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.
Predictive models for COVID-19 detection using routine blood tests and machine learning

Yury V. Kistenev, Denis A. Vrazhnov, Ekaterina E. Shnaider, Hala Zuhayri

Laboratory of Laser Molecular Imaging and Machine Learning, Tomsk State University, 36 Lenin Av., 634050 Tomsk, Russia

ABSTRACT

The problem of accurate, fast, and inexpensive COVID-19 tests has been urgent till now. Standard COVID-19 tests need high-cost reagents and specialized laboratories with high safety requirements, are time-consuming. Data of routine blood tests as a base of SARS-CoV-2 invasion detection allows using the most practical medicine facilities. But blood tests give general information about a patient’s state, which is not directly associated with COVID-19. COVID-19-specific features should be selected from the list of standard blood characteristics, and decision-making software based on appropriate clinical data should be created. This review describes the abilities to develop predictive models for COVID-19 detection using routine blood tests and machine learning.

1. Introduction

The problem of accurate, fast, and inexpensive COVID-19 tests has been urgent till now. The key message of this review is that blood tests contain enough information to detect SARS-CoV-2 infected patients with acceptable accuracy. To establish the corresponding latent dependencies between blood test parameters and COVID-19 presence, a machine learning approach should be applied. The most effective prediction models with accuracy of COVID-19 diagnosis of about 85–95% are based on using Random forest, Gradient boosting, and their variations. Surprisingly, that these models use not more than only 8–10 blood parameters. A set of the most informative features includes mean corpuscular hemoglobin, lymphocytes, leukocytes, basophils, eosinophils, C-reactive protein, bilirubin, D-dimer.

Current “gold standard” for SARS-CoV-2 infection identification is based on a transcription-polymerase chain reaction (RT-PCR) and Enzyme-linked immunosorbent assay (ELISA) tests [1, 2]. RT-PCR has high specificity but low sensitivity [3, 4] and high misclassification in the early stage of disease [5]. The drawbacks are high-cost equipment and consumables, long time of analysis. Additional information can be achieved by computer tomography or X-ray visualization [6, 7, 8], but these methods require high-tech equipment and cannot be used often. A
possible solution is to use more conventional clinical tests combined with sophisticated but powerful analytical methods from the field of Artificial Intelligence implemented as a Medical Decision Support System (MDSS). The attractiveness of this idea is that Artificial Intelligence can compensate a nonspecificity of general clinical tests.

Since the COVID-19 outbreak at the end of 2019, several Medical Decision Support Systems (MDSSs) were developed for this disease diagnostic and severity prediction [9, 10, 11, 12, 13]. MDSSs can be based on data-driven, knowledge-driven, and hybrid (a combination of knowledge driven and data-based) data models. Data driven models (DDMs) rely only on the observed data, while knowledge-driven (KDM) ones use human logic. The hybrid models take advantage of both approaches. For example, DDMs are superior in natural language processing and can extract knowledge from Electronic Health Records, and KDM can be used for the support of disease diagnostics. Usually, DDM MDSSs use machine learning (ML) algorithms like artificial neural networks; KDM MDSSs are based on an expert knowledge and human logic [14]. Most trustworthy are expert systems based on subjective evaluations described as medical treatment protocols. Such systems are the state-of-the-art for a well-known disease [15] but lack the knowledge with COVID-19. ML MDSSs are available to analyze latent relations in data and construct complex predictive models, but this is a purely data-driven approach and couldn’t handle data uncertainty.

ML pipeline is a sequence of algorithms that includes preprocessing, feature extraction/selection, prediction model construction, and its validation (see Figure 1) [13]. A brief description of these steps is presented below.

1.1. Preprocessing

The necessity of preprocessing depends on the quality of input data. Routine blood test results often have missing values, and imputation algorithms based on the statistical mean are used to fill it. Also, detection and removal of outliers, presenting data, significantly different from the major of the dataset, should be conducted. The statistical-based approaches like a parametric Z-score [4, 16], isolation forest [17], density-based spatial clustering of applications with noise [18] are considered to be the most effective.

1.2. Feature extraction

The feature extraction is based on a transformation of initial parameters to a feature vector of less dimension due to removing redundant features. Principal component analysis and multidimensional scaling are the most frequently used [19, 20].

1.3. Feature selection

Feature selection is based on uninformative parameters removing. There are univariate and multivariate methods of feature selection [21, 22]. Univariate methods analyze every feature impact on the predictive model accuracy independently. Multivariate methods do the same for several features jointly. Some classification methods, like Decision trees (DT) and Random forests (RF) can also estimate features’ informativeness [23].

1.4. Prediction model construction

This step allows making a conclusion about a studied object belonging to one of the preliminarily determined classes. Supervised ML methods useful for creating prediction models in medicine are limited by the requirements of model transparency and a small volume of medical datasets [24]. DT, RF, Support vector machine (SVM), and gradient boosting are the most often used [24, 25].

1.5. Prediction model validation

The validation needs to estimate a prediction model accuracy. It should be conducted with a data set, which was not used for data model training. The latter most simple implementation is train-test split [19, 26]. Another option is to get a small standalone dataset from the other source and use it as a final performance measure [27]. K-fold cross-validation is often used [19].

Using routine blood tests as an additional diagnostic tool is very attractive because of its simplicity and availability in all medical institutions of practical healthcare. The routine blood tests issue is that they give general information about a patient’s state, which is not directly associated with COVID-19 presence or absence. COVID-19-specific features should be selected from the list of standard blood characteristics, and decision-making software based on appropriate clinical data should be created to overcome this issue. The latter can be implemented using ML.

This review aims to analyze routine blood test availability for COVID-19 detection and the ways of ML pipeline implementation using routine blood test data.

2. The blood test informative features selection

We analyzed about 50 articles published in journals in the field of medicine and machine learning and indexed in Scopus, Google Scholar, and Web of Science. The search included the following terms: COVID-19; Blood tests; Machine learning; Statistical analysis.

The features met in more than 10 papers and features considered by the authors as informative are presented in Table 1.

The oxygen desaturation is associated with severe respiratory failure in COVID-19 patients. SARS-CoV-2 surface glycoproteins bind to porphyrin on hemoglobin and inhibit heme metabolism. It is a possible mechanism of MCH and MCHC changing during COVID-19 [59].

Lymphopenia (temporary or persistent decrease in the level of lymphocytes in the blood) is a typical symptom for COVID-19 patients [60]. An essential decreasing CD8\(^+\) T cells, a type of lymphocytes, was established for SARS-CoV-2 infected patients compared to vaccinated subjects [61].

Eosinophils play an essential role in SARS-CoV-2 infection associated with high CD62L expression, a lung eosinophil marker [62]. The basophil
| Abb. | Feature          | Description                                                                                                                                                                                                 | Articles:                                                                                     |
|------|------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------|
|      |                  |                                                                                                                                                                                                           | features used                                                                                 |
|      |                  |                                                                                                                                                                                                           | informative features                                                                          |
| HCT  | Hematocrit       | The hematocrit is the calculated volume percentage of red blood cells (erythrocytes) in blood. HCT is a useful but low-specificity biomarker associated with number of diseases like thrombocytopenia, thrombosis, hypoxia. | [20] [29, 30, 31] [32, 33, 34] [35, 36, 37] [38, 39, 40] [41, 42, 43] [44, 45, 46] [47]          |
| HGB  | Hemoglobin       | Hemoglobin is a conjugated protein molecule, which handles oxygen carrier to tissues. Therefore, it is the key component of the red blood cells. Both low/high hemoglobin levels can indicate certain diseases. | [19, 26, 28] [29, 30, 31] [33, 34, 35] [36, 37, 38] [39, 40, 41] [42, 44, 45] [46, 47, 48] [49, 50, 51] [35, 51] |
| PLT  | Platelets        | Platelets are a component of blood whose function is to react to bleeding from blood vessel injury by clumping, that plays a major role in blood clotting. PLT reflects a pathology risk/severity, for example, cancer [87]. | [4, 23, 26] [28, 29, 30] [31, 32, 33] [35, 36, 37] [38, 39, 40] [42, 44, 45] [45, 46, 47] [50, 51, 52] [23, 28] [30, 38] [39, 51] |
| RBC  | Red blood cells  | Red blood cells are the most common type of blood cells that carry oxygen to the body tissues via blood flow through the circulatory system. Low/high RBC levels indicate health problems, for example liver disease. | [19, 26, 27] [28, 29, 30] [31, 33, 49] [35, 36, 37] [38, 42, 43] [46, 47, 53] [30, 35] [38] |
| LYM  | Lymphocytes      | White blood cells are a cellular component of the adaptive (specific) immune response: B-Lymphocytes produce antibodies; T-lymphocytes destruct the infected cells and regulate other leukocytes. Their concentration is low in severe COVID-19. | [4, 19, 26] [28, 29, 30] [31, 33, 34] [35, 36, 38] [39, 40, 41] [42, 44, 45] [46, 47, 49] [50, 51, 52] [54, 55, 56] [34, 39] [41, 44] [50, 51] [52, 55] [56, 57] |
| MCH  | Mean Corpuscular Hemoglobin | The mean corpuscular hemoglobin is the average mass of HGB per a RBC in a blood sample.                                                                                                                | [19, 26, 27] [28, 29, 30] [31, 32, 33] [34, 35, 36] [37, 46, 52] [53] [19] |
| MCHC | Mean Corpuscular Hemoglobin Concentration | The mean corpuscular hemoglobin concentration characterizes an average concentration of HGB in a RBC. MCHC also reflects the amount of hemoglobin per unit volume in a single red blood cell | [19, 26, 28] [29, 30, 31] [34, 35, 36] [37, 42, 46] [52] |
| WBC  | Leukocytes       | White blood cells include: granulocytes (neutrophils, eosinophils, basophils), monocytes, lymphocytes. WBC are the cells of the immune system involved in protecting the body against infectious disease. | [4, 19, 26] [27, 29, 29] [30, 31, 32] [33, 34, 35] [36, 37, 38] [39, 40, 41] [42, 43, 44] [45, 46, 47] [50, 51, 52] [53, 54, 55] [56, 57] [4, 26] [28, 30] [38, 39] [51, 52] [53, 56] [57] |
| BAY  | Basophils        | Basophils are a type of white blood cells. Basophils have granules containing biologically active substances, with the help of which they regulate immune responses. They handle detection of an antigen and its presentation to other cells, stimulating a response on an inflammation. | [4, 19, 23] [26, 27, 29] [29, 30, 31] [32, 33, 34] [36, 42, 43] [46, 47, 49] [50, 52, 53] [23, 34] [42, 50] [52] |

(continued on next page)
| Abb. | Feature | Description | Articles: |
|------|---------|-------------|-----------|
| EOS  | Eosinophils | Eosinophils are a type of white blood cells, responsible for hypersensitive immune responses. They contain inflammatory mediators (prostaglandins, leukotrienes, platelet-activating factor, cytokines) and play a significant role in destroying foreign particles and removing toxic substances. | ([4, 19, 26], [39, 42, 50], [52, 53, 57]) |
| LDH  | Lactate dehydrogenase | Lactate dehydrogenase is an important enzyme, which involves in cellular respiration and energy production. Also, LDH is a marker of cells and tissues damage. | ([4, 23, 26], [28, 29, 30], [31, 32, 37], [39, 40, 41], [44, 45, 46], [49, 50, 51], [53, 57]) |
| MCV  | Mean corpuscular volume | The average volume of an erythrocyte population | ([19, 26, 27], [28, 29, 30], [31, 33, 35], [36, 37, 39], [42, 43, 44], [53]) |
| RDW  | Red blood cell distribution width | Red blood cell distribution width characterizes the variability RBC size. RDW reflects the extent of anisocytosis if is elevated in iron deficiency anemia. | ([19, 26, 27], [28, 29, 30], [31, 33, 35], [36, 37, 42], [46, 50, 52]) |
| MONO | Monocytes | Monocytes are one of the largest among WBC, responsible for attacking and breaking down germs and bacteria that enter the body. Monocytes also influence adaptive immune responses and exert tissue repair functions. | ([4, 19, 26], [28, 29, 30], [31, 33, 34], [38, 39, 41], [42, 43, 49], [56]) |
| MPV  | Mean platelet volume | Mean platelet volume is a measure of the average size of platelets in blood. MPV high value can be observed with increased destruction of platelets and sepsis. | ([19, 26, 28], [29, 30, 31], [33, 34, 35], [42, 46, 52], [53, 57]) |
| NEU  | Neutrophils | Neutrophils are a type of white blood cells, being an immune system’s first line of defense against infections by ingesting microorganisms and releasing enzymes that kill them. | ([4, 19, 26], [28, 29, 30], [31, 33, 34], [38, 43, 44], [45, 46, 47], [49, 51, 56], [57]) |
| CRP  | C-reactive protein | C-reactive protein is produced by the liver and is induced by various inflammatory mediators, such as interleukin-6. In a healthy person, it is absent or present in minimal amounts. Therefore, it is an early marker of an acute phase of inflammatory response. | ([4, 19, 26], [28, 29, 30], [31, 32, 34], [35, 37, 39], [40, 43], [46, 47, 51], [58]) |
| CREAT | Creatinine | Creatinine is a chemical waste molecule resulting from muscle metabolism. It is formed mainly in the liver and excreted by the kidneys. | ([19, 23, 26], [28, 29, 30], [31, 34, 40], [41, 44, 45], [46, 47, 51], [58]) |
| UREA | Urea | Urea is a sub-product of the protein metabolism which is formed in the liver, excreted by the kidney, and is a standard biomarker of renal failure. | ([26, 28, 29], [31, 34, 38], [41, 44, 46], [47, 53, 57]) |
| K+   | Potassium | Potassium is an essential mineral that is needed by all tissues. It participates in the electrophysical processes in the cells, as well as in carbohydrate and protein metabolism, regulation of blood pressure. Potassium deficiency may be caused by increased renal excretion or its loss through the gastrointestinal tract. | ([19, 23, 26], [27, 28, 29], [34, 38, 40], [41, 44, 45], [47, 48, 58]) |

(continued on next page)
Table 1 (continued)

| Abb. | Feature | Description | Articles: |
|------|---------|-------------|----------|
| Na   | Sodium  | Sodium is the main electrolyte in the extracellular space, mainly responsible for osmotic pressure and electrolyte balance in the blood. | [19, 26, 27] [28, 30] [34, 38, 40] [41, 43, 44] [45, 47, 48] [56] |
| AST  | Aspartate transaminase | Aspartate transaminase is an enzyme involved in amino acid metabolism. It is found mainly in the liver, heart, nervous tissue, skeletal muscles, and in smaller amounts in the kidneys, pancreas, spleen, and lung tissue. | [4, 19, 26] [27, 28, 29] [30, 31, 32] [37, 40, 41] [43, 44, 45] [47, 51, 53] [56] |
| ALT  | Alanine transaminase | Alanine transaminase is an endogenous enzyme that commonly measured clinically as biomarker of liver injury. It is found in smallest amounts in the kidneys, heart, skeletal muscles, pancreas, therefore, it may indicate diseases of these organs. | [4, 19, 26] [28, 29, 30] [31, 34, 37] [46, 41, 43] [44, 45, 47] [51, 53] [37, 53] |
| BR   | Bilirubin (Total) | Bilirubin is an intermediate product of hemoglobin metabolism, formed during the normal process of breaking down red blood cells. Higher than normal levels of bilirubin may be associated with liver, bile duct or gallbladder problems. | [19, 23, 26] [27, 29, 30] [31, 32, 34] [46, 43, 44] [45, 46, 51] [54, 59] [23, 29] [34, 51] |
| XDP  | D-dimer | D-dimer is a plasmatic protein essential for dissolving blood clots. D-dimer is considered as a thrombosis marker, its high level indicates increased blood clotting. | [26, 30, 32] [37, 40, 44] [45, 46, 51] [57, 58] [26, 37] [51, 58] |
| Ferritin | Ferritin is a complex consisting of a protein shell (apoferitin) and iron hydroxide. The function of ferritin is to bind free iron ions, neutralising its toxic properties and increasing its solubility. It is used in clinical medicine as an indicator of iron stores. The highest concentrations of ferritin are usually found in hepatocytes and reticuloendothelial cells. | [29, 31, 32] [37, 40, 43] [45, 49, 50] [55] [29, 32] [37, 49] [49, 50] [55] |

The most often used features include: Lymphocytes (LYM), Leukocytes (WBC), Mean corpuscular hemoglobin (MCH), Basophils (BAY), Eosinophils (EOS), C-reactive protein (CRP), Bilirubin, D-dimer.

3. The COVID-19 prediction models construction using blood test data

The dataset description used by the authors, number of selected/extracted informative features, and created COVID-19 prediction model performance are shown in Table 2.

level is correlated with humoral response to SARS-CoV-2 infection [62, 63].

CRP is usually produced by hepatocytes and, additionally, by macrophages in a place of inflammation [64]. The latter is strongly related to COVID-19 disease complications and mortality. A cytokine storm appearing at the most dangerous stage of COVID-19 pneumonia stimulates hepatocytes to produce CRP [65, 66]. CRP is considered an independent indicator of COVID-19 severity [64, 67, 68]. Aspartate transaminase is a statistically significant marker of pulmonary fibrosis caused by SarS-CoV-2 infection [69]. SARS-CoV-2-positive patients (positive test for SARS-CoV-2) showed a statistically significant decrease in calcium concentration compared to SARS-CoV-2-negative patients [70]. High levels of ferritin, as well as severe lymphopenia (decrease in T-, B – lymphocytes), also high levels of C-reactive protein, ferritin, D-dimer, ALT and AST, are signs of a cytokine storm [71]. High ferritin levels were found in autopsies of deceased COVID-19 patients [72].

A high bilirubin content was established in severe Covid-19 patients compared to milder forms (an averaged difference was in interval from 0.27 to 0.95 μmol/L) [73, 74, 75].

As mentioned above, D-dimer is a biomarker of thrombosis. A 3–4 fold D-dimer concentration increasing was associated with poor prognosis [76].

Accuracy, sensitivity (SE), and specificity (SP) are defined as follows. Accuracy reflects how much the prediction is exact [36]:

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)
\]

The sensitivity is an ability of a method identifying true positives, and specificity (SP) is an ability to identify true negatives [44, 49]:

\[
SE = \frac{TP}{TP + FN} \quad (2)
\]

\[
SP = \frac{TN}{TN + FP} \quad (3)
\]

Here:
- True Positive (TP): Participants with a target disease classified as belonging to the target group;
- False Positive (FP): Healthy participants but classified as belonging to the target group;
- True Negative (TN): Healthy participants classified as belonging to the control group;
- False Negative (FN): Participants with the target disease but classified as belonging to the control group.

Machine learning methods met the most papers include Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT), Boosting, Artificial Neural Network (ANN), Logistic regression (LR) (see Figure 2). The popularity of DT-based methods, including Random Forest (RF), Regression DT, etc., is associated with the explainability of the
| Database | Positive cases | Methods | Features | The best method | SP, % | SE, % | Accuracy, % | Refs. |
|----------|----------------|---------|----------|------------------|-------|-------|-------------|-------|
| 279 patients | 126 positive cases | ET, SVM, LR, NB, RF, KNN, TWRF, DT | Used: 12 | RF | 65 | 75 | 92 | 82 | [1] |
| 535 patients | 180 positive cases | KNN, RF, SVM, XGBoost | All features: 117 | XGBoost | 97.9 ± 0.4 | 81.9 ± 6 | [19] |
| 105 patients | RF | All features: 49 | Used: 11 | 94 | 100 | [23] |
| 1st dataset | 2183 positive cases | LR, NB, KNN, SVM, RF | All features: 72 | 1) OSR: RF, SVM 2) COVID: KNN, SVM 3) CBC: KNN, RF | 86, 89 | 89, 83 | 82, 84 | [26] |
| 2nd dataset | 608 positive cases | DTX, LR, XGBoost, RF, SVM SMOTE, MLP | All features: 111 | RF | 91 ± 2 | 66 ± 10 | 88 ± 2 | [28] |
| 564 patients | 257 positive cases | ANN, DT, PLS-DA, KNN | All features: 75 | ANN | 94 | 93 | 94 | [29] |
| 1st data set: 279 patients | 739 positive cases | LR, KNN, DT, SVM, NB, ET, RF, XGBoost, DNN, CNN, RNN, LSTM | Used: 52 | DNN | 84.56 | 93.02 | 95.27 | 96.14 | 93.16 | 93.33 | [30] |
| 2nd data set: 1624 patients | 886 positive cases | SMOTE, RF, KNN, GNB, Ridge, Elastic net, DT, ET, AdaBoost, SMOTE, SV-LAR (ensemble based on LR, RF and Adalboost) | Used: 48 | SV-LAR | 91 | 83 | 91 | [31] |
| 5644 patients | 558 positive cases | LR, RF, KNN, GNB, Ridge, Elastic net, DT, ET, AdaBoost, SMOTE, SV-LAR (ensemble based on LR, RF and Adalboost) | Used: 48 | GB | 86.8 | 82.4 | 86.4 | [32] |
| 92,254 patients | 7335 positive cases | LR, RF, KNN, GNB, Ridge, Elastic net, DT, ET, AdaBoost, SMOTE, SV-LAR (ensemble based on LR, RF and Adalboost) | Used: 48 | GB | 86.8 | 82.4 | 86.4 | [33] |
| 294 patients | 111 patients | SMOTE, DT, RF, KNN, SVC, GB, GNB, MLP, GP | All features: 111 | GB | 94 | 93 | 94 | [29] |
| 405 patients | 212 positive cases | EBT | Used: 22 | EBT | 76.65 | 85.85 | 81.79 | [34] |
| The initial dataset had 9004 records. | RF, LR, ANN, SVM, KNN, Ensemble. | Used: 11 | KNN, kStar, and RF | 86.8 | 82.4 | 86.4 | [35] |
| 5644 patients | 558 positive cases | SVM, Adalboost, RF, KNN, Ensemble | Used: 15 | SVM, Adalboost, RF, KNN, Ensemble | 86.8 | 82.4 | 86.4 | [36] |
| 305 patients | 421 positive cases | CatBoost, SVM, LR, Radiologist + ML model | All features: 13 | CatBoost, SVM, LR, Radiologist + ML model | 91.5–98.3 | Radiologist + ML model | 52.7–66.7 | ML | 89.3–96.9 | [39] |
| 250 positive cases | SIMPLS | Used: 32 | SIMPLS | 86.8 | 82.4 | 86.4 | [40] |

(continued on next page)
**Table 2 (continued)**

| Database | Methods | Features | The best method | SP, % | SE, % | Accuracy, % | Refs. |
|----------|---------|----------|-----------------|-------|-------|-------------|-------|
| 2034 patients | DT, RF, GB, XGBoost, SVM, LGBM, KNN, ANN | Used: 59 Informative: 3 | 1) RF and GB 2) RF, LGBM 3) RF, LGBM, XGBoost | SP, % | SE, % | Accuracy, % | Refs. |
| Total cases | Positive cases | | 1) 89, 89 2) 93, 91 3) 91, 93, 87 | | | | | |
| 1218 patients | XGBoost, LR, SVM, RF, DT, Ensemble | Used: 15 Informative: 3 | DT | SVM | 91, 93 | | | | |
| 137 patients | RF, SVM, KNN, AdaBoost | Used: 30 Informative: 7 | SVM | | 81 | | | | |
| 422 patients | RF, NB, SVM, KNN, LR, ANN | Used: 38 Informative: 6 | NB | | 75.0 | 85.9 | | | |
| 162 patients | CRT, RF, ANN | Used: 25 | CRT 92.7 RF 95.8 ANN 96.3 | | | | | | |
| 5644 patients | 559 positive cases | MLP, SVM, RF, DT, Bayes Net, NB, RF | All: 107 Used: 41 | RF | 92.1 ± 1.2 | 93.6 ± 1.1 | 92.891 ± 0.851 | | | |
| 5644 patients | CNN, DT, NB, KNN, All features: 111 Used: 18 | CNN | | | 80 | | | | |
| 51 positive cases | HHOSRL | Used: 20 Informative: 4 | | | 100.00 | 100.00 | 100.00 | | | |
| 1455 patients | 182 positive cases | RF, LR, SVM, MLP, SGD, XGBoost, AdaBoost, Ensemble. | | | 64 (95 confidence interval 0.59–0.69) | 93 (95 confidence interval 0.84–0.98) | | | |
| 127,115 patients | 1573 positive cases | SNN, KNN, LR, SVM, RF, XGBoost | All features: 100 Used: 19 Informative: 11 | XGBoost RF | | | | | | |
| 412 patients | 326 positive cases | RF, SVM | Used: 19 Informative: 11 | SVM | 80 | 88 | 84 | | | |
| 1500 patients | LR, EMB, RF, SVM, SHAP | Used: 16 Informative: 4 | EBM RF | | (70–90) | | | | | | |
| 5644 patients | 559 positive cases | MLP, SVM, DT, Bayes Net, NB, RF | Used: 24 Informative: 6 | Bayes Net | 93.6 ± 1.1 | 96.8 ± 0.7 | 95.159 ± 0.693 | | | |
| 88 positive cases | SVM, XGBoost, RF, GB, DT | In: 3 Informative: 3 | Subsemble | 72 | 85 | | | | | | |
| 300 patients | 87 positive cases | LR, SVM, XGBoost | Used: 8 Informative: 6 | XGBoost | 69 | 94 | 87 | | | |
| 196 positive cases | MCDM | Used: 6 Informative: 3 | | | 82 | | | | | | |
| 404 patients | XGBoost | Used: 20 Informative: 3 | | | 90 | | | | | | |
| 212 patients | RF, MLP, SVM, GB, ET, AdaBoost | Used: 17 Informative: 5 | RF | | 74 | 73 | | | | | | |
| 250 patients | 126 positive cases | GB, MLP, KNN, DT, ET, LR, AdaBoost, SVM, RF, | All features: 100 | Subsemble | 98.6 | | | | | | |
| 117 positive cases | SVM-P, SVM-R, KNN, NNET, NB, RF | Informative: 6 | RF | | 92 | 93 | 92 | | | | | | |
| 4009 COVID-19 patients | 489 severe and 3520 non-severe cases | MLP, DT | Used: 30 Informative: 6 | MLP | 96.4 | 96.5 | | | | | | |

Remarks and abbreviations.
- AdaBoost: Adaptive Boosting,
- ANN: Artificial Neural Networks,
- CatBoost: Categorical gradient boosting,
- CNN: Convolutional Neural network,
- CRT: Classification and Regression Decision Tree,
- DNN: Deep Neural Network,
- DT: Decision Trees,
- DTX: Decision Tree-based Explainer,
- EBT: Ensemble Bagged Tree model,
- EMB: Explainable Boosting Machine,
- ET: Extremely Randomized Trees,
created data model and excellent accuracy. In 9 cases, the DT-based methods demonstrated the best prediction accuracy. A RF-based demonstrated COVID-19 diagnosis accuracy of 92.9% [76].

The SVM generalization potential and good performance are the reasons of its frequent use [26, 31, 37, 43, 51]. The Boosting family is in the third place in popularity and in the second place of prediction accuracy (7 studies). The Logistic regression is effective for informative features’ extraction and robust to overfitting. Researchers are still using KNN and Naïve Bayes (NB), which were among the early appeared supervised learning methods and do not show high results. An exception is the Bayesian network with diagnosis performance of 95.2% [76]. Researchers are still aware of ANN and Deep Neural Network (DNN) usage because the number of features in raw data is small (tens) compared to a number of samples (hundreds), which is not suitable for these methods.

A predictive model accuracy depends on the data volume and quality. In the analyzed papers, the smallest dataset included 51 COVID-19 patients, the largest – 7335 ones [32]. The average number of COVID-19 patients was 919. The most large control group included 125,542 samples [50]. The average number of informative blood features was 8; the median was 6. The mean accuracy was 88%. The simplicity of the best predictive models is an argument that overfitting was avoided. A relation between the number of informative features and a predictive model accuracy is shown in Figure 3.

4. Discussion

Circulating blood contains the metabolic biomarkers of inflammatory processes in a human body, so the metabolic profile potentially can be used for COVID-19 diagnostics and severe prediction. It can be seen from

![Figure 2](https://example.com/figure2.png)  
*Figure 2. The frequency of various classification used methods.*
Table 1 that the most full profile contains up to a hundred metabolites. Removing redundant ones can significantly improve the robustness of the predictive models on the one hand and reveal the most informative blood features for diagnostic purposes.

Preprocessing of input data includes missing data imputations based on statistics. Feature extraction was not usually used. The used machine learning systems provide the following result: there is a definite relation between mean corpuscular hemoglobin, lymphocytes, leukocytes, basophils, eosinophils, C-reactive protein, bilirubin, D-dimer, and COVID-19 (see Table 1). The most effective prediction models were based on using Random forest, Gradient boosting, and their variations. The predictive model achieved the highest accuracy score with a relatively small dataset size can be a subject of overfitting. Datasets with over 500 samples demonstrate an accuracy of about 85–95%, that can be considered as very good (see Table 2). Such results should be deemed positive, and specific blood metabolites-oriented diagnostics may be used in the clinical practice.

Significant progress in blood data analysis became available due to open datasets, also cited in this paper (see Table 2). Though merging the small datasets into large ones can benefit data scientists, it is still a problem from the ethical and medical points of view. Federative machine learning (for instance, OpenFL project by Intel Corporation) can solve a problem with data security, but the medical protocol unification issue is persisting. Different machine learning pipelines and the absence of the standard validation strategy that proves the quality of the predictive data model diminish the trust in such results. The open-source software, which allows the reproduction of achieved results, is a key to high-quality data models.

5. Conclusion

PCR test is still the best choice for diagnosing SARS-COV-2 infection [80] but, despite its usefulness, it requires expensive laboratory equipment, highly qualified laboratory personnel, and the time between sampling and the result is relatively long. Other clinical and immunodiagnostic tests should also be taken into account when interpreting the results of RT-PCR tests for a more reliable diagnosis [81]. It is necessary to improve RT-PCR tests to decrease false-negative and false-positive results [82]. COVID-19 pandemic force rapid development of a new method to study disease metabolic pathways.

Routine blood tests, including microfluidic devices, are minimally invasive, minimizing the sample volume to one drop. From this perspective, proposed predictive models allow us to understand the connections among human blood metabolites, COVID-19 diagnosis and prognosis of severe. Such knowledge is not a clinic-ready system, but a step before it. Once the verified set of metabolites is found, a simple expert knowledge-based system can be constructed to be used routinely.

Some ideas for the future development of COVID-19 blood tests are as follows. Classes imbalance can lead to a biased predictive model. Synthetic Minority Over-sampling Technique allows generating additional data samples through variations of each blood feature [28, 31, 33]. A game theory-based Shapley value method can provide a reliable, feature extraction [83, 84] An ensemble learning based on combining several classification algorithms and generating a final prediction, for instance, by the voting procedure allows improving a predictive model performance [38, 54, 85, 86, 87, 88].

Another achievement of such studies is the proof that machine learning methods coupled with thoroughly developed protocols and large datasets can provide rapid and reliable techniques to generate fundamental knowledge about socially significant diseases. We encourage scientific groups to share their datasets and develop machine learning models because humanity can overcome the pandemic by uniting efforts.

The essential proofs, which may lead to a clinical trial and further introduction of such methods into the clinical practice, are improvements and conclusions by physiologists found in many reviewed papers.

Declarations

Author contribution statement

All authors listed have significantly contributed to the development and the writing of this article.

Funding statement

Prof. Yury V. Kistenev was supported by the Tomsk State University Development Programme (Priority-2030).


Data associated with this study has been deposited at www.kaggle.com.

References

1. K. Delibas, E. Czaczyk, D. Nidzworski; Detection methods of human and animal influenza virus—current trends, Biosensors (2018) 8, 94.
2. K.S. George, Diagnosis of influenza virus, in: influenza Virus, Y. Kawaoka, G. Neumann (Eds.), Methods in Molecular Biology, vol. 865, Humana Press, Totowa, NJ, 2012, pp. 53–69.
3. T. Ai, Z. Yang, H. Heo, C. Zhang, C. Chen, W. Lv, Q. Tao, Z. Sun, X. Xia, Correlation of chest CT and RT-PCR testing for coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases, Radiology 296 (2020) E52–E60.
4. D. Biniati, A. Campagnaro, D. Ferrari, M. Locatelli, G. Banfi, F. Cabitza, Detection of COVID-19 infection from routine blood exams with machine learning: a feasibility study, J. Med. Syst. 44 (2020) 135.
5. D. Ferrari, A. Motta, M. Srollo, G. Banfi, M. Locatelli, Routine blood tests as a potential diagnostic tool for COVID-19, Clin. Chem. Lab. Med. (CCLM) 58 (2020) 1059–1065.
6. F. Mohammad-Rahimi, M. Nadimi, A. Ghahreynia-Langeroodi, M. Taheri, S. Ghafouri-Fard; Application of machine learning in diagnosis of COVID-19 through X-ray and CT images: a scoping review, Front. Cardiovasc. Med. (2021).
7. A. Zargar Khuzani, M. Heidari, S.A. Shariat; COVID-Classifier: An automated machine learning model to assist in the diagnosis of COVID-19 infection in chest X-ray images, Sci. Rep. 11 (2021) 9887.
8. L. Husain, T. Nguyen, H. Li, A.A. Abbasi, K.J. Lone, Z. Zhao, M. Zaib, A. Chen, T.Q. Duong; Machine-learning classification of texture features of portable chest X-ray accurately classifies COVID-19 lung infection, Biomed. Eng. Online 19 (2020) 88.
9. M.A. Callejo-Leblé, R. Moreno-Luna, A. Del Cuvillo, I.M. Reyes-Tejero, M.A. García Villaran, M. Santos-Peza, J.M. Maza-Solano, D.I. Martín-Jimenez, J.M. Palacios-Garcia, C. Fernandez-Velez, et al.; Loss of smell and taste can accurately predict COVID-19 infection: a machine-learning approach, J. Clin. Pathol. 10 (2021) 570.
10. I. Arpaci, S. Huang, M. Al-Emran, M.N. Al-Kahi, M. Peng; Predicting the COVID-19 infection with fourteen clinical features using machine learning classification algorithms, Multimed. Tool. Appl. 80 (2021) 11143–11157.
11. Y. Wan, H. Zhou, X. Zhang; An interpretation architecture for Deep learning models with the application of COVID-19 diagnosis, Entropy 23 (2021) 2012.
12. A. Imamz, I. Posokhova, U. Masood, M.S. Riaz, K. Ali, C.N. John, M. Celik, F. Dadaser-Celik, A.S. Dokuz; Anomaly detection in temperature data with the application of COVID-19 diagnosis, Entropy 23 (2021) 204.
13. D. Cousineau, S. Chartier, Outliers detection and treatment: a review, Int. J. Psychol. Res. 3 (2010) 58–67.
14. F.T. Liu, K.M. Ting, Z.H. Zhou, Isolation forest, in: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, IEEE, Pisa, Italy, 2008, pp. 413–422.
15. M. Celik, F. Dadaser-Celik, A.S. Dokuz, Anomaly detection in temperature data using DBSCAN algorithm, in: Proceedings of the 2011 International Symposium on Innovations in Intelligent Systems and Applications, IEEE, Istanbul, Turkey, 2011, pp. 91–95.
16. M. Mukar, G. Guncar, T. Vovko, S. Podnar, P. Cernelec, M. Brvar, M. Zalaznik, M. Notar, S. Moekon, M. Notar, COVID-19 diagnosis by routine blood tests using machine learning, Sci. Rep. 11 (2021), 10738.
17. J.B. Tenenbaum, V. de Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (2000) 2319–2323.
18. T. Hsing, S. Attoor, E. Dougherty, Relation between Permutation-Test F Values and Classifier Error Estimates, 2003.
