Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Application of machine learning for the diagnosis of COVID-19

Prajoy Podder¹, Subrato Bharati¹, M. Rubaiyat Hossain Mondal¹, Utku Kose²

¹INSTITUTE OF ICT, BANGLADESH UNIVERSITY OF ENGINEERING AND TECHNOLOGY, DHAKA, BANGLADESH; ²SULEYMAN DEMIREL UNIVERSITY, ISPARTA, TURKEY

1. Introduction

In late Dec. 2019, the world came to know about a deadly coronavirus disease in Wuhan, China. Soon, this disease started to spread in different countries. This was initially named the 2019 novel coronavirus by the World Health Organization (WHO). Later, in Feb. 11, 2020, the WHO officially named the disease coronavirus disease 2019 (COVID-19). The Coronavirus Study Group of the International Committee termed the virus that caused COVID-19 a severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) on Feb. 11, 2020 [1] The Chinese scientists rapidly isolated a SARS-CoV-2 from a patient within a short time on Jan. 7, 2020 and genome sequenced the SARS-CoV-2 [2]. As of Apr. 22, 2020, there were 24,71,136 confirmed cases and 1,69,006 deaths globally in which a total of 7,76,907 cases of COVID-19 were confirmed in the United States, including 37,602 deaths. On the other hand, a total of 84,287 cases of COVID-19 were confirmed in China, the origin of coronavirus, including 4642 deaths [3].

SARS-CoV-2 has a number of structural proteins: spike (S), envelop (E), membrane (M), and nucleocapsid (N) [4]. Of these four proteins, the RNA genome is held by the N protein, which helps to form a coiled tubular structure. The other three proteins (S, E, and M) are responsible for developing the envelope of the virus. The E protein surrounds the helical nucleocapsid. The S protein helps the virus get into a host by attaching to the cell membrane of that host. This SARS-CoV-2 has five essential genes, which are for the four structural proteins and for viral replication. The genome organization of SARS-CoV-2 is 5'-RdRp-S-E-M-N-3' [5,6]. This genome sequencing indicates that SARS-CoV-2 is a novel bat coronavirus different from that of SARS-CoV [4,7–11]. It was also revealed that SARS-CoV-2 has 79% identity with SARS-CoV and 51.8% similarity with Middle East respiratory syndrome coronavirus.
The increase in the number of cases in COVID-19 is threatening to overwhelm health systems around the world with the demand for intensive care unit beds far above the existing capacity. According to WHO, more than 50 COVID-19 vaccine candidates are existing in trial stage. Therefore, preventive measures need to be taken to avoid becoming infected by this virus. Usually, the virus gets into human body through the eyes, throat, and nose, carried by the hand. The WHO has asked people to maintain personal hygiene by washing hands with soap and water frequently, and covering the mouth and nose with an elbow or a tissue when coughing or sneezing. Moreover, cleaning surfaces with disinfectants and maintaining social distancing are important to prevent this disease. In many countries, the spread of SARS-CoV-2 has resulted in the lack of testing kits to diagnose the virus. There are not enough testing kits or trained personnel to test the virus in suspected patients. In some cases, doctors, nurses, and medical support team members are also becoming infected, or they have to quarantine, which makes the scenario more difficult for patients. Therefore, there is a need for the data-driven diagnosis of COVID-19 patients. This chapter focuses on the automatic detection of COVID-19 using a dataset available in https://www.kaggle.com/einsteindata4u/covid19 [12]. First, the condition of COVID-19 in the world is visualized using a number of tables and graphs. This indicates confirmed deaths and recovered cases in different countries and continents. Second, machine learning techniques are used on a dataset [12] of COVID-19 patients. Both feature selection and classification algorithms are applied, and it is shown that for the given dataset, the COVID-19 disease can be predicted reliably.

2. Visualization of the spread of coronavirus disease 2019

In this section, the spread of COVID-19 is presented through data visualization. For visualization purposes, the dataset is collected from a website [13,14]. Fig. 9.1 illustrates

![FIGURE 9.1 Continentally reported cases (up to Apr. 23, 2020).]
continentally reported cases for which confirmed deaths and recovered or active cases are discussed. The mortality rate per 100 people is also described. Fig. 9.1 shows that the number of confirmed cases is the highest on the European continent (11, 65, and 661, respectively). The mortality rate is also the highest in Europe (9.55%). Fig. 9.2 depicts the list of top 10 countries of confirmed COVID-19 cases, recovered cases, death cases and active cases.

3. Methodology

This dataset [12] was generated from patients at the Hospital Israelita Albert Einstein in São Paulo, Brazil. The samples are collected anonymously performing the SARS-CoV-2 reverse transcriptase polymerase chain reaction and additional laboratory tests. The data were standardized by converting the samples so that the mean value of the samples is zero whereas the standard deviation is unity.

The dataset has 5644 rows and 111 columns. The dataset is imbalanced. There are a number of missing values in the data samples. For this, features with more than 99.8% of
null values in positive cases are dropped because they are unlikely to contribute to the prediction of COVID-19. Fig. 9.3 shows that the dataset was originally imbalanced with 90.1% samples representing negative cases. After removing attributes with at least 99.8% null values, the dataset became balanced with 51.1% representing negative cases. Table 9.1 shows the list of dropped features that have at least 99.8% null values. The details of feature filtering are illustrated in Fig. 9.4.

The figures show the percentage of negative and positive cases after undersampling. Thus the dataset can be considered balanced. The new dataset has 1091 rows and 61 columns; it will only have numerical features. The target feature is converted to 0 or 1, in which 1 means positive and 0 means negative.

In this section, experiments are performed to classify normal and COVID-19 patients using samples in the dataset. This research work is implemented using the scikit-learn library of Python programming language. Steps followed in this implementation are shown in Fig. 9.5. A number of processes are performed, including data labeling and data filtering, which are part of preprocessing. Next, important features are selected.

![Figure 9.3](image)

**FIGURE 9.3** Positive and negative cases of data samples: (A) imbalanced case in original dataset; (B) balanced case after processing.

| Table 9.1 | List of primarily dropped features. |
| --- | --- |
| FiO$_2$ (venous blood gas analysis) | Vitamin B12 |
| o-Dimer | *Mycoplasma pneumoniae* |
| Urine—nitrite | Urine—sugar |
| Partial thromboplastin time | Albumin |
| Prothrombin time, activity | Phosphor |
Classification algorithms are then applied on the selected features. A number of popular classification algorithms such as random forest (RF), logistic regression (LR), decision tree (DT), and XGBoost are considered. Both cross-validation (cv) and holdout methods are considered. For cv, the KFold() function, and for holdout, the train_test_split() function from scikit-learn library are used to split the dataset. Next, the classification models are fitted with the training data and the models are then used to predict COVID-19 samples.
4. Feature importance and feature scoring

There are a number of feature selection algorithms. In this case, a univariate feature selection method is considered. For this, the SelectKBest() function of the scikit-learn library is used. Table 9.2 shows the top 25 features with their corresponding scores.

**FIGURE 9.5** Workflow diagram. DT, decision tree; LR, logistic regression; RF, random forest; XGB, XGBoost.
obtained using SelectKBest(). Table 9.3 shows the ranking of these top 25 features. For the dataset considered, serum glucose is the best-ranked attribute, or the most influential feature in predicting a COVID-19 patient.

| Name of feature                                      | Score      |
|------------------------------------------------------|------------|
| Protein C-reactive, mg/dL                            | 20,857.493593 |
| Leukocytes                                           | 5800.140368 |
| Lymphocytes                                          | 2888.050022 |
| Neutrophils                                          | 2774.036881 |
| Alanine transaminase                                 | 2237.563824 |
| pO2 (venous blood gas analysis)                      | 1835.797530 |
| γ-Glutamyltransferase                               | 1584.636027 |
| Platelets                                            | 1505.770002 |
| Monocytes                                            | 1462.219391 |
| Eosinophils                                          | 1441.501397 |
| Red blood cells                                      | 1243.403849 |
| Mean corpuscular volume                              | 1128.205213 |
| Aspartate transaminase                              | 1030.129529 |
| Indirect bilirubin                                   | 949.508049  |
| Hematocrit                                           | 934.136514  |
| pCO2 (partial pressure of carbon dioxide within venous blood) | 870.907532  |
| Red blood cell distribution width                    | 807.758742  |
| Creatinine                                           | 733.620615  |
| Serum glucose                                        | 729.902624  |
| pH (venous blood gas analysis)                       | 664.791565  |
| Total bilirubin                                      | 587.205740  |
| Parainfluenza 3                                     | 518.255556  |
| Urea                                                 | 450.905016  |
| Hb saturation (venous blood gas analysis)            | 439.066476  |
| Parainfluenza 4                                     | 362.666667  |

5. Classification using machine learning

After selecting top features by the feature selection method, the feature subset is then taken into the classifier training stage. In the training stage, XGBoost [15], RF [16–18], LR [19,20], and DT are employed. Fig. 9.5 illustrates the stages of this implementation. Fig. 9.5 shows that the dataset is initially preprocessed, followed by the feature selection process. Next, the data samples are split into training and testing samples. Then, the training data is used to fit a classifier model. The testing data are then applied to the model to predict the target: in this case, COVID-19. Finally, the testing target value that is the actual value is compared with the predicted value.
XGBoost is a popular form of gradient boosting algorithm designed for optimal hardware use. It is an implementation of gradient-boosted DTs. XGBoost can penalize a model for complexity using L1 and L2 regularization in which regularization prevents overfitting of the XGBoost model. Regularization helps prevent overfitting. Algorithm 1 describes how XGBoost is used to classify COVID-19 patients.

**Algorithm 1. Detection of positive COVID-19 patient using XGBoost**

Input: A list of features
Output: Classification report, confusion matrix, receiver operating characteristic (ROC) curve

Process:

1. Standardize the selected features using StandardScaler() function
2. Apply XGBoost classifier using XGBClassifier (base_score=0.5, booster='gbtree', gamma=0, learning_rate=0.1, max_depth=3, n_estimators=100, objective='reg:linear', random_state=0) function on the selected features

### Table 9.3 Top 25 features with their associated ranking.

| Name of feature                  | Associated rank |
|----------------------------------|-----------------|
| Serum glucose                    | 1               |
| Respiratory syncytial virus      | 2               |
| Influenza A                      | 3               |
| Influenza B                      | 4               |
| Coronavirus NL63                 | 5               |
| Coronavirus HKU1                 | 6               |
| Parainfluenza 3                  | 7               |
| Adenovirus                       | 8               |
| Parainfluenza 4                  | 9               |
| Coronavirus 229E                 | 10              |
| Influenza A H1N1 2009            | 11              |
| Metapneumovirus                  | 12              |
| Influenza B, rapid test          | 13              |
| Influenza A, rapid test          | 14              |
| Alanine transaminase             | 15              |
| Aspartate transaminase           | 16              |
| Total bilirubin                  | 17              |
| Direct bilirubin                 | 18              |
| Indirect bilirubin               | 19              |
| Hb saturation (venous blood gas analysis) | 20 |
| Base excess (venous blood gas analysis) | 21 |
| pO2 (venous blood gas analysis)  | 22              |
| Total CO2 (venous blood gas analysis) | 23 |
| pH (venous blood gas analysis)   | 24              |
| HCO3 (venous blood gas analysis) | 25              |
3. Train the model using selected features
4. Predict result using test dataset
5. Evaluate the accuracy of the classifier using accuracy_score() function
6. Use confusion_matrix() function to evaluate true negative (TN), false positive (FP),
   false negative (FN), and true positive (TP).
7. Use classification_report() function to calculate precision, recall, and F1 score

5.2 Random forest

RF is a combination of multiple DTs. Two important concepts make this algorithm
random: the randomness in the sampling of the training portion of the data and the
randomness in the selection of features for the splitting nodes. The RF algorithm
maintains the reliability of a large part of the dataset by handling any missing sample
values. Algorithm 2 describes the stages of RF in classifying COVID-19 patients.

Algorithm 2. Detection of positive COVID-19 patient using RF
Input: A list of features according to rank
Output: Classification report, confusion matrix, accuracy
Process:
1. Standardize the selected features using StandardScaler() function
2. Apply RF using RFClassifier (n_estimators=100, criterion=’gini’) function with
   some parameter on the selected features
3. Train the model using selected features
4. K-fold parameters for K-fold cv: thresh = 0.5, k_fold_seed = 13, n_folds = 10
5. Predict the result using test dataset
6. Evaluate the accuracy of the classifier function
7. Use confusion_matrix() function to evaluate TN, FP, FN, and TP
8. Use classification_report() function to calculate precision, recall, and F1 score

5.3 Decision tree and logistic regression

Other popular classification algorithms are DT and LR. Algorithm 3 shows important
steps of DT and LR classifiers in predicting patients affected by COVID-19.

Algorithm 3. Detection of positive COVID-19 patient using DT and LR
Input: A list of features according to rank
Output: Classification report, confusion matrix, accuracy
Process:
1. Standardize the selected features using StandardScaler() function
2. Apply DT using DTClassifier (criterion=’entropy’, max_depth=5, random_state=0)
   function with some parameter on the selected features
Or apply LR using LogisticRegression() function

3. Train the model using selected features
4. K-fold parameters for K-fold cv: thresh = 0.5, k_fold_seed = 13, n_folds = 10
5. Predict the result using test dataset
6. Evaluate the accuracy of the classifier function
7. Use confusion_matrix() function to evaluate TN, FP, FN, and TP
8. Use classification_report() function to calculate precision, recall, and F1 score

6. Performance parameters

For the analysis and diagnosis of medical data including COVID-19 samples, total accuracy is insufficient for evaluating the machine learning algorithm. Furthermore, it is important to diagnose the patient correctly because an incorrect prediction of a COVID-19—affected patient as a normal patient can be a serious issue. Several popular metrics are used in this chapter for the data-driven diagnosis of COVID-19 patients. To evaluate the performance, TP refers to COVID-19 samples correctly classified as SARS-CoV-2—affected patients. TN denotes the number of normal people who correctly have negative predictions. That means they are classified as normal patients. FN denotes the number of undetected patients who actually have COVID-19, and FP denotes the number of samples without COVID-19 but that are wrongly classified as those of patients with COVID-19. With this consideration, they can be defined as shown in Fig. 9.6.

Fig. 9.6 shows a performance matrix and its mathematical illustration including precision, recall, TP rate, TN rate, accuracy, miss rate (FN rate), and F1 score. The performance evaluation is conducted for several classifiers including XGBoost, RF, LR, and DT for several cv numbers. For this, three- to 10-fold cv is considered. Table 9.4 presents the performance results for XGBoost using cv. A number of metrics such as precision, recall, F1 score, and testing accuracy are presented in Table 9.4. Table 9.4 shows that XGBoost provides the highest accuracy, precision, recall, and F1 score for cv 5, for which the highest accuracy of XGBoost is 97.2477%.

The performance results of RF, LR, and DT are evaluated in Tables 9.5—9.7, respectively. The accuracy of RF, LR, and DT are 95.4128%, 98.1651%, and 95.4128%, respectively, where the number of cvs is 5. Hence, fivefold cv provides the highest accuracy, precision, and recall for these four classifiers for the given dataset. Compared with other three classifiers, LR provides the highest accuracy (98.1651%) and the same value of precision, recall, and F1 score (98%). XGBoost provides the second highest accuracy of 97.2477% (in Table 9.4). Hence, the performance results vary with the difference in cv fold value. Next, we compare the classifiers by taking the average of results for cv values of 3—10. For example, we take the average of values of XGBoost for cv 3—10 shown in
Table 9.4  Performance results of XGBoost using cross-validation.

| Fold (cross-validation) | Precision (%) | Recall (%) | F1 score (%) | Testing accuracy (%) |
|-------------------------|--------------|------------|--------------|-----------------------|
| 3                       | 92           | 92         | 92           | 91.7431               |
| 4                       | 91           | 91         | 91           | 90.8257               |
| 5                       | 97           | 97         | 9            | 97.2477               |
| 6                       | 93           | 93         | 93           | 92.6606               |
| 7                       | 94           | 94         | 94           | 94.4954               |
| 8                       | 90           | 89         | 89           | 88.9908               |
| 9                       | 93           | 93         | 93           | 92.6606               |
| 10                      | 94           | 94         | 94           | 93.5779               |
Table 9.4 and compute records for average cv. Similarly, this is done for other classifiers using Tables 9.5–9.7. Table 9.8 compares the classifiers in terms of precision, recall, F1 score, testing accuracy, miss rate, specificity, and confusion matrix using an average value of cv. XGBoost has the highest accuracy value of 92.67%, and LR has the second

**Table 9.5** Performance results of random forest using cross-validation.

| Fold (cross-validation) | Precision (%) | Recall (%) | F1 score (%) | Testing accuracy (%) |
|-------------------------|---------------|------------|--------------|-----------------------|
| 3                       | 91            | 90         | 90           | 89.9083               |
| 4                       | 92            | 91         | 91           | 90.8257               |
| 5                       | 96            | 95         | 95           | 95.4128               |
| 6                       | 92            | 92         | 92           | 91.7431               |
| 7                       | 94            | 94         | 94           | 93.5779               |
| 8                       | 88            | 86         | 86           | 86.2385               |
| 9                       | 91            | 91         | 91           | 90.8257               |
| 10                      | 93            | 93         | 93           | 92.6606               |

**Table 9.6** Performance results of logistic regression using cross-validation.

| Fold (cross-validation) | Precision (%) | Recall (%) | F1 score (%) | Testing accuracy (%) |
|-------------------------|---------------|------------|--------------|-----------------------|
| 3                       | 91            | 91         | 91           | 90.8257               |
| 4                       | 92            | 92         | 92           | 91.7431               |
| 5                       | 98            | 98         | 98           | 98.1651               |
| 6                       | 93            | 93         | 93           | 92.6606               |
| 7                       | 96            | 96         | 96           | 96.3303               |
| 8                       | 90            | 90         | 90           | 89.9083               |
| 9                       | 88            | 88         | 88           | 88.0734               |
| 10                      | 93            | 93         | 93           | 92.6606               |

**Table 9.7** Performance results of decision tree using cross-validation.

| Fold (cross-validation) | Precision (%) | Recall (%) | F1 score (%) | Testing accuracy (%) |
|-------------------------|---------------|------------|--------------|-----------------------|
| 3                       | 89            | 89         | 89           | 88.9908               |
| 4                       | 85            | 85         | 85           | 85.3211               |
| 5                       | 96            | 95         | 95           | 95.4128               |
| 6                       | 89            | 89         | 89           | 88.9908               |
| 7                       | 96            | 95         | 95           | 95.4128               |
| 8                       | 84            | 83         | 83           | 83.4862               |
| 9                       | 87            | 87         | 87           | 87.1559               |
| 10                      | 89            | 88         | 88           | 88.0734               |
The highest accuracy of 92.58%. Both XGBoost and LR have the highest values of precision, recall, and F1 score, all of which are 93%. Hence, XGBoost and LR are good choices for classifying this specific dataset and thus predict COVID-19 patients reliably.

Next, the classifiers are evaluated in term of ROC curves and the area under the curve (AUC). Fig. 9.7 shows the ROC curves for (a) XGBoost, (b) RF, (c) LR, and (d) DT, where the AUC values for XGBoost, RF, LR, and DT are 97.1%, 96.8%, 96.4%, and 94.4%, respectively.

Next, the overall performance of the classifiers is shown for the holdout method, in which the dataset is split into different portions of testing and training samples. The results vary with the difference in the splitting. In this case, we split the dataset so that 80% of the data samples are used for training and the remaining 20% are used for testing. Table 9.9 shows the performance results for different classifiers when 20% data samples are used for testing. LR has the best testing accuracy of 94.06%. Moreover,

### Table 9.8 Overall performance comparison of classifiers using cross-validation.

| Classifier         | Precision (%) | Recall (%) | Miss rate or false-negative rate (%) | Specificity (%) | F1 score (%) | Accuracy (%) | Confusion matrix |
|--------------------|---------------|------------|--------------------------------------|-----------------|--------------|--------------|------------------|
| XGBoost            | 93            | 93         | 7                                    | 94.183          | 93           | 92.67        | Predicted 0.0 1.0 Actual 0.0 502 49 1.0 31 509 |
| Random forest      | 92            | 92         | 8                                    | 97.373          | 92           | 91.84        | Predicted 0.0 1.0 Actual 0.0 519 75 1.0 14 483 |
| Logistic regression| 93            | 93         | 7                                    | 93.058          | 93           | 92.58        | Predicted 0.0 1.0 Actual 0.0 496 44 1.0 37 514 |
| Decision tree      | 90            | 90         | 10                                   | 92.308          | 90           | 89.73        | Predicted 0.0 1.0 Actual 0.0 492 71 1.0 41 487 |
FIGURE 9.7 Receiver operating characteristic (ROC) and area under the curve (AUC) comparison of (A) XGBoost, (B) random forest, (C) logistic regression, and (D) decision tree classifiers.

Table 9.9 Performance comparison of the classifiers using holdout method.

| Classifier       | Precision (%) | Recall (%) | Miss rate or false negative rate (%) | F1 score (%) | Accuracy (%) | Area under the curve (%) |
|------------------|---------------|------------|-------------------------------------|--------------|--------------|--------------------------|
| XGBoost          | 92            | 92         | 8                                   | 92           | 92.8374      | 92                       |
| Random forest    | 91            | 91         | 9                                   | 91           | 91.7574      | 90                       |
| Logistic regression | 94           | 94         | 6                                   | 94           | 94.0639      | 94                       |
| Decision tree    | 89            | 89         | 11                                  | 89           | 88.5845      | 89                       |
LR outperforms other classifiers in terms of precision, recall, miss rate, F1 score, and AUC. Table 9.8 shows that for the case of cv, XGBoost and LR have high testing accuracies of 92.67% and 92.58%, respectively, whereas for the case of holdout (20% testing), LR has the highest accuracy value of 94.06%.

According to WHO, as of December 2020, over 200 vaccine candidates are being developed for fighting against COVID-19. Of these, at least 52 candidate vaccines are in human trials. Scientists across the world are trying to find appropriate drugs. Table 9.10 lists the pharmacologic parameters of different therapies for COVID-19. Some promising drug targets include nonstructural proteins and viral entry and immune regulation pathways. Nonstructural proteins such as 3-chymotrypsin-like protease, papain-like protease, and RNA-dependent RNA polymerase share homology with other novel coronaviruses.

According to the report in Wu [44], so far, the S protein in the genome of SARS-COV-2 is the major target for COVID-19 vaccine development [44]. Major vaccine candidates in development for prevention of COVID-19 are listed in RAPS [45] and https://www.clinicaltrials.gov/ct2/show/NCT04283461 [46]. Fig. 9.8 lists the five most active vaccine candidates for COVID-19 as of Apr. 23, 2020, as reported by the WHO.

### Table 9.10  Pharmacology for select proposed coronavirus disease 2019 treatment.

| Agent | Target | Special case |
|-------|--------|-------------|
| Chloroquine phosphate (Aralen/generic) [21–26] | The target is to block the virus by inhibiting glycosylation of host receptors, proteolytic processing, and endosomal acidification. The other target is to consider additional immunomodulatory effects. | This can be applied in pregnancy if risks are less compared with prospective benefits. |
| Hydroxychloroquine sulfate (Plaquenil/generic) [21–23], [25–32] | Same mechanism as chloroquine | Can be used for pregnant women if benefits are more than risks |
| Lopinavir/ritonavir (Kaletra) [33–38] | 3CL protease | May be applied pregnant women; oral solution is to be avoided because it contains ethanol |
| Remdesivir [39–41] | RNA polymerase inhibitor | Should not be used in case of pregnancy |
| Favipiravir [42,43] | RNA polymerase inhibitor | Must be avoided during pregnancy because metabolite has been found in breast milk |
7. Conclusion

This chapter provides an overview of the spread of COVID-19. The United States and some European countries such as Italy, Spain, the United Kingdom, and Germany are heavily affected by the disease. This chapter uses machine learning algorithms to predict COVID-19 for a given dataset. For this particular dataset, our experimental results indicate that serum glucose is the most influential attribute in predicting COVID-19. Our results also show that for the case of cv, XGBoost has the highest accuracy value of 92.67% and LR has the second highest accuracy of 92.58%, whereas both XGBoost and LR have the same 93% value for precision, recall, and F1 score. For the case of the holdout method with 20% testing data samples, LR exhibits the highest testing accuracy of 94.06%. Hence, XGBoost and LR can be used to predict COVID-19.

FIGURE 9.8 Some of vaccine candidates for COVID-19.
The reliability of the diagnosis results presented in this chapter depends on the reliability of the dataset used. In future, with the availability of more reliable datasets, machine learning algorithms [51,52] should be applied to those new datasets to validate the effectiveness of the classifiers. Hybrid deep learning algorithms [47–49,53] can also be successfully applied in various chest X-ray or computed tomography image [50] datasets to detect COVID-19 patients.

References

[1] Y. Guo, Q. Cao, Z. Hong, et al., The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak – an update on the status, Mil. Med. Res. 7 (2020) 11, https://doi.org/10.1186/s40779-020-00240-0.
[2] R. Lu, X. Zhao, J. Li, P. Niu, B. Yang, H. Wu, et al., Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding, Lancet 395 (10224) (2020) 565–574.
[3] WHO, Coronavirus Disease (COVID-2019) Situation Reports, 2020. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports. (Accessed 23 April 2020).
[4] N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, et al., A Novel coronavirus from patients with pneumonia in China, 2019, N. Engl. J. Med. 382 (2020) 727–733.
[5] L.L. Ren, Y.M. Wang, Z.Q. Wu, Z.C. Xiang, L. Guo, T. Xu, et al., Identification of a novel coronavirus causing severe pneumonia in human: a descriptive study, Chin. Med. J. 1 (2020).
[6] J.F. Chan, S. Yuan, K.H. Kok, K.K. To, H. Chu, J. Yang, et al., A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster, Lancet 395 (2020) 514–523.
[7] Coronaviridae Study Group of the International Committee on Taxonomy of Viruses, The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2, Nat. Microbiol. 5 (2020) 536–544.
[8] K.K.-W. To, O.T.-Y. Tsang, C.C.-Y. Yip, K.-H. Chan, T.-C. Wu, J.M.-C. Chan et al., Consistent Detection of 2019 Novel Coronavirus in Saliva, Clinical Infectious Diseases, ciaa149. https://doi.org/10.1093/cid/ciaa149.
[9] M. Letko, A. Marzi, V. Munster, Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses, Nat. Microbiol. 5 (2020) 562–569.
[10] M.R.H. Mondal, S. Bharati, P. Podder, P. Podder, Data analytics for novel coronavirus disease, Inform. Med. Unlocked (June 15, 2020) 100374. https://doi.org/10.1016/j.imu.2020.100374.
[11] B.E. Young, S.W.X. Ong, S. Kalimuddin, J.G. Low, S.Y. Tan, J. Loh, et al., Epidemiologic features and clinical course of patients infected with SARS-CoV-2 in Singapore, J. Am. Med. Assoc. 323 (15) (2020) 1488–1494. https://doi.org/10.1001/jama.2020.3204. In this issue.
[12] https://www.kaggle.com/einsteindata4u/covid19. (Accessed 28 April 2020).
[13] https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data. (Accessed 23 April 2020).
[14] https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset. (Accessed 23 April 2020).
[15] T. Chen, C.G. Xgboost, A Scalable Tree Boosting System, 2016 arxiv e-prints. arXiv preprint arXiv: 1603.02754.
S. Bharati, P. Podder, R. Mondal, A. Mahmood, M. Raihan-Al-Masud, Comparative performance analysis of different classification algorithms for the purpose of prediction of lung cancer, in: A. Abraham, A. Cherukuri, P. Melin, N. Gandhi (Eds.), Intelligent Systems Design and Applications. 2018. Advances in Intelligent Systems and Computing, vol. 941, 2020, pp. 447–457.

S. Bharati, P. Podder, P.K. Paul, Lung cancer recognition and prediction according to random forest ensemble and RUSBoost algorithm using LIDC data, Int. J. Hybrid Intel1. Syst. 15 (2) (2019) 91–100.

S. Bharati, M.R. Robel, M.A. Rahman, P. Podder, N. Gandhi, Comparative performance exploration and prediction of fibrosis, malign lymph, metastases, normal lymphogram using machine learning method, in: International Conference on Innovations in Bio-Inspired Computing and Applications, Springer, Cham, December 16, 2019, pp. 66–77. https://doi.org/10.1007/978-3-030-49339-4_8.

M. Raihan-Al-Masud, M.R.H. Mondal, Data-driven diagnosis of spinal abnormalities using feature selection and machine learning algorithms, PLoS One 15 (2) (2020) e0228422.

S. Bharati, M.A. Rahman, P. Podder, Breast cancer prediction applying different classification algorithm with comparative analysis using WEKA, in: 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT), 2018.

D. Zhou, S.M. Dai, Q. Tong, COVID-19: a recommendation to examine the effect of hydroxychloroquine in preventing infection and progression, J. Antimicrob. Chemother. (2020). https://doi.org/10.1093/jac/dkaa114 dkaa114.

C.A. Devaux, J.M. Rolain, P. Colson, D. Raoult, New insights on the antiviral effects of chloroquine against coronavirus: what to expect for COVID-19? Int. J. Antimicrob. Agents (March 11, 2020). https://doi.org/10.1016/j.ijantimicag.2020.105938.

P. Colson, J.M. Rolain, J.C. Lagier, P. Brouqui, D. Raoult, Chloroquine and hydroxychloroquine as available weapons to fight COVID-19, Int. J. Antimicrob. Agents (March 4, 2020). https://doi.org/10.1016/j.ijantimicag.2020.105932.

National Health Commission and State Administration of Traditional Chinese Medicine, Diagnosis and treatment protocol for novel coronavirus pneumonia. https://www.chinalawtranslate.com/wp-content/uploads/2020/03/Who-translation.pdf. (Accessed 23 April 2020).

Chloroquine [Database Online], Lexicomp Inc, Hudson, OH, 2016. http://online.lexi.com. (Accessed 23 April 2020).

Aralen (Chloroquine Phosphate) [package Insert], Sanofi-Aventis, Bridgewater, NJ, 2008. https://www.accessdata.fda.gov/drugsatfda_docs/label/2018/006002s045lbl.pdf. (Accessed 17 March 2020).

X. Yao, F. Ye, ZhangM, et al., In vitro antiviral activity and projection of optimized dosing design of hydroxychloroquine for the treatment of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), Clin. Infect. Dis. (March 9, 2020). https://doi.org/10.1093/cid/ciaa237.

P. Gautret, J.C. Lagier, P. Parola, et al., Hydroxychloroquine and azithromycin as a treatment of COVID-19: results of an open-label non-randomized clinical trial, Int. J. Antimicrob. Agents (March 20, 2020). https://doi.org/10.1016/j.ijantimicag.2020.105949.

J. Chen, D. Liu, L. Liu, et al., A pilot study of hydroxychloroquine in treatment of patients with common coronavirus disease-19 (COVID-19), J. Zhejiang. Univ. (Med. Sci). (March 6, 2020). https://doi.org/10.3785/j.issn.1008-9292.2020.03.03.

Hydroxychloroquine [Database Online], Lexicomp Inc, Hudson, OH, 2016. http://online.lexi.com. (Accessed 23 April 2020).

Plaquenil (Hydroxychloroquine Sulfate) [Package Insert], Concordia Pharmaceuticals Inc, St Michael, Barbados, 2018. https://www.accessdata.fda.gov/drugsatfda_docs/label/2019/009768Orig1s051lbl.pdf. (Accessed 23 April 2020).
Chapter 9 • Application of machine learning for the diagnosis of COVID-19 193

[32] H.S. Lim, J.S. Im, J.Y. Cho, et al., Pharmacokinetics of hydroxychloroquine and its clinical implications in chemoprophylaxis against malaria caused by Plasmodium vivax, Antimicrob. Agents Chemother. 53 (4) (2009) 1468–1475. https://doi.org/10.1128/AAC.00339-08.

[33] C.M. Chu, V.C. Cheng, I.F. Hung, et al., HKU/UCH SARS Study Group. Role of lopinavir/ritonavir in the treatment of SARS: initial virological and clinical findings, Thorax 59 (3) (2004) 252–256. https://doi.org/10.1136/thorax.2003.012658.

[34] A.H. de Wilde, D. Jochmans, C.C. Posthuma, et al., Screening of an FDA-approved compound library identifies four small-molecule inhibitors of Middle East respiratory syndrome coronavirus replication in cell culture, Antimicrob. Agents Chemother. 58 (8) (2014) 4875–4884. https://doi.org/10.1128/AAC.03011-14.

[35] B. Cao, Y. Wang, D. Wen, et al., A trial of lopinavir-ritonavir in adults hospitalized with severe COVID-19, N. Engl. J. Med. (March 18, 2020). https://doi.org/10.1056/NEJMoa2001282.

[36] Lopinavir/ritonavir [Database Online], Lexicomp Inc, Hudson (OH), 2016. http://online.lexi.com. (Accessed 17 March 2020).

[37] Kaletra (Lopinavir and Ritonavir) [Package Insert], Abbvie, North Chicago, IL, 2019. https://www.accessdata.fda.gov/drugsatfda_docs/label/2019/021226s048lbl.pdf. (Accessed 17 March 2020).

[38] J.M. Sanders, M.L. Monogue, T.Z. Jodlowski, J.B. Cutrell, Pharmacologic treatments for coronavirus disease 2019 (COVID-19): a review, J. Am. Med. Assoc. 323 (18) (2020) 1824–1836. https://doi.org/10.1001/jama.2020.6019.

[39] D. Siegel, H.C. Hui, E. Doerfler, et al., Discovery and synthesis of a phosphoramidate prodrug of a pyrrolo[2,1-f][triazin-4-amino] adenine C-nucleoside (GS-5734) for the treatment of Ebola and emerging viruses, J. Med. Chem. 60 (5) (2017) 1648–1661. https://doi.org/10.1021/acs.jmedchem.6b01594.

[40] J.A. Al-Tawfiq, A.H. Al-Homoud, Z.A. Memish, Remdesivir as a possible therapeutic option for the COVID-19, Trav. Med. Infect. Dis. (March 5, 2020). https://doi.org/10.1016/j.tmaid.2020.101615.

[41] T.P. Sheahan, A.C. Sims, S.R. Leist, et al., Comparative therapeutic efficacy of remdesivir and combination lopinavir, ritonavir, and interferon beta against MERS-CoV, Nat. Commun. 11 (1) (2020) 222. https://doi.org/10.1038/s41467-019-13940-6.

[42] F.G. Hayden, N. Shindo, Influenza virus polymerase inhibitors in clinical development, Curr. Opin. Infect. Dis. 32 (2) (2019) 176–186. https://doi.org/10.1097/QCO.0000000000000532.

[43] Avigan (Favipiravir) [package Insert], Taisho Toyama Pharmaceutical Co Ltd, Tokyo, Japan, 2017, 4th version. (Accessed 25 March 2020).

[44] S.-C. Wu, Progress and concept for COVID-19 vaccine development, Biotechnol. J. 15 (2020) 2000147. https://doi.org/10.1002/biot.202000147.

[45] RAPS, Regulatory Focus, COVID-19 Tracker. https://www.raps.org/news-and-articles/news-articles/2020/3/covid-19-vaccine-tracker. (Accessed 23 April 2020).

[46] https://www.clinicaltrials.gov/ct2/show/NCT04283461. (Accessed 23 April 2020).

[47] S. Bharati, P. Podder, M.R.H. Mondal, Hybrid deep learning for detecting lung diseases from X-ray images, Inform. Med. Unlocked (July 4, 2020) 100391. https://doi.org/10.1016/j.imu.2020.100391.

[48] S. Bharati, P. Podder, M. Rubaiyat Hossain Mondal, Artificial neural network based breast cancer screening: a comprehensive review, Int. J. Comput. Inform. Syst. Ind. Manag. Appl. 12 (2020) 125–137.

[49] Aditya Khamparia, Subrato Bharati, Prajoy Podder, Deepak Gupta, Ashish Khanna, Thai Kim Phung, Dang N.H. Thanh, et al., Diagnosis of breast cancer based on modern mammography using hybrid transfer learning, Multidimensional Systems and Signal Processing (2021). https://doi.org/10.1007/s11045-020-00756-7.
[50] Dang N.H. Thanh, V.B. Surya Prasath, Minh Hieu Le, A Review on CT and X-Ray Images Denoising Methods, Informatica 43 (2) (2019) 151−159. https://doi.org/10.31449/inf.v43i2.2179.

[51] Subrato Bharati, Prajoy Podder, M. Rubaiyat Hossain Mondal, Diagnosis of Polycystic Ovary Syndrome Using Machine Learning Algorithms, 2020 IEEE Region 10 Symposium (TENSYMP) (2020) 1486−1489. https://doi.org/10.1109/TENSYMP50017.2020.9230932.

[52] Prajoy Podder, Subrato Bharati, M. Rubaiyat Hossain Mondal, Automated gastric cancer detection and classification using machine learning, in: Deepak Gupta, Utku Kose, Bao Le Nguyen, Siddhartha Bhattacharyya, Bao Le Nguyen (Eds.), Artificial Intelligence for Data-Driven Medical Diagnosis, De Gruyter, Berlin, Boston, 2021, pp. 207−224.

[53] Subrato Bharati, Prajoy Podder, Performance of CNN for predicting cancerous lung nodules using LightGBM, in: Deepak Gupta, Utku Kose, Bao Le Nguyen, Siddhartha Bhattacharyya, Bao Le Nguyen (Eds.), Artificial Intelligence for Data-Driven Medical Diagnosis, De Gruyter, Berlin, Boston, 2021, pp. 1−18.