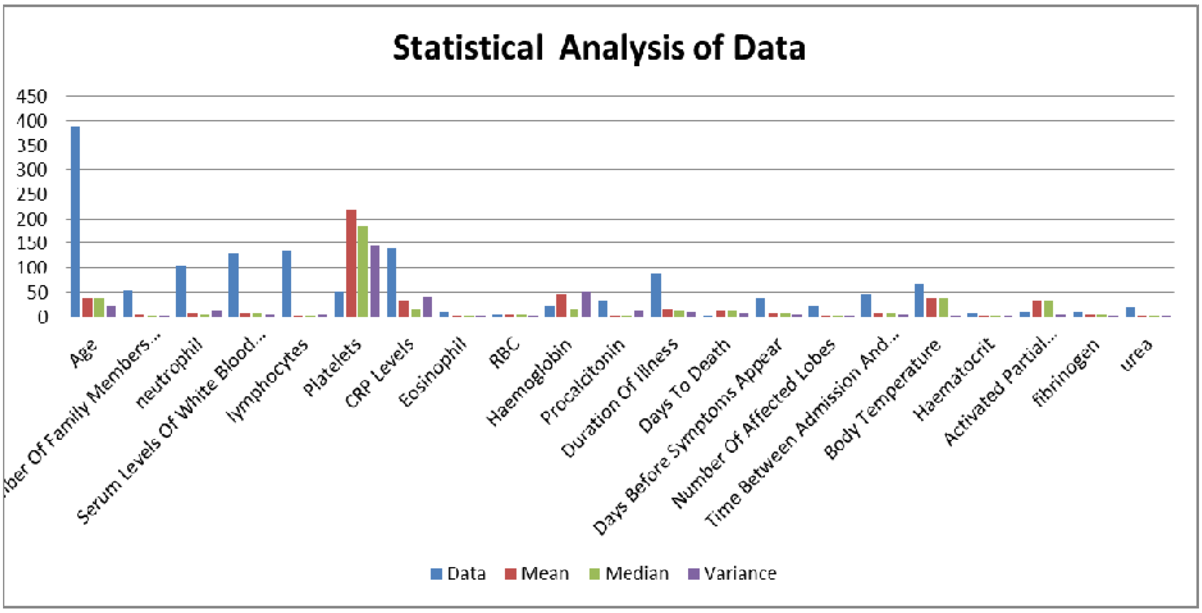

Figure 3.Visualization of statistical information about data

ii. **Statistical analysis of the Discrete attributes:** The discrete attributes have some categorical values that denote the presence of symptoms, level of symptoms.

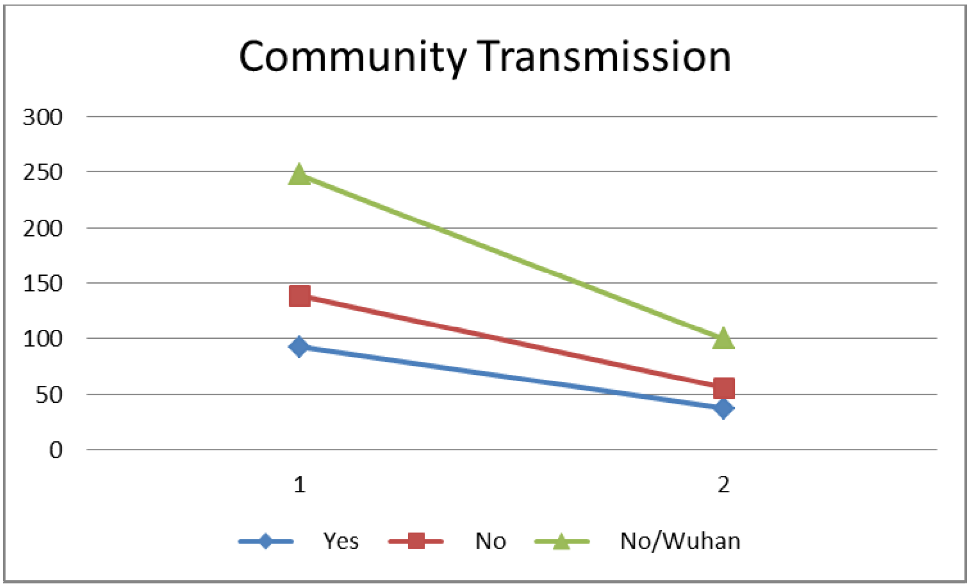

Figure 4.Frequency of disease transmission through community

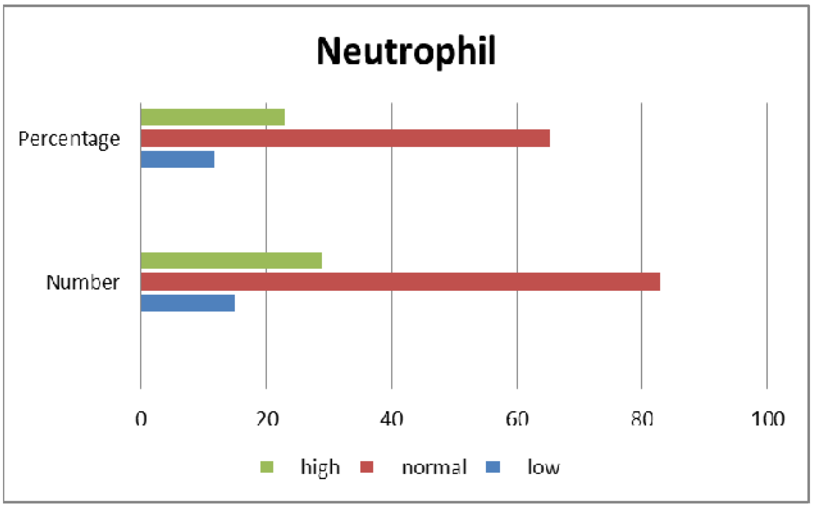

Figure 5.Frequency of Neutrophil level in patients

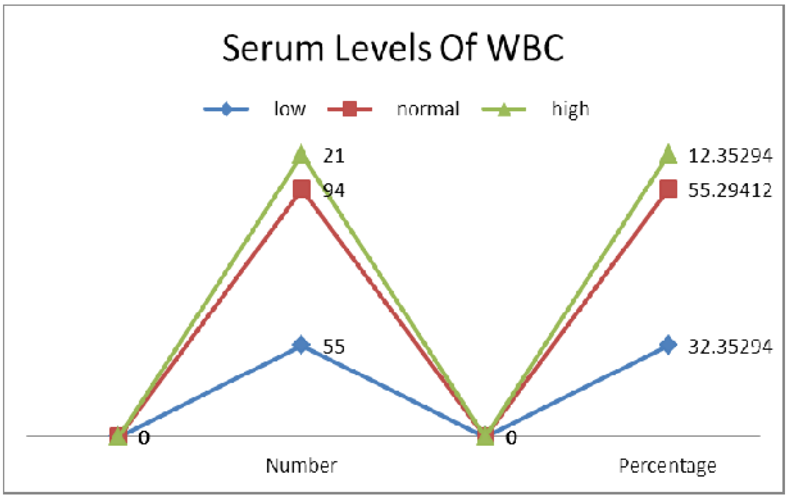

Figure 6.Frequency of Serum level of WBC in patients

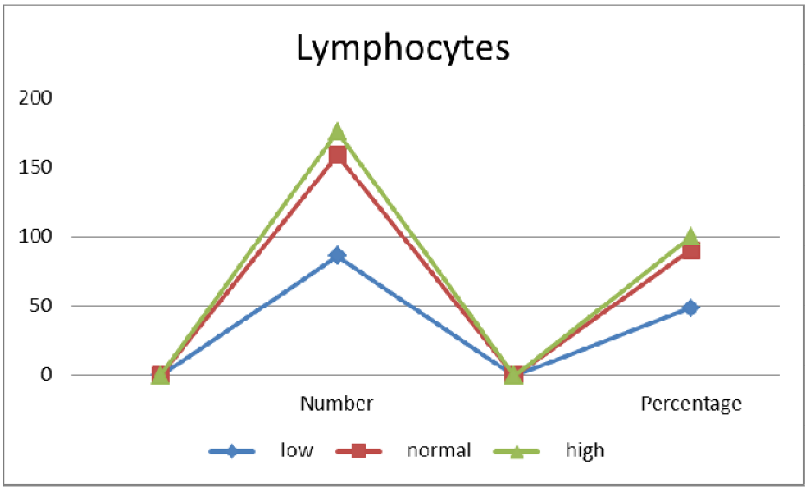

Figure 7.Frequency of Lymphocytes level in patients

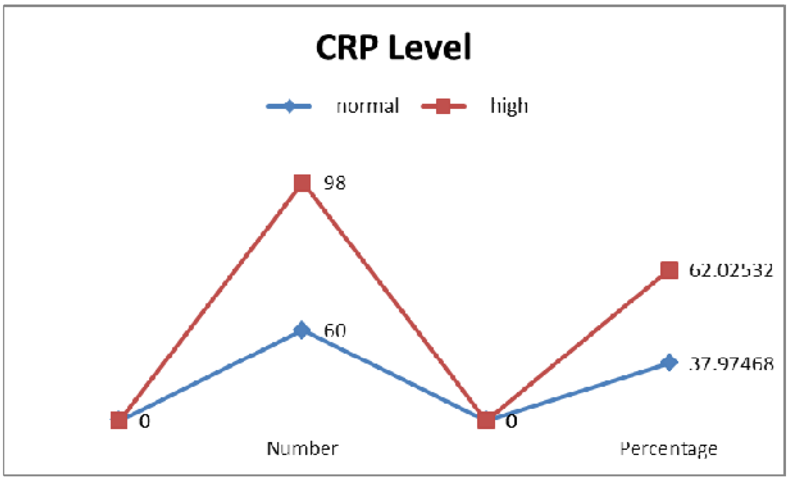

Figure 8.Frequency of CRP level in patients

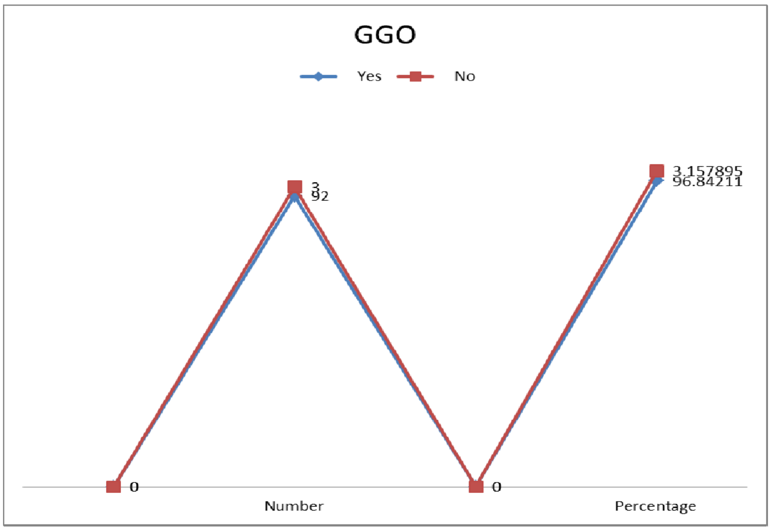

Figure 9.Frequency of GGO in patients

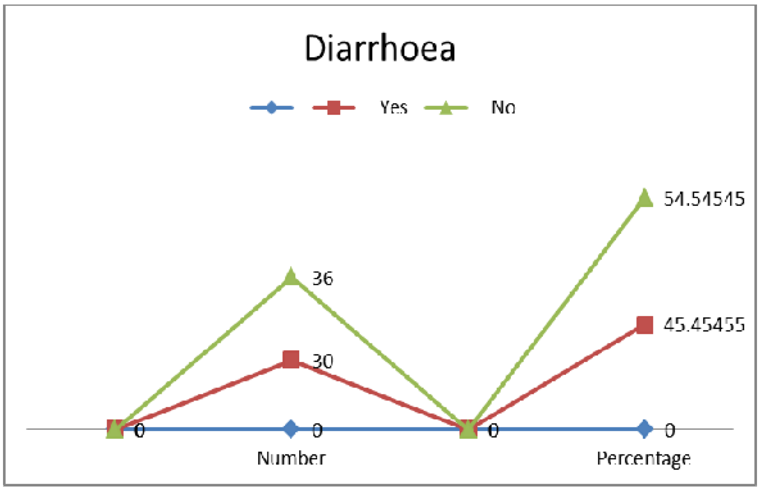

Figure 10.Frequency of Diarrhoea in patients

In the above visualizations figure 5 depicts that the level of Neutrophil found 11.81102%, 65.35433%, 22.83465% for low, normal, and high level respectively. Figure 6 depicts the frequency of WBC serum level with 32.35294%, 55.29412%, and 12.3529% for low, normal, and high level respectively COVID-19 patients. Figure 7 depicts that the level of Lymphocytes found 448.8636%, 41.4773%, and 9.65909% for low, normal, and high level respectively. In figure 8 it can be seen that the level of CRP found 37.97468%, 62.02532% fornormal and high respectively. In figure 9 it can be seen that 96.84211% and 3.157895% were the patient frequency for Yes and No categories respectively. Figure 10 shows the frequency for Yes and No categories 45.45455%, 54.54545%of Diarrhoea in patients respectively.

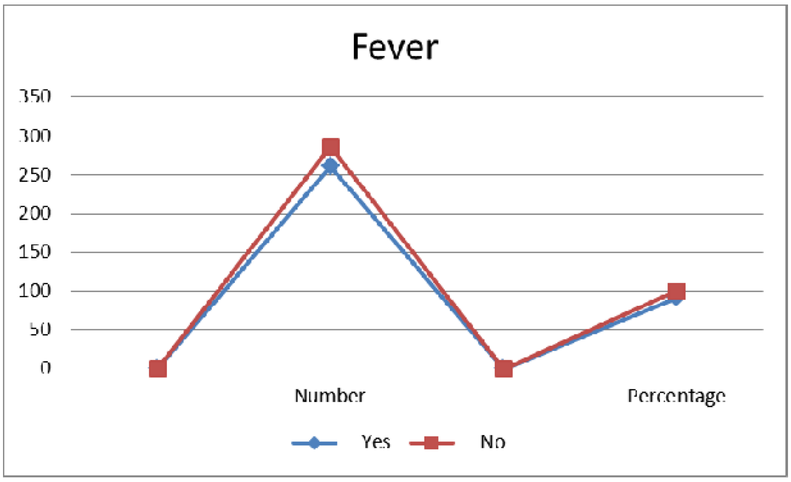

Figure 11.Presence and frequency of Fever in patients

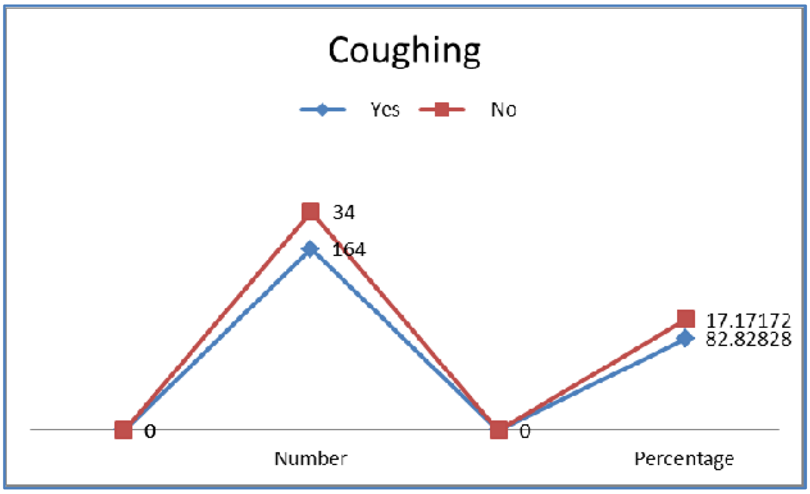

Figure 12.Presence and frequency of Coughing in patients

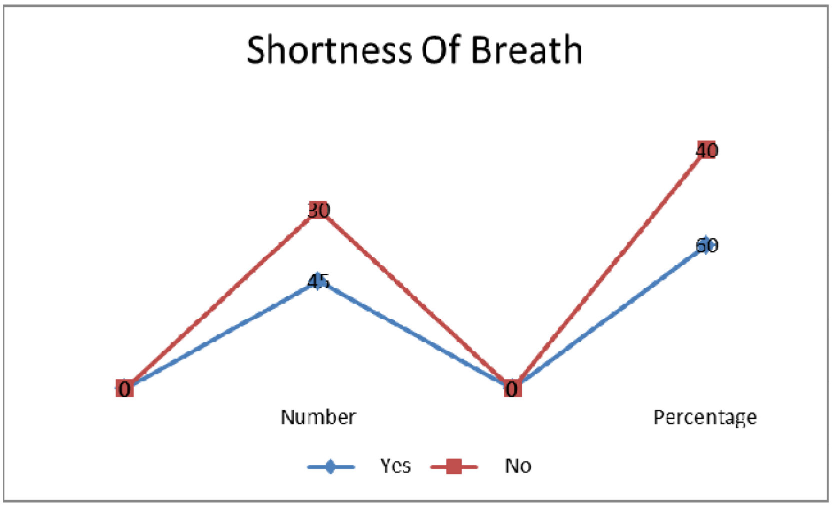

Figure 13.Frequency of breath shortness in patients

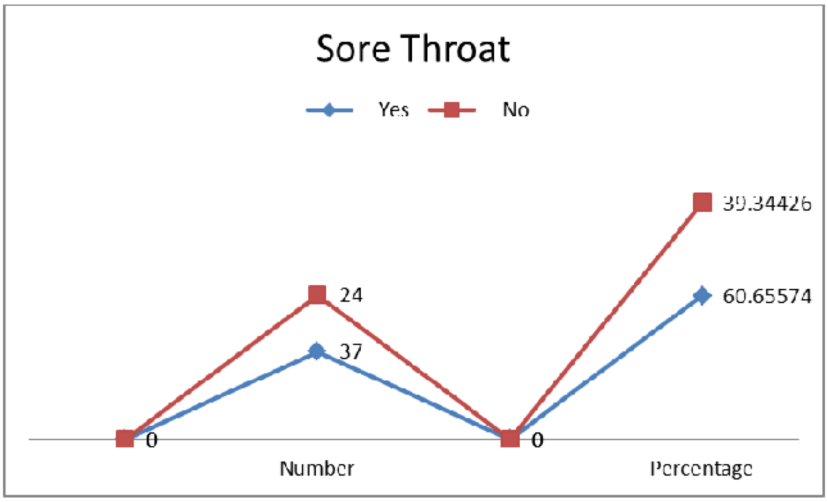

Figure 14.Presence and Frequency of soring in the throat in patient

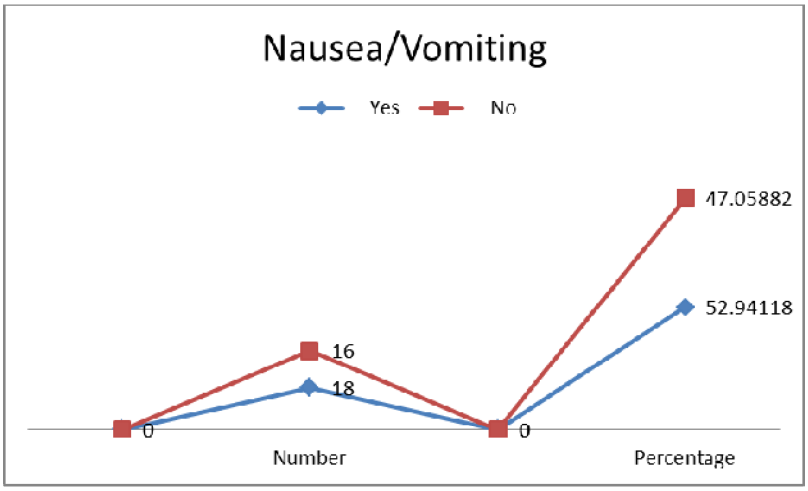

Figure 15.Presence and frequency of vomiting in patients

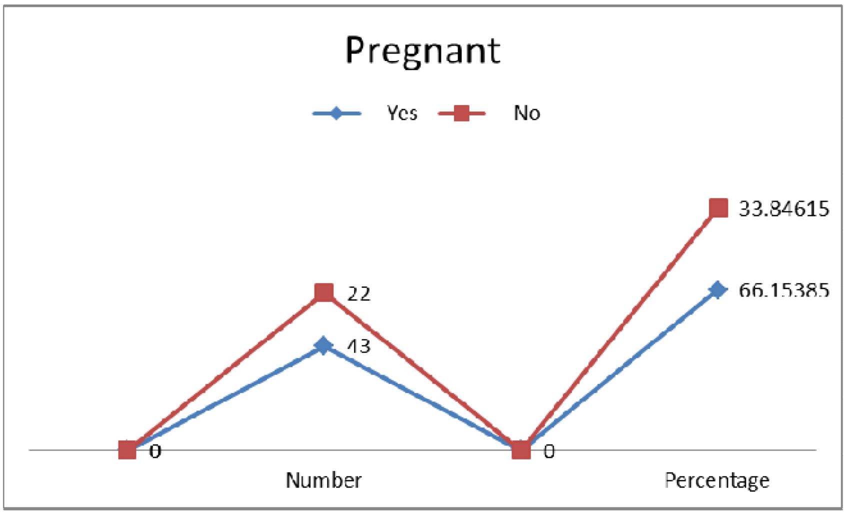

Figure 16.Presence and Frequency of pregnancy in female patients

In the above visualizations figure 11 depicts that the level of fever found 91.25874%, 8.741259% for Yes and No categories respectively. Figure 12 depicts the frequency of coughing with 82.82828%, 17.17172%, Yes and No categories respectively COVID-19 patients. Figure 13 depicts that the level of breathing problem found 60% and 40%, for Yes and No category respectively. In figure 14 it can be seen that the level of soring found 60.65574% and 39.34426 % for Yes and No category respectively. In figure 15 it can be seen that 52.94118% and 47.05882%were the patient frequency for Yes and No categories of vomiting problem respectively. Figure 16 shows the frequency for Yes and No categories 66.15385% and 33.84615 %of pregnancy cases in patients respectively.

B) **Correlation of variables' pairs:** The process of correlation is performed among all possible pairs of clinical variables to find out the relationship among various variables exhibited in Table 1. In the above table, both Continues and discrete variables are listed. For finding the correlations between two continuous variables, the Spearman test was performed. Total 143 Spearman test was performed and as the outcome of this correlation test, the 27 out of 143 depicts the significant correlations. As the conclusion of these tests, it is shown that the age factor plays a significant and important role in the COVID-19 disease development.

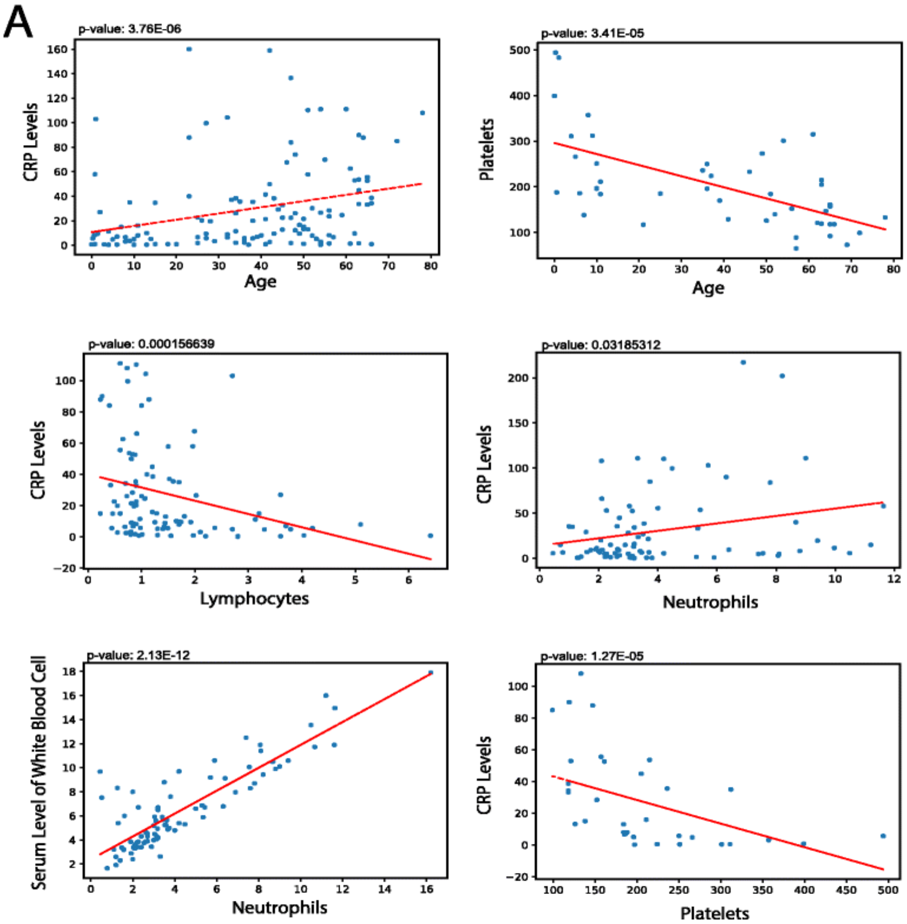

Figure 17.Visualization of Continues variables correlations

From the above visualization, it is seen that the correlation between the variable CRP level and serum platelets has the most significant relationship with the variable age. As the age increases, the correlation of CRP level significantly increases that makes a positive correlation with age.

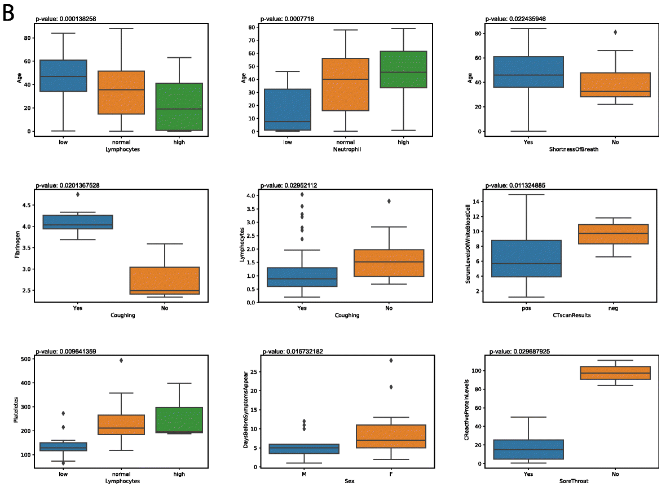

Figure 18.Visualization of correlations between 1 continues and 1 discreet variable

From the above visualization, it is seen the variable age is both positively and negatively correlated with other various variables. The table given below exhibits the relationship of age with another variable.

Table 2.Relationship of Age with Other Variables

Variable	Positive	Negative
lymphocyte level	No	Yes
neutrophil levels	Yes	No
shortness of breath	Yes	No

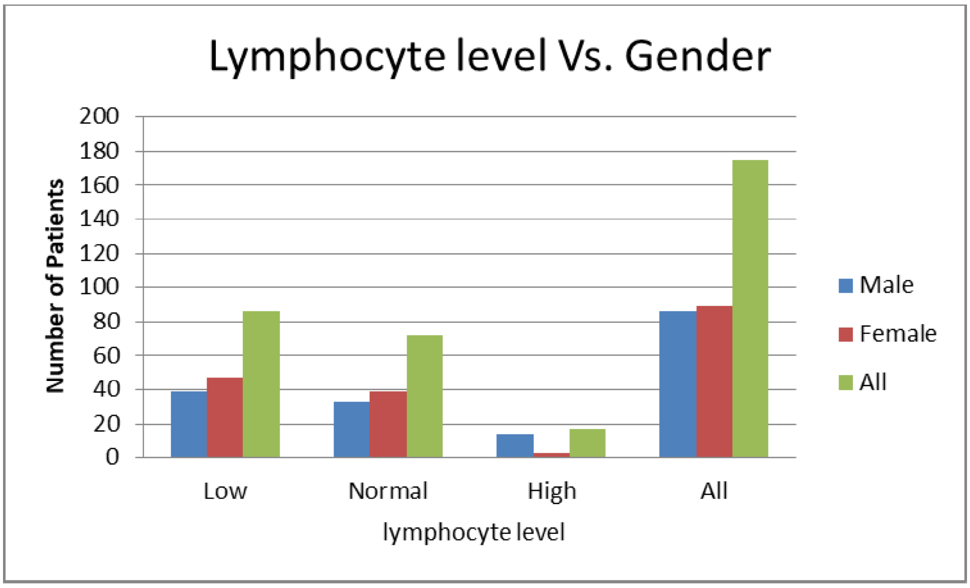

Figure 19.Visualization of correlation between Lymphocyte and Gender

The analysis from the above visualization depicts that the Male has a higher correlation with the Lymphocyte whereas the female has less correlation.

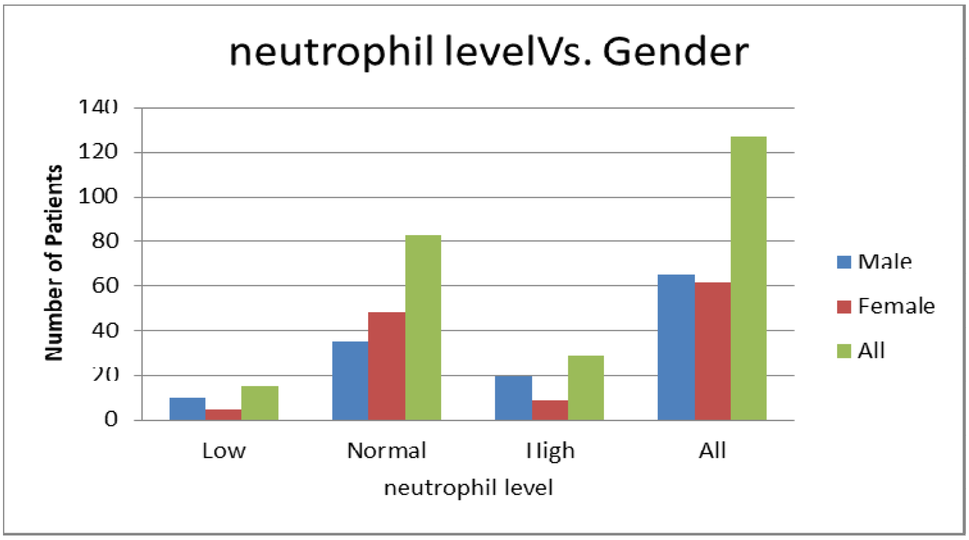

Figure 20.Visualization of correlation between Neutrophil and Gender

The analysis from the above visualization depicts that the Male has a higher correlation with the Neutrophil level whereas the females have less correlation as the level of Lymphocyte.

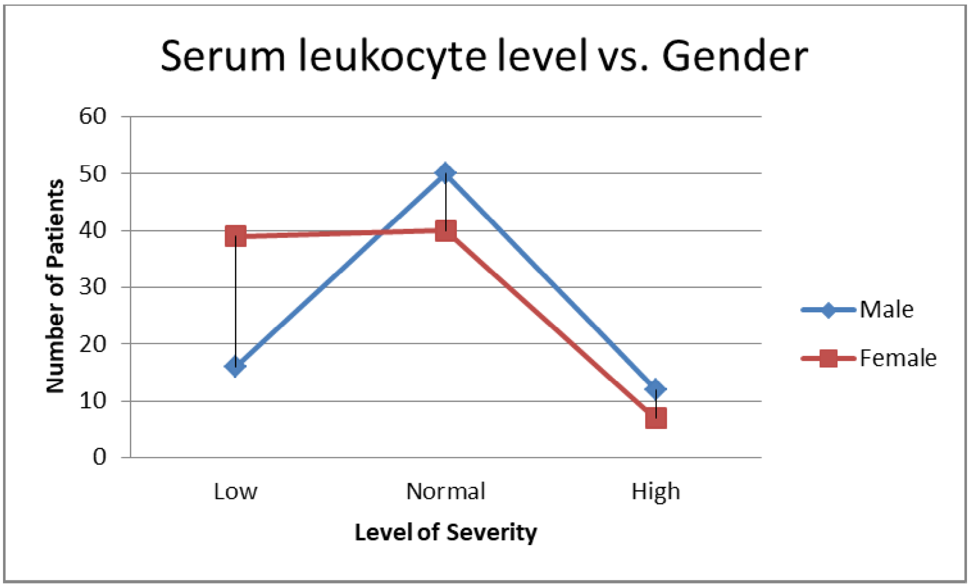

Figure 21.Visualization of correlation between Serum Leukocyte level and Gender

The analysis from the above visualization depicts that Females have a lower possibility of Serum leukocyte level in white blood cells as compare to males.

Conclusion

As we all know that a deadly epidemic like COVID-19 has come before the whole world. It is very necessary to know about this epidemic before fighting this deadly epidemic. It is very important to develop technologies to know and understand such fatal and terrible epidemics. Even today, many countries around the world are unable to recognize the disease among the citizens of their country and the health system of those countries is seen to be devoted to the right and proper treatment of life-threatening diseases like COVID-19. This study significantly focuses on the identification of patterns of this epidemic that can help mankind to identify this Aggressive disease in patients. According to the used dataset for this study, it can be concluded that the major symptoms of this epidemic are highly correlated with the Age factor and directly affect the older age patients. The neutrophil levels and shortness of breath are the symptoms that positively correlated with the gender of the patient.66.15385% of pregnant females are fighting with this COVID-19. The symptoms like fever, coughing, sour throat, shortness of breath are the major factor that creates the possibilities of COVID-19 infections in patients. It is analyzed in the study that 91.25847 patients had a fever, 82.82828% patients had a coughing problem, 60% of patients were facing the breathing problem, 60.65574% of patients had soring in the throat, 52.94118% patients had vomiting due to infection of COVID-19. The experiments conducted in this study and the facts derived from them can be helpful For people working in the medical field from all over the world to fight a terrible epidemic like COVID-19.

References

1. Abbasi, B., Saraf, D., Sharma, T., Sinha, R., Singh, S., Gupta, P. et al. (2020). Identification of vaccine targets; design ofvaccine against SARS-CoV-2 coronavirus using computational and deep learning-based approaches.
2. Ahmed, N., Michelin, R. A., Xue, W., Ruj, S., Malaney, R., Kanhere, S. S., & Jha, S. K. (2020). A survey of COVID-19 contact tracing apps. IEEE Access, 8, 134577-134601.
3. Alsdurf, H., Bengio, Y., Deleu, T., Gupta, P., Ippolito, D., Janda, R., Jarvie, M., Kolody, T., Krastev, S., Maharaj, T., and Obryk, R. (2020). COVI White Paper. arXiv preprint 2020,arXiv:2005.08502.
4. Ardakani AA, KanafiAR, Acharya UR, Khadem N, Mohammadi A. (2020). Application of deep learning technique tomanage COVID-19 in routine clinical practice using CT images: results of 10 convolutional neural networks. ComputBiol Med 2020; 121: 103795. 2020 https://doi.org/10.1016/j.compbiomed.2020.103795.
5. Ahuja, A., Reddy, V., Marques, O. (2020). Artificial intelligence and COVID-19: A multidisciplinary approach. Integrative Medicine Research, 9(3), 100434. https://doi.org/10.1016/j.imr.2020.100434.
6. Apostolopoulos, I. D., & Mpesiana, T. A. (2020). COVID-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. Physical and Engineering Sciences in Medicine.
7. Ardabili, S. F., Mosavi, A., Ghamisi, P., Ferdinand, F., Varkonyi-Koczy, A. R., Reuter, U., & Atkinson, P. M. (2020).COVID-19 outbreak prediction with machine learning. Available at SSRN 3580188.
8. Keshavarzi Arshadi, A., Webb, J., Salem, M., Cruz, E., Calad-Thomson, S., Ghadirian, N. et al. (2020). Artificial Intelligence for COVID-19 Drug Discovery and Vaccine Development. Frontiers in Artificial Intelligence, 3, https://doi.org/10.3389/frai.2020.00065.
9. Arora, P., Kumar, H., & Panigrahi, B. K. (2020). Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India. Chaos, Solitons & Fractals, 139, 110017.
10. Beck, B., Shin, B., Choi, Y., Park, S., Kang, K. (2020). Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model. Computational and Structural Biotechnology Journal, 18, 784-790. https://doi.org/10.1016/j.csbj.2020.03.025.
11. Burdick, H., Lam, C., Mataraso, S., Siefkas, A., Braden, G., Dellinger, R. P., & Hoffman, J. (2020). Prediction of respiratory decompensation in COVID-19 patients using machine learning: The READY trial. Computers in biology and medicine, 124, 103949.

12. COVID-19 Rapid Response Virtual Agent-Google Cloud. (n.d.). Retrieved October 22, 2020, fromhttps://cloud.google.com/solutions/contact-center/COVID19-rapid-response.
13. Dar, A. B., Lone, A. H., Zahoor, S., Khan, A. A., & Naaz, R. (2020). Applicability of mobile contact tracing in fighting pandemic (COVID-19): Issues, challenges, and solutions. Computer Science Review, 100307.
14. Eva, L. H., Lui, C., Woo, P. P., CHEUNG, A. T., Lam, P. K., Tang, T. W., & Lee, L. H. (2020). Development of a data-driven COVID-19 prognostication tool to inform triage and step-down care for hospitalized patients in Hong Kong: Apopulation-based cohort study. medRxiv.
15. Direkoglu, C., & Sah, M. (2020). Worldwide and Regional Forecasting of Coronavirus (COVID-19) Spread using a DeepLearning Model. medRxiv.
16. Ardakani AA, KanafiAR, Acharya UR, Khadem N, Mohammadi A. Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: results of 10 convolutional neural networks. Comput Biol Med 2020; 121:103795. 2020 https://doi.org/10.1016/j.compbiomed.2020.103795.
17. Rachna Murthy, David Eccleston, Darren Mckeown, Apul Parikh, Sophie Shotter, Improving aseptic injection standards in aesthetic clinical practice, Dermatologic Therapy, 10.1111/dth.14416, 34, 1, (2020).
18. Rustam, Furqan & Reshi, Aijaz & Mehmood, Arif & Ullah, Dr. Saleem & On, Byungwon & Aslam, Waqar & Choi, Gyu Sang. (2020). COVID-19 Future Forecasting Using Supervised Machine Learning Models. IEEE Access. PP. 1-1. 10.1109/ACCESS.2020. 2997311.
19. Direkoglu, Cem & Sah, Melike. (2020). Worldwide and Regional Forecasting of Coronavirus (COVID-19) Spread using a Deep Learning Model. 10.1101/2020. 05.23.20111039.
20. Chakraborty T, Ghosh I. Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: a data-driven analysis. Chaos, Solitons Fractals 2020:109850. DOI: 10.1016/j.chaos.2020.109850.
21. Ke Y-Y, Peng T-T, Yeh T-K, Huang W-Z, Chang S-E, Wu S-H, Hung H-C, Hsu T-A, Lee S-J, Song J-S, Lin W-H, Chiang T-J, Lin J-H, Sytwu H-K, Chen C-T. Artificial intelligence approach fighting COVID-19 with repurposing drugs. Biomed J 2020. DOI: 10.1016/j.bj.2020.05.001.
22. Beck BR, Shin B, Choi Y, Park S, Kang K. Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model. Comput Struct Biotechnol J 2020; 18: 784-90. DOI: 10.1016/j.csbj.2020.03.025.
23. Ekins S, Freundlich J, Coffee M . A common feature pharmacophore for DA-approved drugs inhibiting the Ebola virus. F10 0 0Research 2014; 3: 277.
24. Ekins S, Mottin M, Ramos PRPS, Sousa BKP, Neves BJ, Foil DH, Zorn KM, Braga RC, Coffee M, Southan C, Puhl CA, Andrade CH. Déjàvu: stimulating open drug discovery for SARS-CoV-2. Drug Discov Today 2020. DOI: 10.1016/j. drudis.2020.03.019.