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.

Smart Health 26 (2022) 100331

Contents lists available at ScienceDirect

Smart Health

journal homepage: www.elsevier.com/locate/smhl

A machine learning study of COVID-19 serology and molecular tests and predictions

Magdalyn E. Elkin, Xingquan Zhu *

Department of Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL 33431, USA

ARTICLE INFO

MSC:
92C50
92C60
92C55
68T05
68T50

Keywords:
COVID-19
Serology test
Molecular test
Symptoms
Classification
Machine Learning

ABSTRACT

Serology and molecular tests are the two most commonly used methods for rapid COVID-19 infection testing. The two types of tests have different mechanisms to detect infection, by measuring the presence of viral SARS-CoV-2 RNA (molecular test) or detecting the presence of antibodies triggered by the SARS-CoV-2 virus (serology test). A handful of studies have shown that symptoms, combined with demographic and/or diagnosis features, can be helpful for the prediction of COVID-19 test outcomes. However, due to nature of the test, serology and molecular tests vary significantly. There is no existing study on the correlation between serology and molecular tests, and what type of symptoms are the key factors indicating the COVID-19 positive tests.

In this study, we propose a machine learning based approach to study serology and molecular tests, and use features to predict test outcomes. A total of 2,467 donors, each tested using one or multiple types of COVID-19 tests, are collected as our testbed. By cross checking test types and results, we study correlation between serology and molecular tests. For test outcome prediction, we label 2,467 donors as positive or negative, by using their serology or molecular test results, and create symptom features to represent each donor for learning. Because COVID-19 produces a wide range of symptoms and the data collection process is essentially error prone, we group similar symptoms into bins. This decreases the feature space and sparsity. Using binned symptoms, combined with demographic features, we train five classification algorithms to predict COVID-19 test results. Experiments show that XGBoost achieves the best performance with 76.85% accuracy and 81.4% AUC scores, demonstrating that symptoms are indeed helpful for predicting COVID-19 test outcomes. Our study investigates the relationship between serology and molecular tests, identifies meaningful symptom features associated with COVID-19 infection, and also provides a way for rapid screening and cost effective detection of COVID-19 infection.

1. Introduction

In 2019, a novel coronavirus disease (COVID-19) caused by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) emerged in Wuhan City, China and quickly spread globally (Singhal, 2020). As of May, 2022, there has been over 515 million COVID-19 cases and over 6 million deaths worldwide (Worldometer, 2022). COVID-19 infection is mainly transmitted through aerosol droplets from coughing or sneezing from an infected persons to a non-infected person. Transmission also can occur from asymptomatic persons (Singhal, 2020). The incubation period of SARS-CoV-2 ranges from 2 to 14 days, those developing symptoms typically do so within 12 days of infection.

* Corresponding author.
E-mail address: xzhu3@fau.edu (X. Zhu).

https://doi.org/10.1016/j.smhl.2022.100331
Received 29 September 2022; Accepted 29 September 2022
Available online 20 October 2022

Similarly to other highly pathogenic human coronaviruses (hCoVs), such as Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV) and Middle East Respiratory Syndrome-related Coronavirus (MERS-CoV), SARS-CoV-2 primarily affects the respiratory system. Once infected, patients often experience similar symptoms, and research has estimated that one third of SARS-CoV-2 cases are asymptomatic (Oran & Topol, 2021). On the other hand, symptoms are also the first warning signs triggering individuals to seek for clinical test, and such correlation has shed light for early (fast) prediction of COVID-19 infection using symptoms.

The symptomology of COVID-19 varies, from minor symptoms to Acute Respiratory Distress Syndrome (ARDS), and can also be fatal (Yuki, Fujiogi, & Koutsogiannaki, 2020). Fever, cough, difficulty breathing and fatigue are commonly reported symptoms of COVID-19. Other symptoms may include nausea/vomiting, diarrhea, myalgia, sore throat and congestion/runny nose (Alimohamadi, Sepandi, Taghdir, & Hosamirudsari, 2020; for Disease Control & Prevention). Neurological symptoms, such as headache, dizziness, impaired consciousness, smell and/or taste dysfunction, are also commonly reported in COVID-19 subjects (Chen bibetal, 2021). Nevertheless, these symptoms can also be presented in other viral infections and respiratory diseases, such as MERS, SARS and influenza. Previously, research has shown that the order of symptoms may be used to distinguish between COVID-19, MERS, SARS and influenza (Larsen, Martin, Martin, Kuhn, & Hicks, 2020). While the order of the symptoms themselves may provide evidence for which respiratory disease is indicated, not all symptoms may be experienced for each case. Symptom occurrence and severity differ between different COVID-19 patients. More severe symptoms are indicative of more severe COVID-19 cases and certain symptoms can be associated with more severe cases. For example, the presence of gastrointestinal symptoms may be associated with higher hospital visit rate (Sudre et al., 2021).

When experiencing symptoms, it is recommended to receive confirmatory COVID-19 testing immediately. Mass diagnostic testing is necessary for containing COVID-19 outbreaks (Weissleder, Lee, Ko, & Pittet, 2020). Vaccinations, testing, contact tracing and quarantining positive persons are shown to be effective to stop the spread of COVID-19.

While molecular testing is the only accepted method of confirming COVID-19 infection, research has been conducted to predict COVID-19 infection. These types of models may benefit health care systems by understanding the risk of disease of patients and by identifying key factors associated with COVID-19 diagnosis. Predictive models may also aid in screening patients and identifying those that need isolation so as to prevent them from spreading the virus to healthy individuals.

1.1. Research questions and contributions

Motivated by the above observations, the goal of our study is to create a predictive model using easy-to-obtain symptom features, along with demographic features; such as number of days PSO (post-symptom onset), temperature, age, and gender, to accurately predict whether a COVID-19 test might be positive or not. The predictive modeling is complicated by many puzzling questions currently unanswered in the literature. Some of important research questions to be addressed in our study are summarized as follows:

- What are correlation between different types of COVID-19 tests, such as serology tests *vs.* molecular tests? Are they consistent in the test? If not, which ones are more or less consistent?
- What are easy-to-obtain symptom features possibly useful for the prediction of COVID-19? Which symptoms features are more informative and useful for prediction, and how accurate can a symptom based model make predictions?
- Given identified symptom features and samples, do different machine learning models' performance vary significantly in their prediction? Can we derive clinically transparent (interpretable) models for symptoms based prediction?

In our study, we use test results from 2,467 donors collected from Boca Biolistics, LLC to create a testbed. Combining symptoms and simple demographic information, we design a set of features for predictive modelings. Five machine learning models, including Random Forest, XGBoost, Logistic Regression, Support Vector Machine (SVM), and Neural Network, are used in our study for evaluation. Their performance are compared by using three performance metrics: Accuracy, F1-score, and AUC (Area Under the ROC, Receiver Operating Characteristic, Curve). The main contribution, compared to existing research in the field is summarized as follows:

- **COVID-19 test relationship:** Our testbed is unique in the sense that each donor has one or multiple tests, allowing us to study COVID-19 test relationship. Our study clearly shows correlation between different types of COVID-19 tests.
- **COVID-19 symptom features:** Our data source contains a set of easy-to-obtain but noisy symptom features. We design new way to narrow down noisy symptom features for clinical interpretation and predictive modeling. The study also shows discriminative power of different features in the prediction.
- **COVID-19 predictive modeling:** By using symptom and simple demographic features, combined with the testbed, we use five types of machine learning models to validate their performance for COVID-19 prediction.

The remainder of the paper is structured as follows. Section 2 reviews related work on computational methods for COVID-19 prediction, with a brief discussion on the difference between our research *vs.* existing work in the field. Section 3 discusses methods and approaches used in our study, including COVID-19 test types and data used in our study. Section 4 reports experiments and results, with a focus on relationship between different types of COVID-19 tests, the importance of symptoms features, and the COVID-19 prediction results using symptom features. A short discussion and the data availability are reported in Section 6, and we conclude the paper in Section 6.

Table 1
Summary of related work using machine learning and symptom data for COVID-19 infection and/or patient outcome. Symptom, demographic and activity features are denoted with (+) to indicate their low cost of acquiring. Physiology features are denoted with (+++) to represent their higher cost of acquisition. For research that utilizes more than one classification algorithm, the symbol * denotes the model that demonstrates the highest performance.

Paper	Symptoms (+)	Demographics/ Activity (+)	Diagnosis (+++)	Model	Prediction	Data	Derived Symptom Features
Zoabi et al. (2021)	Cough, Fever, Sore Throat,Headache, Shortness of Breath	Age, Gender, Contact with COVID Patient		Light GBM	RT-PCR Test	Israeli Ministry	Symptom Questionnaire
Iwendi bibetal (2020)	Unnamed Symptoms numbered 1–6	Age,Gender, Days PSO, City, Country, Visited/from Wuhan		Boosted Random Forest	Patient Death	Kaggle	Categorical encoding given symptoms
Ahamad bibetal (2020)	Fever, Runny Nose, Pneumonia, Cough, Lung infection, Diarrhea, Muscle Soreness	Age, Gender, Travel History, Isolation		Decision Tree, SVM, GBM, Random Forest, XGBoost*	Confirmed COVID Infection by Doctor	BDBC-KG-NLP/ COVID-19 tracker	String Matching unstructured text for keywords
Mei et al. (2020)	Fever, Cough, Temp., Cough with sputum	Age,Gender Exposure history	Chest CT, Laboratory Findings	CNN, MLP CNN+MLP*	RT-PCR Test	Medical Records	Symptoms from clinical data
Tostmann bibetal (2020)	Anosmia, muscle ache, Ocular Pain, Malaise, Runny Nose, Fatigue, Sore Throat, Cough, Common cold, Headache Fever, Shortness of breath Nausea, Sneeze, Diarrhea,	Age, Gender, Comorbidities, Profession		Lasso Regression	RT-PCR Test	Netherlands Healthcare Workers	Symptom Questionnaire
Quer et al. (2021)	Stomach ache, Sore Throat, Gastrointestinal, Fatigue, Body aches, Congestion, Neck Pain, Headache, Decreased taste/smell, Fever, Difficulty breathing	Gender, Age, Fitbit or Apple User	RHR, Sleep, Activity (steps)	Logistic Regression	COVID-19 Diagnostic test	Subjects self reported symptoms, tests, sensor data via DETECT app	Symptom Questionnaire
Menni et al. (2020)	Loss of Smell/Taste, Fatigue, Fever, Diarrhea, Chest pain, Stomach Pain, Skipped meals, Hoarse Voice,Delirium, Cough, Shortness of Breath	Gender, Age, BMI		Logistic Regression	RT-PCR test	Subjects self reported symptoms and tests via smartphone app	Symptom Questionnaire

2. Related work

Using computational approaches for COVID-19 prediction has been investigated in a handful of research studies, by using different type of information, such as symptoms, demographics, travel history, or computerized tomography (CT) scan.

Several research has created COVID-19 prediction models using symptom data (Ahamad bibetal, 2020; Iwendi bibetal, 2020; Mei et al., 2020; Menni et al., 2020; Quer et al., 2021; Tostmann bibetal, 2020; Zoabi et al., 2021). These models have been used to predict patient outcome (Iwendi bibetal, 2020); or COVID-19 diagnosis (Ahamad bibetal, 2020; Mei et al., 2020; Quer et al., 2021; Tostmann bibetal, 2020; Zoabi et al., 2021). The datasets used in these studies are often RT-PCR or PCR confirmed positive COVID-19 samples (Menni et al., 2020; Tostmann bibetal, 2020; Zoabi et al., 2021).

In Table 1, we summarize several related work using machine learning and symptom data to predict COVID-19 infection or patient outcomes. The columns Symptoms (+), Demographics/Activity (+) and Physiology (+++) list the features used in the research paper along with their cost of acquiring. Symptom, demographic and activity features represent a lower cost of acquisition as they can often be captured with a simple questionnaire. Physiology features include diagnostic factors, such as Chest CT and laboratory results, or physiological sensors. Sensor data can include resting heart rate (RHR), sleep, or physical activity (number of steps). Physiology features represent a high cost of acquisition. While the usage of Chest CT and laboratory findings have been shown to have high predictive power for COVID-19 infection, these can be costly to obtain and are not readily available for all patients presented for early screening. It was also shown that the best model combined a multi-layer perceptron (MLP) to analyze symptom and demographic data combined with a convolutional neural network (CNN) for Chest CT (Mei et al., 2020).

In addition to symptom features, a separate study has shown that sensor data can aid in the prediction of COVID-19 diagnosis, because sensor data provide additional monitoring of the health condition of each individual. The highest predictive performance was shown to include symptoms and sensor data. The sensor data used is commonly captured by common smartwatch devices, such as Fitbit and Apple watches (Quer et al., 2021). However, these devices are still costly, and the utilization of the data requires software applications to capture sensor information.

In addition to the COVID-19 prediction, we have has also previously proposed to use natural language processing and feature engineering to predict completion and cessation of COVID-19 clinical trials (Elkin & Zhu, 2021), by leveraging data collected from ClinicalTrial.gov, the largest clinical trials database.

In summary, the above studies demonstrate the high predictive power of using features, as well as physiological data, to model COVID-19 for prediction.

Table 2
Basic summary of diagnostic tests for COVID-19. Category (cost) indicates if the test is serology or molecular and the relative cost basis of the two categories. Serology assays have a general lower cost basis than molecular costs due to equipment and reagents that are involved in molecular assays. Sensitivity factors display the specific factors that have a role in the sensitivity of the diagnostic assay. Detection basis is what the diagnostic test is measuring the presence of. # Days PSO lists the median number of days after symptoms started required for detection with the diagnostic assay. Result indication lists the general interpretation of a positive COVID-19 test result.

Category (Cost)	Sample Type	Sensitivity Factors	Detection Basis	# Days PSO	Result Indication
Serology (++)	Serum	Seroconversion	IgA	4–6	Early Immune Response
			IgM	4–6	Early Immune Response
			IgG	5–10	Later Immune Response
Molecular (+++)	Swab/ Saliva	Site & quality of specimen; Viral Load	Viral RNA	0–4	Current Infection

In this study, we propose to combine easy-to-obtain symptoms and simple demographic information for COVID-19 prediction. Our research differs from the past research in the following ways:

- Our study is based on a unique dataset that has combined serology and molecular testing. Donors in the dataset have a variety of serology (IgG, IgA and IgM) and molecular testing or a combination of test results. To consolidate different test results, we create a binary label indicating if the sample has a positive COVID-19 test. The flexibility of using multiple diagnostic tests in such a model allows the model to access more samples.
- From the entire dataset, there are a total of 121 separate reported symptoms. As many of these symptoms are related to each other (for example, Congestion and Stuffy Nose), we create 26 "binned" symptom features that combine similar symptoms into a single grouping. This aids in decreasing the feature space and decreases features sparsity, while keeping many "uncommon" symptoms that may not have been captured previously, such as neurological symptoms. Since only a few samples may have a specific single feature, combining similar features into a bin will increase the relative number of samples with the binned feature.

3. Methods & approaches

3.1. COVID-19 diagnostic tests

There are two different main categories of COVID-19 testing, molecular and serology tests. Table 2 briefly summarizes two types of tests in terms of their costs, sensitivity, detection base, post symptom onset (PSO), and the indication of a positive result.

Molecular tests measure the presence of viral SARS-CoV-2 RNA. They can be done by polymerase chain reaction (PCR), reverse transcription polymerase chain reaction (RT-PCR), or transcription mediated amplification (TMA). These tests amplify the virus's genetic material for detection. Molecular tests are the most common method to detect SARS-CoV-2 (Weissleder et al., 2020). Molecular tests typically have high sensitivity (true positive rate) and specificity (true negative rate). However, the sensitivity rate is dependent on factors such as specimen collection and viral load (Weissleder et al., 2020). Previously it was shown that the highest virus detection rate using RT-PCR was 89% between 0–4 days post symptom onset (PSO), which drops to 54% after 10–14 days PSO (Mallett et al., 2020). False negative test results can be common if the infected individual is tested early in incubation period (Böger et al., 2021). The sensitivity of molecular tests are also dependent on site and quality of specimen collection. Molecular tests are commonly performed using nasal or throat swab or saliva from a given individual. PCR tests can be conducted with other sample types (*e.g.* rectal swabs, urine, plasma, blood), however the sensitivity of the test is highest with respiratory biospecimens (Böger et al., 2021).

Serology tests, such as enzyme-linked immunosorbent assays (ELISAs), chemiluminescent immunoassays (CLIA) and chemiluminescent microparticle immunoassay (CMIA), detect the presence of immunoglobulin G (IgG), IgM or IgA antibodies, which represent an immune response to SARS-CoV-2 (*i.e.* an indirect indication of infection) (Theel, Slev, Wheeler, Couturier, Wong, & Kadkhoda, 2020). The three antibodies have different roles in infection and seroconversion time. The sensitivity and specificity of serology tests are dependent on immune response to infection and seroconversion time (Weissleder et al., 2020). Previous research reported that IgM and IgA antibodies can be detected as soon as 1 day after symptom onset (Theel et al., 2020), while the median seroconversion time for IgA/IgM is 4–6 days after symptom onset (Ma bibetal, 2020). IgG antibodies have a larger role to play in the host immune response as they are associated with viral neutralization (Theel et al., 2020). IgG antibodies have a seroconversion time of 5–10 days after symptom onset (Ma bibetal, 2020). The presence/absence of certain antibodies represent different immune time responses to the infection. With IgA and IgM antibodies representing earlier infection, and IgG representing later infection. The levels of antibody may also reflect the severity of COVID-19 infection. Previous reports found that IgM and IgG levels were significantly higher in moderate and severe cases compared to mild cases; and IgA levels were significantly higher in severe cases than mild or moderate cases (Ma bibetal, 2020). Serology tests are commonly performed using serum from the blood sample of a given individual.

Because serology testing is an indirect measure of infection, molecular testing remains the primary method of verification of current infection of COVID-19 (Theel et al., 2020). However, serology testing can play an important role in COVID-19 diagnostics.

Table 3
Summary of positive samples and total number of samples tested per COVID-19 assay. t/d denotes that the numbers are based on tests/donors, respectively.

COVID-19 test	# Tested positive	# Donors^{d} (# Tests^{t})
IgA	187^{d} (187^{t})	247^{d} (247^{t})
IgM	212^{d} (212^{t})	473^{d} (473^{t})
IgG	997^{d} (997^{d})	1,738^{d} (1,738^{t})
Serology	1,052^{d} (1,396^{t})	1,831^{d} (2,458^{t})
Molecular	294^{d} (294^{t})	1,452^{d} (1,452^{t})
COVID-19	1,255^{d}	2,467^{d}

Serology tests represent easier, cheaper and rapid methods of testing for COVID-19 (Böger et al., 2021). This makes serology tests ideal in times of PCR test kit shortages in areas of high outbreak. Additionally, the window for detection is longer for serology tests, compared to molecular tests; thus serology tests can be beneficial in determining infection rates for larger communities (Böger et al., 2021). However, with mass testing communities, the prevalence of COVID-19 has a role in sensitivity and specificity rates. In populations with low disease prevalence, false positive rates may be higher compared to populations with high disease prevalence (Kumleben bibetal, 2020).

3.2. Data

Biomedical sample collection is essentially a complicated process, due to the interaction with human subjects. For ease of understanding, we define following three concepts related to our data collection.

Definition 1. Donor: A donor (d) is a human subject from whom biomedical specimens are taken for COVID-19 test. Each donor may contribute one or multiple specimens.

Definition 2. Specimen: A specimen (s) is a biomedical sample taken from a donor. A donor may contribute one or multiple specimens, in same or different types. For example, a nasal swab test will collect one type of specimen. A blood draw will collect a different type of specimen.

Definition 3. Test: A test (t) refers to a COVID-19 lab assay on a specific specimen. Depending on the nature of the testing mechanism, *e.g.* antibody or viral RNA, each test will return a numerical response value, from which the technician will determine whether the test is positive or negative, by comparing the output value to a reference value provided by the manufacturer.

The dataset used in our study consists of 2,467 clinically collected donors from Boca Biolistics, LLC. Donors were enrolled in collections for either having a confirmed COVID-19 molecular test, experiencing COVID-19 symptoms or other respiratory infection symptoms. At the day of sample collection from the donors, basic demographic information, such as age, gender, as well as easy-to-obtain symptoms, and date of symptom onset, are also collected through a simple questionnaire form. Donors may provide blood specimens (for serum specimens) and/or nasal swabs. Specimen collection may vary for donors.

After data collection, specimens are tested in a lab environment at Boca Biolistics, LLCto confirm COVID-19 infection. Serum specimens underwent serology testing (for IgA, IgM or IgG antibodies or a combination of the three); swab specimens underwent molecular testing. After testing, donors have serology and/or molecular test results available. The specimens tested are often tested on different test methods. For serology tests, methods include ELISA, CMIA or CLIA. For molecular tests, methods include RT-PCR, PCR, or TMA. Positive and negative test interpretations depend on the test method and reference values as specified by the test manufacturer. Donors may receive multiple tests in one grouping (such as IgG). For each donor, if any test of the donor is positive, the donor is labeled as positive, or negative otherwise. As a result, we create a binary indicator denoting COVID-19 positive or negative for each donor.

Table 3 reports tested numbers and tested positive numbers with respect to donors and tests, respectively. Each row in Table 3 lists the number of tested donors and number of tests for each testing type. For example, the first row shows that 247^{d} donors received IgA tests (so there are also 247 IgA tests: ^{t}). Among them, 187 donors are IgA positive (187^{d}).

For each type of test, such as IgA, IgM, IgG, and molecular test, we only collect one result for each donor. Therefore the number of donors and number of tests are the same for IgA, IgM, IgG, and molecular test, respectively. For serology tests (including IgA, IgM, and IgG), each donor may receive more than one type of tests. Therefore, the number of serology donors are far less than the sum of IgA, IgM, and IgG donors (same for the serology tests). The overlapping between IgA, IgM, IgG, and molecular tests are further detailed in Fig. 2.

For each donor, a serology positive label indicates that the donor has a positive serology test (positive for at least one of IgA, IgM, IgG). Likewise, molecular positive label indicates the donor has a positive molecular test. For each donor, if any of the serology test or molecular test is positive, we will label the donor as positive. In Table 3, the last row reports the number of donors (n = 2,467) and the number of tested positive donors (n = 1,255). Overall, the positive rate of the donors are 50.9% (1,255/2,467), so the class distributes are well balanced.

Table 4
Summary of binned symptom features. Group name shows a common medical construct to represent the feature, symptoms lists the individual symptoms in the grouping.

Group name	Symptoms
Chills	Chills, Shaking chills, Shaking, Rigors, Cold, Sweats, Night sweats
Sore Throat	Sore throat, Itchy and sore throat, Tightness in gland, Burning in the throat
Gastrointestinal	Diarrhea, Constipation, Bloated, Gas, Reflux, Stomach ache, Upset stomach
Abdominal pain	Abdominal pain, Abdominal cramps, Abdominal distention
Loss of Smell/Taste	Loss of smell and taste, Altered taste, Decreased taste, Loss of taste, No taste, Loss of smell, Anosmia
Cough	Cough, Dry cough
Muscle pain	Muscle pain, Muscle ache, Myalgia
Chest Pain	Chest pain, Chest pressure, Chest pain while coughing, Chest sensitivity, Chest tightness, Chest burning
Body pain	Body ache, Body pain, Joint pain, Leg pain, Aches, Back pain, Backache, Punches in the back
Fatigue	Fatigue, Body fatigue, Severe fatigue, Lingering fatigue, Tiredness
Anemia	Anemia, Weakness
Headache	Headache, Head pain
Nausea	Nausea, Nausea and vomiting, Vomiting, Nausea or vomiting
Congestion	Congestion, Nasal congestion, Stuffy nose, Nasal pressure, Sinus pressure, Nose bleeds, Discomfort in the nasal passage
Runny nose	Runny nose, Rhinorrhea, Postnasal drip, Excess mucus, Sneezing
Breathing	Shortness of breath, Difficulty breathing, Wheezing
Mental	Anxiety, Confusion, Hallucination, Insomnia, Memory loss
Appetite	Decreased appetite, No appetite, Poor appetite, Loss of appetite, Weight loss
White blood cells	Low white blood cells, Leukopenia, Elevated white blood cells
Discomfort	General discomfort, General malaise, Malaise
Rash	Rash, Rash on back and face, Rash on legs and chest, Shingles
Eye symptoms	Itching eyes, Eye pain, Eye irritation, Blurry vision
Ear pain	Ear pain, Fluid in ears
Dizziness	Dizziness, Light headed, Imbalance, Syncope episode
Other	Burning sensation, Itching, Numbness, Tingling in feet and toes, Dry mouth, Thirst, Tongue ulcers Acute encephalitis, Central uremia, Viremia, Pericarditis, Pneumonia of both lungs

Table 5
Summary of all features (COVID-19+ represents the predictive label) used for predictive modeling and feature statistics. Feature lists the feature name; Type indicates if the feature was binary or continuous; Count/Range shows the total number of samples with the feature (for binary features) or range of values for continuous features.

Feature	Type	Count/Range	Feature	Type	Count/Range	Feature	Type	Count/Range
COVID-19+	Binary	1,255	Gastrointestinal	Binary	334	Breathing	Binary	646
Age 0–19	Binary	27	Abdominal Pain	Binary	149	Mental	Binary	11
Age 20–34	Binary	512	Loss Smell/Taste	Binary	523	Appetite	Binary	19
Age 35–49	Binary	682	Cough	Binary	1,065	White Blood Cells	Binary	9
Age 50–64	Binary	828	Muscle Pain	Binary	489	Discomfort	Binary	508
Age 65–79	Binary	354	Chest Pain	Binary	98	Rash	Binary	6
Age 80+	Binary	64	Body Pain	Binary	154	Eye Symptoms	Binary	7
Female	Binary	1,277	Fatigue	Binary	278	Ear Pain	Binary	3
Male	Binary	1,190	Anemia	Binary	101	Dizziness	Binary	15
No Symptoms	Binary	284	Headache	Binary	1,304	Other	Binary	15
Fever	Binary	1,347	Nausea	Binary	328	Fever Temperature	Continuous	0-107
Chills	Binary	340	Congestion	Binary	344	Days PSO	Continuous	0-260
Sore Throat	Binary	939	Runny Nose	Binary	470			

In the predictive modeling part of our study, we create machine learning models to predict whether a donor is likely going to be tested positive, by using simple demographics and symptom features. A given sample in our dataset represents one donor, where features consist of symptoms and demographics; the predictive label is COVID-19+, which indicates that the donor is positive for at least one of IgA, IgM, IgG, or molecular.

3.3. Features for test outcome prediction

We use two types of features in our study: simple demographic features and easy-to-obtain symptom features. Demographic features include age, gender, and days PSO. Six binary features are created for age depending on if the sample's age fell within the

following categories: 0–19, 20–34, 35–49, 50–54, 54–79, 80 +. Two binary features are created for gender: Female and Male. Days PSO is calculated from date of symptom onset and date of specimen collection. For asymptomatic donors, days PSO is equal to 0.

Symptom features are created from symptom data field of the questionnaire form. The form provides standardized symptoms, but also allow donors to provide additional information, using simple keywords. This allows more valuable/accurate input, but also results in significant noise in the data. For example, some donors may describe chest pain as "chest burning". This is further complicated with inputs that may have typos and errors.

As a result of the data collection and processing, symptom data is stored in a string field with a list of symptoms separated by a semi-colon for each donor. For example, a donor's symptom description may be "Muscle aches; Chills; Nausea/Vomiting". The string is separated into tokens by splitting on punctuation. Token terms become individual symptoms. In this example, the donor has four symptom features: "Muscle Aches", "Chills", "Nausea", "Vomiting". The end result, after considering all donors' symptoms, is 121 separate symptom features. To decrease the symptom feature space, similar symptoms are placed into bins. For example, "Nausea", "Nausea and Vomiting", "Vomiting" and "Nausea or Vomiting" are all separate symptom features that are ultimately placed within a "Nausea" bin. A total of 26 binned features are generated. These are binary features, 1 indicates that the sample has at least 1 symptom within the corresponding bin; 0 otherwise. Table 4 lists all the grouped features and their corresponding symptoms.

Binned features ultimately group together symptoms that are the same, yet described differently, such as "Headache" and "Head pain". Binned features also group together symptoms that have a similar underlying medical construct, such as the grouping of eye symptoms. Some symptoms have very low occurrence and ambiguous connection to other bins. For example, "Acute Encephalitis", is a symptom listed for a single sample. The inclusion of this uncommon symptom may not provide additional useful information to distinguish between COVID-19+ and COVID-19-; however, it still represents a valid symptom. While COVID-19 related encephalitis may be rare, previously, it was shown that seven case reports of encephalitis have been published (Chen bibetal, 2021).

In order to include rare or uncommon features, the group "Other" was created. This group represents very uncommon features, often only present for a single sample, with no obvious placement into the other symptom groupings. The presence of these features demonstrate the wide variety of symptoms that may accompany COVID-19 infection. The uncommon presence of a symptom feature provides very little information for a classification model. However, grouping them together in the "Other" bin allow the models more instances of the uncommon features, which provides more information. This also greatly decreases the sparsity of the feature space. As with using symptoms alone, there are 12 highly uncommon features, however with using binned features, their presence is captured in a single feature.

There are two features to indicate a fever was present. Fever temperature indicates the recorded temperature of the donor at time of sample collection. The feature, "Fever Temperature" indicates the recorded temperature. In some samples, "Fever" was listed as symptom, however fever temperature was not indicated. A separate binary feature, "Fever" indicates 1 if the sample lists fever as a symptom or 1 if "Fever Temperature" is greater than 100° F; 0 otherwise. Lastly, a binary feature, "No Symptoms" is created where 1 indicates if the sample is not experiencing symptoms, 0 otherwise. For asymptomatic samples, all binned symptom features are set to 0.

As a result of the feature construction process, Table 5 lists features, feature type, and their count (for binary features) or range (for continuous features) used for the predictive modeling. The first item, COVID-19+, represents the predictive label.

3.4. Predictive models

In order to train classification models to predict whether a donor is COVID-19 positive or not, using simple demographic and symptom features, we use five machine learning algorithms, Random Forest, XGBoost, Logistic Regression, Support Vector Machine and Neural Network, in our study. The five models represent a balance of best-performance methods in the literature (such as Random Forest, XGBoost, SVM), models with good interoperability (such as Random Forest and Logistic Regression), and well sought-after Neural Networks. We did not use any deep learning methods because our data are in tabular format, and existing research has shown that deep learning models has limited improvement on tabular data (Han, Li, & Zhu, 2019).

3.4.1. Random forest

A Random Forest is an ensemble method built of randomized K decision tree classifiers, with each tree being created using a random subset of features. Decision tree models classify samples based on tree structured paths using features in the dataset. An example of a Random Forest classifier and its decision trees are shown in Fig. 1. Each feature shown in the decision tree and each unique color/intensity denotes different feature. Because trees are created using random subsets of features, they do not necessarily share common features. For each test instance, its prediction is generated by applying the instance to each single tree, and combine the predicted results from all trees to generate final prediction. The strength of the Random Forest is that it relies on a large number of trees (typically more than 100), each trained from a random subset of features, and the trees have transparent decision logic. As a result, Random Forest represents one of the most accurate machine learning models with transparent decisions, in the literature.

For our implementation, we set decision trees maximum depth as 4. A minimum of 2 samples are required for each split at an internal node. A minimum of 4 samples are required for each leaf node. The maximum number of features considered for the best split is $\log_2(m)$, where m is total number of features in the training dataset. The Gini criterion is used to determine the best split. The random forests consists of 200 decision trees. These parameters are chosen after completing an exhaustive gridsearch to optimize random forest AUC scores.

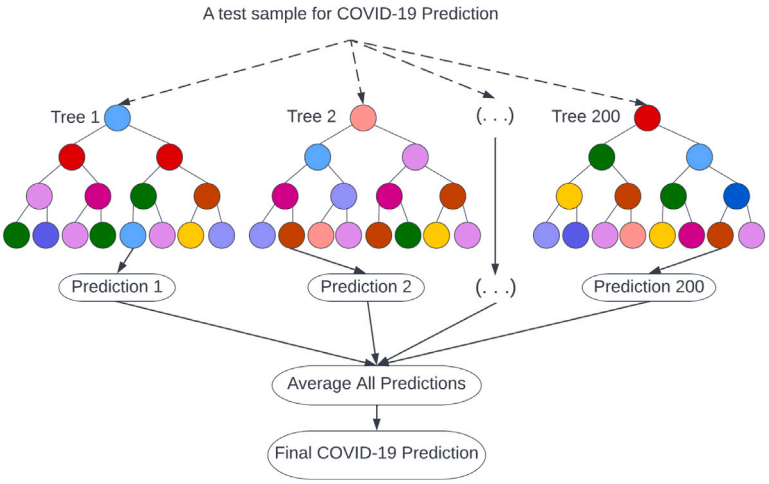


Fig. 1. A conceptual view of random forest and its prediction mechanism. The forest contains 200 trees, each tree is created using a subset of randomly selected features (each node color/intensity denotes a unique feature). For each test input, the predictions from all trees are combined to generate final prediction.

3.4.2. XGBoost

Extreme Gradient Boosting (XGBoost) is a decision tree ensemble algorithm with a gradient boosting framework (Chen & Guestrin, 2016). The decision tree ensemble uses K additive functions to predict output $\hat{y}_i$ for instance x_i.

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathcal{F} \tag{1}$$

Each decision tree f_k has an independent tree structure that utilize the decision rules to determine the final prediction by summing up scores, denoted by w in the leaves. In XGBoost, decision tree functions are learned by minimizing the objective function as defined in Eq. (2). Where l is a differentiable convex loss function to measure the difference between the predicted value, $\hat{y}_i$ and true value, y_i of instance x_i (Chen & Guestrin, 2016).

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{2}$$

The model complexity of tree f with T leaves is penalized by $\Omega(f)$, as defined in Eq. (3) (Chen & Guestrin, 2016). Splitting nodes in tree f increases the number of leaves and the complexity of the model. Thus nodes are split only if there is a positive reduction in loss function $\mathcal{L}(\phi)$. The amount of complexity cost for each additional leaf is controlled by γ. Increasing γ increases the complexity costs.

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{3}$$

For our implementation of XGBoost, we use $\gamma = 5$, the number of trees in ensemble is set to 100, and the subsample ratio of training instances used to build each tree is 0.6. The feature space subsampling ratio is 0.6, the maximum depth of each tree is 2, and the minimum weight in child nodes is 10.

3.4.3. Neural network

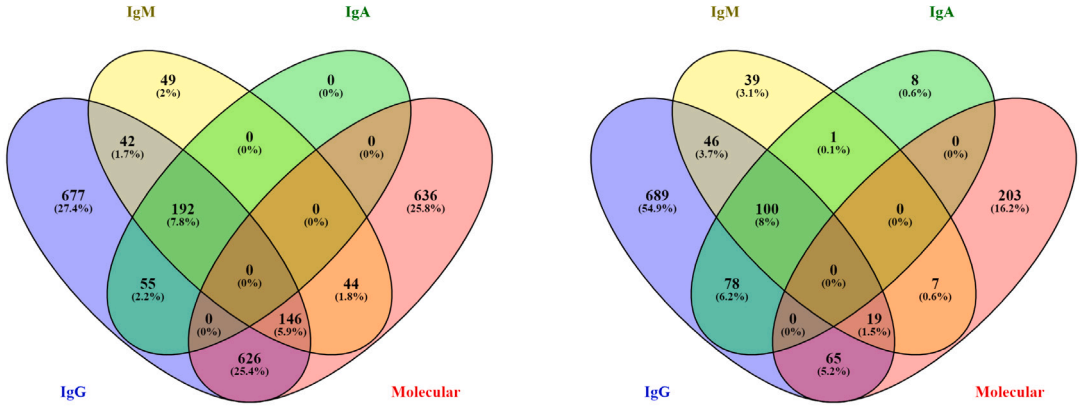
Our neural network implementation consists of a three layer feed forward network. The first layer is the input layer where features are feed in. The second layer is a hidden layer with 20 nodes and the third layer consists of the output layer with a single node, which produces a probability score of COVID-19 positive *vs* COVID-19 negative.

Nodes i in a hidden layer and output layer each connects to node j in the previous layer and has associated weight w_{ij} and bias input b_i. Nodes in hidden layer and output layer create their output a weighted sum of inputs, a_i, for all n nodes connected to node i, as defined in Eq. (4). The output value y_i is determined by applying the sigmoid function to a_i as defined in Eq. (5) (Wang, Fu, Yao, & Li, 2018).

$$a_i = \sum_{j=1}^{n} (w_j \cdot x_j) + b_i \tag{4}$$

$$y_i = \frac{1}{1 + e^{-a_i}} \tag{5}$$

After the feed forward pass, the second phase of training utilizes the *Adam* activation function to minimize the loss function and update weight values (Kingma & Ba, 2015).



(a) Samples tested (b) Samples with Positive Results

Fig. 2. Venn diagram to demonstrate (a) Samples that received 1 or more of IgG, IgM, IgA or molecular testing; (b) samples that were tested positive for IgG, IgM, IgA or molecular testing. Venn diagrams are constructed using Venny (Oliveros, 2021).

3.4.4. Logistic regression

Logistic Regression is a classification algorithm that models the conditional probability of output y for sample x_i as defined in Eq. (6) (Yu et al., 2011). The output $y = 1$ indicates COVID-19+ and $y = -1$ indicates COVID-19- and w is the weight vector.

$$\mathcal{P}_{\mathbf{w}}(y = \pm 1|\mathbf{x}) = \frac{1}{1 + e^{-y\mathbf{w}^T\mathbf{x}_i}} \tag{6}$$

Weight values w are learned by utilizing a binary class l_2 penalization function to minimize the cost function defined in Eq. (7), where $C > 0$ is a penalty parameter (Yu et al., 2011). In our implementation, $C = 0.001$.

$$\arg\min_{\mathbf{w}} \mathcal{P}(\mathbf{w}) = C \sum_{i=1}^{n} \log\left(1 + e^{-y\mathbf{w}^T\mathbf{x}_i}\right) + \frac{1}{2}\mathbf{w}^T\mathbf{w} \tag{7}$$

3.4.5. Support vector machine

Support vector machine (SVM) classifies samples into two separate classes (COVID-19+ and COVID-19−) by creating a set of hyper-planes to create a decision boundary that separates the two classes by maximizing the margin. The margin is the smallest distance between the decision boundary and any of training samples (Bishop, 2009). The kernel function in a SVM projects the data from a low-dimensional space to a higher dimensional space. This allows the data to be separable in the higher dimensional space (Noble, 2006). In our study, we utilize the Radial Basis Function (RBF) kernel as defined in Eq. (8), where $\|x_i - x_j\|^2$ represents the squared Euclidean distance between two feature vectors of training sample x_i and x_j.

$$K(x_i, x_j) = e^{-\gamma\|x_i - x_j\|^2} \tag{8}$$

The parameter γ is defined in Eq. (9), where M indicates the number of features in the training set and σ is the variance of the training set.

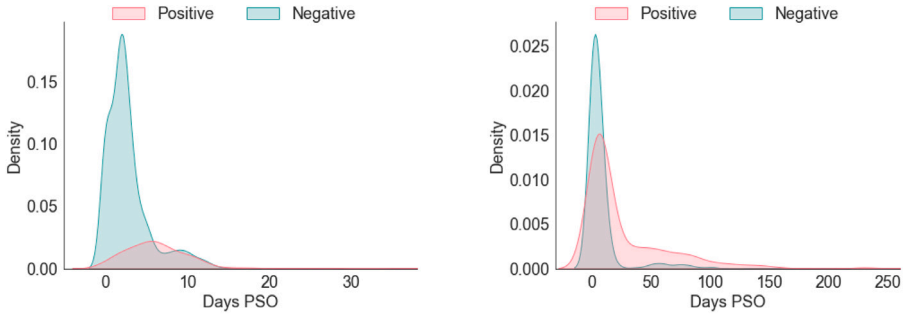
$$\gamma = \frac{1}{M \times \sigma^2} \tag{9}$$

4. Results

4.1. Serology vs. Molecular test relationships

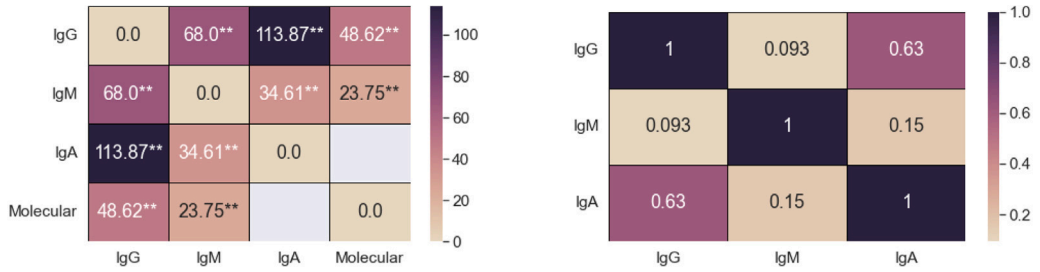
One unique feature of our testbed is that some donors may have multiple test results, as shown in Table 3. This allows us to analyze relationship between serology tests *vs.* molecular tests, and also understand consistency within each type of test.

To detail the overlapping between testing results, Fig. 2(a) reports the number of samples tested on serology and molecular tests, respectively. Fig. 2(b) reports the number of positive samples tested on serology and molecular tests (the Venn diagrams are constructed using Venny (Oliveros, 2021)). In each Venn diagram, the overlapped region report the number (and percentage) observed for the respective test combinations. For example, in Fig. 2(a), out of total 2467 donors, 677 donors have IgG tests only, 42 donors have both IgG and IgM tests, 192 donors have IgG, IgM, and IgA tests, and so on. Therefore, the total number of donors



(a) Samples with Molecular Testing (b) Samples with Serology Testing

Fig. 3. Kernel Density Estimation plot of days PSO with respect to (a) Samples with molecular testing; (b) Samples with serology testing.



(a) Chi-squared test of serology vs. molecular tests. (b) Pearson Correlation of serology tests.

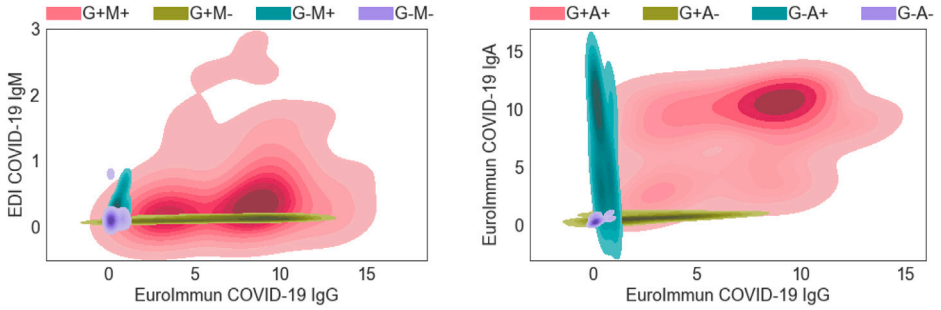
Fig. 4. Statistical tests of different test mechanisms. (a) Chi-Squared test to show statistical significance between pairs of COVID-19 diagnostic tests. The symbol ** indicates the result was statistically significant with $p < 0.001$. No sample in the dataset has both IgA and molecular test results, thus the corresponding cell is empty. (b) Pearson correlation matrix of optical density values from IgG, IgM and IgA tests.

with IgG tests is 1738 (677+42+192+55+146+626). In this way, we can clearly understand number of donors (and the positive percentage ratios) with respect to different test combinations. In Fig. 2(b), we further report the number (and percentage) of positive rate, with respect to each test combinations. For example, out of all 1738 donors receiving IgG tests (including combinations with all other tests), 689 donors shows COVID-19 positive only in IgG test, 46 donors show positive in both IgG and IgM, and 100 donors show COVID-19 positive in IgG, IgM, and IgA. Please note that 689 donors are positive in IgG test only in Fig. 2(b), but there are 677 donors with IgG test only in Fig. 2(b). This is not an error, because the Venn diagrams only show results for one specific condition. In this case, the interpretation is that 677 donors only receive IgG tests. Among all 1738 donors receiving IgG tests, including combinations of IgG with other tests, 689 donors only show IgG positive.

Overall, the results in Fig. 2 show that the positive ratios with respect to different tests vary significantly. For serology tests, the positive ratio for IgA is 187/247 = 75.7%, whereas the positive ratios for IgM and IgG are 44.8% (212/473), and 57.4% (997/1738), respectively. As shown in Fig. 2(a), 192 donors receive three serology tests (IgG, IgM, and IgA), and 100 of these donors are positive in all three tests, representing about 52.1% positive rate for Serology tests. For molecular test, the positive ratio 20.2% (294/1452) is far less than any of the serology tests.

COVID-19 tests are known to be dependent on specific factors such as viral load and seroconversion, therefore days PSO can be an important factor. To visualize the relationship between test results and days PSO, Fig. 3 shows a kernel density estimation plot of days PSO with respect to samples with molecular testing (a); and samples with serology testing (b). For samples with positive molecular testing, days PSO ranged from 0–37, with a median of 6 days for positive result and a maximum of 37 days for a positive result. The majority of the positive molecular samples fall between 3–8 days PSO. For serology tested samples, days PSO ranged from 0–260, with a median of 9 days for a positive result and a maximum of 260 days for a positive result. The majority of positive serology samples fall between 5–38 days PSO. The window for days PSO is shortened when using molecular testing, compared to serology testing. These differences in days PSO for detection is expected, as molecular testing indicates a current infection *vs.* an indirect indication of infection for serology tests. This illustrates the importance of days PSO when creating a model that uses a specific COVID-19 diagnostic test. In cases where molecular testing may not be available, donors may be forced to wait for tests results. The longer the wait, the higher the risk of obtaining a false negative. Similarly, tests done too early may risk a false negative due to low viral load.

Despite of significant variance between IgG, IgA, IgM, and molecular tests, Fig. 4(a) reports the Chi-squared tests (χ^2) between each pair of tests, using binary test outcomes (0/1 denotes negative/positive respectively). Because no donor receives both IgA and molecular tests, the corresponding Chi-squared value is empty. The results in Fig. 4(a) confirm that all tests are statistically significantly correlated with each other, with IgG and IgA tests showing highest confidence in correlation.



(a) IgG vs. IgM (b) IgG vs. IgA

Fig. 5. Pairwise Kernel Density Estimation of optical density values from (a) samples tested with EuroImmun COVID-19 IgG *vs* EDI COVID-19 IgM; and (b) samples tested with EuroImmun COVID-19 IgG *vs* EuroImmun COVID-19 IgA. The three tests are all ELISA tests. The samples are color coded to indicate their result interpretation. Darker colored densities indicate more samples in the area. G+ indicates IgG positive, G- indicates IgG negative; M+ indicates IgM positive, M- indicates IgM negative; A+ indicates IgA positive, A- indicates IgA negative.

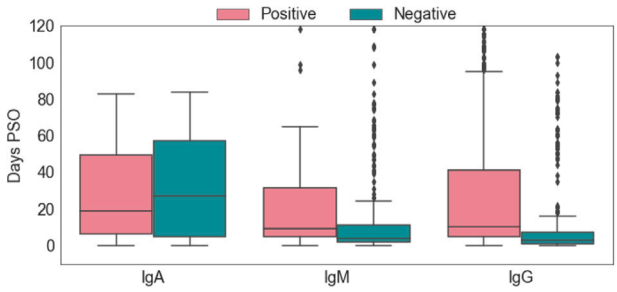


Fig. 6. Boxplot distribution of days PSO with respect to samples tested on IgA, IgM and IgG.

4.1.1. Serology test relationships

Immunoassay results can be expressed quantitatively in terms of optical density (OD) which correspond to antibody concentration. The interpretation of the assay is based on the measured OD value and the cut-off values as dictated by different test manufacturers. While OD values can vary based on different methods, manufacturers, and reference control values; in general, the higher the OD, the more antibody concentration. With this consideration, we can compare OD values from different serology tests to better understand their relationships, agreement, etc.

Fig. 4(b) demonstrates a pearson correlation matrix of OD values from IgG, IgM and IgA tests. For donors that received more than 1 test in a single group (IgG, IgM, IgA), the OD values were averaged. IgG and IgA have a strong positive relationship, $r(245) = 0.63$, $p < 0.001$. IgM and IgA are slightly positively correlated, $r(245) = 0.15$, $p = 0.03$. And IgM and IgG have no meaningful relationship, $r(351) = 0.093$, $p = 0.08$.

To further study the correlation between different types of serology tests, we visualize the distribution of samples immunoassay OD for three different ELISA tests; EuroImmun COVID-19 IgG, EDI COVID-19 IgM, and EuroImmun COVID-19 IgA in Fig. 5. Since the range of OD can vary between different immunoassay methods (ELISA *vs* CMIA *vs* CIA), thus we only consider comparable ELISA serology tests to visualize pairwise density distributions. Fig. 5(a) displays pairwise kernel density estimation of OD values from donors tested on EuroImmun COVID-19 IgG and EDI COVID-19 IgM ($n = 192$); Fig. 5(b) displays pairwise kernel density estimation of OD values from donors tested on EuroImmun COVID-19 IgG and EuroImmun COVID-19 IgA ($n = 192$). Fig. 5 displays a hue based on if the sample was positive or negative for the compared combination of IgG (G), IgM (M) and IgA (A).

Samples that have low optical density values for a given serology test will incur a negative interpretation. For example, samples that are negative for both serology tests, have OD values around 0 (purple areas marked G-M- in Fig. 5(a) and G-A- in Fig. 5(b)).

When looking at tests marked positive in each pair-wise tests (*i.e.* the pink areas marked G+M+ in Fig. 5(a) and G+A+ in Fig. 5(b)), we can find that there is a large variance between serology tests. For example, in Fig. 5(a) the long horizontal green bar (G+M-) indicates that for tests marked as IgM negative, there are significant spread out in IgG values making IgG tests being positive. IgG tests show the largest variance, followed by IgA tests, then IgM tests. Variance can be caused by different tests, different test manufacturer and even different runs.

To visualize the distribution of days PSO with respect to different serology tests, Fig. 6 displays a boxplot distribution of Days PSO for positive and negative samples tested on IgA, IgM, and IgG. Note that Fig. 6 range shows days PSO from 0–120; 56 samples had days PSO greater than 120, these are excluded to better visualize the main distribution.

With regards to IgM and IgG, positive IgM samples show an earlier window of detection with majority of samples between 4–31 days PSO, and median of 9 days. Positive IgG samples show a longer, and slightly later, window of detection; majority of samples

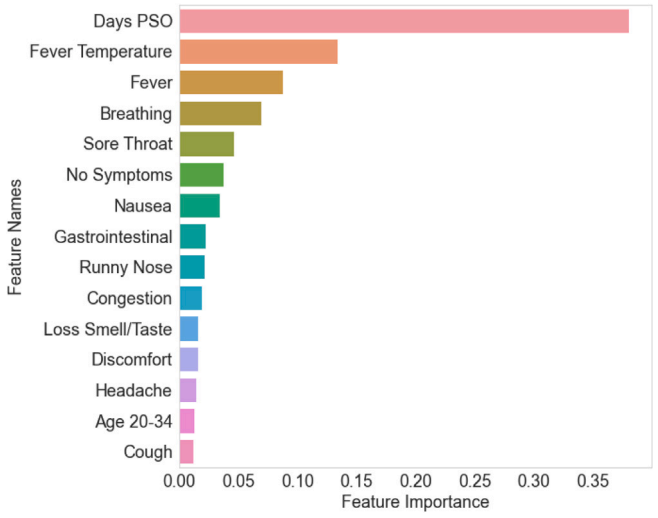


Fig. 7. Top 15 most informative features from Random Forest model.

fall between 4–41 days PSO with a median of 10 days. The upper extreme of positive IgG samples is greater than positive IgM, and IgA, demonstrating the ability of IgG tests to detect past infection even as much as 3 months post symptomatic. Negative IgM and IgG samples show an earlier window of days PSO. This suggests that obtaining an IgM or IgG test too early may increase the risk of a false negative.

Interestingly, the distribution of days PSO for positive and negative IgA samples are similar. Additionally, the majority of positive IgA samples are between 6–50 days PSO, with a median of 19 days. This suggests a later immune response than previously reported (Ma bibetal, 2020). However, note there is a smaller sample size ($n = 247$) of donors tested on IgA compared to donors tested on IgM and IgG, with large variability in the range of days PSO prior to testing.

Overall, these results show that IgM detection is closest to onset of symptoms; IgG detection can be done over a longer duration; days PSO shows a clearer separation for IgM and IgG compared to IgA. Days PSO is not as clearly determinant in positive IgA samples. The detection of IgA may be dependent on other factors previous reports suggest that IgA levels are significantly higher in severe COVID-19 cases *vs* mild and moderate cases (Ma bibetal, 2020). The severity of the infection may play a role in IgA detection for our dataset. However without additional information, such as subject outcome or symptom severity scales, its difficult to define mild, moderate or severe COVID-19 infection.

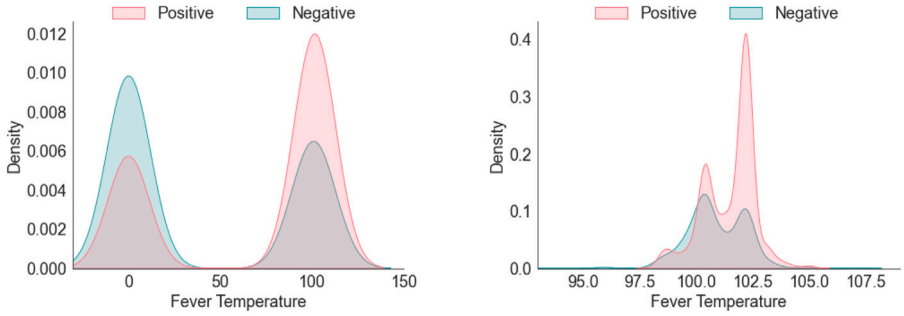
4.2. Symptom feature importance

To analyze which features have the highest importance for classification, the feature importance is averaged over the 5 Random Forest models trained on 5-fold cross validation. The top 15 important features are displayed in Fig. 7. Overall, Days PSO, Fever Temperature and Fever are ranked the highest for feature importance.

The high importance of Days PSO is likely due to the detection limits of the COVID-19 assays. Molecular tests are dependent on the viral load and serology tests are dependent on seroconversion, both of these require a certain range of days PSO. One of the major difficulties with COVID-19 is the number of asymptomatic cases, which is reported to be as many as one third of total COVID-19 cases (Oran & Topol, 2021). In our study, if a sample is asymptomatic, days PSO is set to 0. A total of 284 samples are asymptomatic. The asymptomatic samples introduce extra difficulties to predictive models that utilize symptom data for COVID-19 test predictions. Other than basic demographic features (age and gender), the asymptomatic samples would only have values for "No Symptoms". It is likely that a combination of "No Symptoms" and Days PSO is most informative for asymptomatic samples, which may be a reason for "No Symptoms" ranking at 6th place for feature importance.

Fever temperature and Fever (binary indicator) are ranked number 2 and 3 in feature importance, respectively. This is in concordance with past research showing that fever is an important feature for a COVID-19 discriminatory model (Ahamad bibetal, 2020; Tostmann bibetal, 2020; Zoabi et al., 2021). To view the density distribution of fever temperature between COVID-19 positive and COVID-19 negative samples, Fig. 8(a) shows a density plot of all samples; Fig. 8(b) shows density plot of samples with fever temperature information alone. In our study, if samples do not have a temperature listed, then "Fever Temperature" is recorded as 0. As shown in Fig. 8(a), presence of a fever clearly separates the COVID-19+ and COVID-19- samples. With respect to samples that had temperature data (Fig. 8(b)), higher fever temperatures are more likely to be COVID-19 positive.

Other symptoms ranked on the top 15 feature importance include Breathing, Sore Throat, Nausea, Gastrointestinal, Runny Nose, Congestion, Loss Smell/Taste, Discomfort, Headache and Cough.



(a) All Samples (b) Samples with Fever only.

Fig. 8. Kernel Density Estimation plot of fever temperature with respect to (a) all samples; (b) samples with fever only.

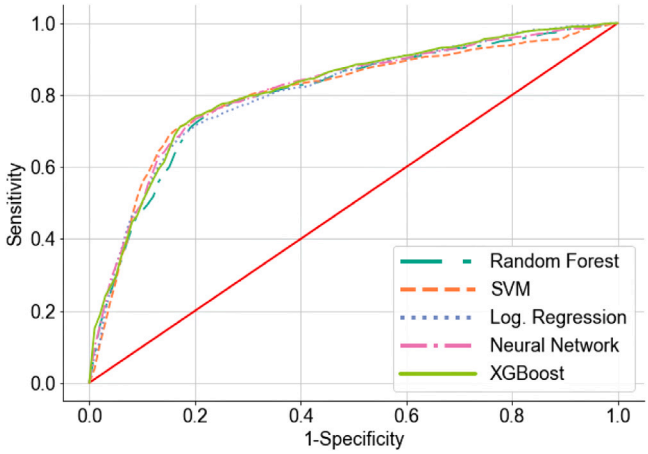


Fig. 9. Receiver Operating Characteristic (ROC) curves for the five classification models.

Table 6
Predictive model results.

Model	Accuracy	F1	AUC
Random Forest	76.37%	76.73%	80.07%
XGBoost	76.85%	77.04%	81.40%
Logistic Regression	70.57%	74.21%	80.67%
SVM	76.69%	76.61%	80.32%
Neural Network	75.88%	76.08%	81.02%

4.3. Predictive modeling results

After creating the feature set, samples are separated into 5 sets of training and test data for 5-fold cross validation. Five models are trained on each training set and tested on the test data set. Models used for classification are Random Forest, XGBoost, Logistic Regression, Support Vector Machine (SVM) and Neural Network. In our experiments, Random Forest, Logistic Regression and Support Vector Machine are implemented with Sci-Kit Learn (Pedregosa et al., 2011). XGBoost is implemented with the XGBoost Python Package. Neural Network is implemented using Keras and Tensorflow Python packages.

Three performance metrics are used to determine the classification performance, Accuracy, F1-score and AUC. The results after averaged 5-fold cross validation are reported in Table 6, and the ROC (receiver operating characteristic) curves for the five classification models are reported in Fig. 9.

Overall, the results indicate that XGBoost has the highest classification performance with 76.85% accuracy, 77.04% F1-score and 81.40% AUC scores.

4.3.1. Random forest case study

In order to validate the decision logic learned from Random Forest, Fig. 10 reports one decision tree. Because we limit the maximum depth of the tree to 4, the tree in Fig. 10 only has four layers. The tree is rooted from "Days PSO", because it is identified

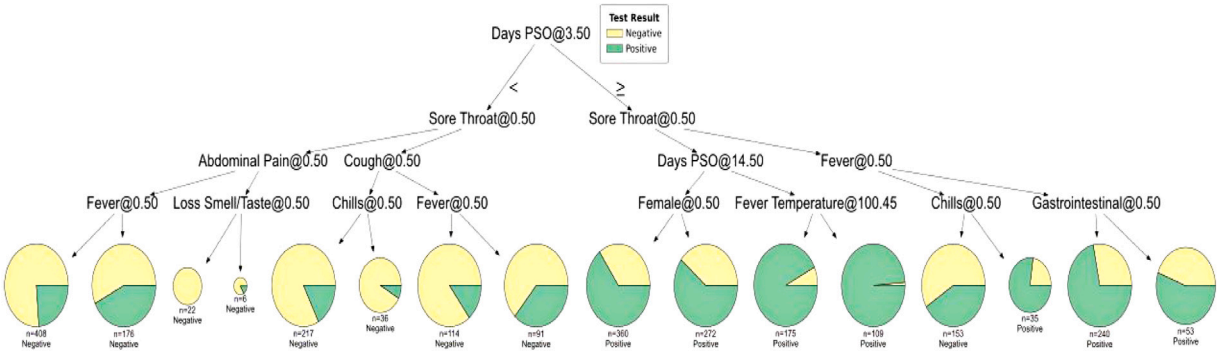


Fig. 10. A COVID-19 prediction decision tree learned from Random Forest.

as the most informative feature. The numerical value next to each feature shows the threshold for tree splitting. For example "Days PSO@3.50" means that the left branch contains the ones whose "Days PSO" values are less than 3.50, and the right branch has donors whose "Days PSO" are greater and equal to 3.50. For binary features, such as "Sore Throat", the threshold 0.5 is the middle point between two binary (0/1) values.

The tree in Fig. 10 shows clear decision logic possibly useful for clinical decision, as the decisions are transparent and the confidence of the decision can both be derived by looking at the path of the tree and the sample distributions at the leaf node of each path.

5. Discussion

In this study we aim to predict a positive COVID-19 test by using symptom and basic demographic features. In our study, symptom information is captured with a questionnaire with limited number of symptoms as well as a free text field for symptoms not contained on the defined list. Without a standardization of symptom reporting, the symptom feature space greatly increases. To combat this, we utilize a binning approach that groups together similar features into bins. This was able to decrease symptom feature space while keeping sample feature information. The limitation in this approach requires manually grouping symptoms, which can be subjective as well.

A limitation in using symptoms for predictive models is the subjective nature of reporting symptoms. Subjects may describe similar symptoms differently and may be biased in reporting symptoms. Additionally, there are limitations in recording symptoms. Questionnaires that limit symptom choices have a benefit of more standardized responses, however they have a disadvantage of potentially missing important or valid symptoms.

The recent advancement in deep neural language models (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) may provide alternatively solutions to learn features from noisy symptoms, without putting feature into bins. This, however, would result in difficulty in the interpretation of the model, and compromise the decision transparency.

Feature importance provides evidence in what features are most informative in distinguishing between COVID-19 positive (+) and COVID-19 negative (−). The feature importance ranking from Random Forest shows that "Days PSO" is the most informative for the classification model. This suggests that the number of days experiencing symptoms plays a large role in COVID-19 test results. This is to be expected, as molecular tests depend on viral load and serology tests depend on seroconversion. They both are time dependent. These results suggest that the number of days post symptomatic are highly important for a positive COVID-19 test and should be under careful consideration when screening patients.

Asymptomatic COVID-19 cases pose an additional difficult component for symptom predictive models and for estimating incubation period or immune response timeline. For asymptomatic subjects, other factors, such as exposure history or risk factors (healthcare workers, etc.) maybe very useful for predictive models.

6. Conclusion

In this paper, we carried out a machine learning study of COVID-19 serology and molecular tests, and proposed to use simple demographic and symptom features to train classification models for COVID-19 infection predict. By using test results from 2467 donors as our test bed, we analyzed correlation between serology tests and molecular tests. Our study shows that molecular tests have much narrower PSO days (between 3–8 days), comparing to PSO days of serology tests (between 5–38 days). As a result, molecular test has the lowest positive rate, because it measures current infection. On the other hand, COVID-19 tests vary significantly, partially because donors' immune response and viral load, the target of different test methods, continuously change. Even for the same donor, it might be possible to observe different positive/negative results from two types of tests. Our study finds a handful of informative symptom features, such as days PSO, fever temperature, fever, etc, strongly related to COVID-19. By using created bin features, combined with five machine learning algorithms, our predictive models achieve over 81% AUC scores, and over 76% classification accuracy. This study shows that machine learning models, trained using simple symptom and demographic features, can help predict COVID-19 infections.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

A processed version of the data may be available from the authors upon reasonable request and with permission from Boca Biolistics, LLC.

Acknowledgments

This work was supported by the U.S. National Science Foundation under Grants IIS-1763452 and IIS-2027339. Any opinions, findings, and conclusions or recommendations expressed in this research are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Institutional Review Board (IRB)

The project is limited to analysis of de-identified data points. The proposed study was reviewed and approved the Florida Atlantic University IRB under IRBNET ID #1950592-1.

References

Ahamad, M. M., Aktar, S., Rashed-Al-Mahfuz, M., Uddin, S., Liò, P., Xu, H., et al. (2020). A machine learning model to identify early stage symptoms of SARS-Cov-2 infected patients. *Expert Systems with Applications*, *160*, Article 113661. http://dx.doi.org/10.1016/j.eswa.2020.113661.

Alimohamadi, Y., Sepandi, M., Taghdir, M., & Hosamirudsari, H. (2020). Determine the most common clinical symptoms in COVID-19 patients: a systematic review and meta-analysis. *Journal of Preventive Medicine and Hygiene*, *61*(3), E304–E312. http://dx.doi.org/10.15167/2421-4248/jpmh2020.61.3.1530.

Bishop, C. M. (2009). *Pattern recognition and machine learning*. Springer: 781058134, OCLC.

Böger, B., et al. (2021). Systematic review with meta-analysis of the accuracy of diagnostic tests for COVID-19. *American Journal of Infection Control*, *49*(1), 21–29. http://dx.doi.org/10.1016/j.ajic.2020.07.011.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proc. of the 22nd ACM SIGKDD Conf.* (pp. 785–794). New York, NY, USA: http://dx.doi.org/10.1145/2939672.2939785.

Chen, X., Laurent, S., Onur, O. A., Kleineberg, N. N., Fink, G. R., Schweitzer, F., et al. (2021). A systematic review of neurological symptoms and complications of COVID-19. *Journal of Neurology*, *268*(2), 392–402. http://dx.doi.org/10.1007/s00415-020-10067-3.

Centers for Disease Control and Prevention *Symptoms of COVID-19*. From: https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html.

Elkin, M., & Zhu, X. (2021). Understanding and predicting COVID-19 clinical trial completion vs. cessation. *PLoS ONE*, *16*(7), Article e0253789. http://dx.doi.org/10.1371/journal.pone.0253789.

Han, H., Li, Y., & Zhu, X. (2019). Convolutional neural network learning for generic data classification. *Information Sciences*, *477*, 448–465. http://dx.doi.org/10.1016/j.ins.2018.10.053.

Iwendi, C., Bashir, A. K., Peshkar, A., Sujatha, R., Chatterjee, J. M., Pasupuleti, S., et al. (2020). COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm. *Frontiers Public Health*, *8*, 357. http://dx.doi.org/10.3389/fpubh.2020.00357.

Kingma, D. P., & Ba, J. L. (2015). Adam: A method for stochastic optimization. (pp. 1–15). CoRR, arXiv:1412.6980.

Kumleben, N., Bhopal, R., Czypionka, T., Gruer, L., Kock, R., Stebbing, J., et al. (2020). Test, test, test for COVID-19 antibodies: The importance of sensitivity, specificity and predictive powers. *Public Health*, *185*, 88–90. http://dx.doi.org/10.1016/j.puhe.2020.06.006.

Larsen, J. R., Martin, M. R., Martin, J. D., Kuhn, P., & Hicks, J. B. (2020). Modeling the onset of symptoms of COVID-19. *Frontiers Public Health*, *8*.

Ma, H., Zeng, W., He, H., Zhao, D., Jiang, D., Zhou, P., et al. (2020). Serum IgA, IgM, and IgG responses in COVID-19. *Cellular & Molecular Immunology*, *17*(7), 773–775. http://dx.doi.org/10.1038/s41423-020-0474-z.

Mallett, S., et al. (2020). At what times during infection is SARS-CoV-2 detectable and no longer detectable using RT-PCR-based tests? A systematic review of individual participant data. *BMC Medicine*, *18*(1), 346. http://dx.doi.org/10.1186/s12916-020-01810-8.

Mei, X., et al. (2020). Artificial intelligence–enabled rapid diagnosis of patients with COVID-19. *Nature Medicine*, *26*(8), 1224–1228. http://dx.doi.org/10.1038/s41591-020-0931-3.

Menni, C., et al. (2020). Real-time tracking of self-reported symptoms to predict potential COVID-19. *Nature Medicine*, *26*(7), 1037–1040. http://dx.doi.org/10.1038/s41591-020-0916-2.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). *Distributed representations of words and phrases and their compositionality* (pp. 3111–3119).

Noble, W. S. (2006). What is a support vector machine? *Nature biotechnology*, *24*(12), 1565–1567. http://dx.doi.org/10.1038/nbt1206-1565.

Oliveros, J. C. (2021). Venny. An interative tool for comparing lists with Venn's diagrams. Available from: https://bioinfogp.cnb.csic.es/tools/venny/index.html.

Oran, D. P., & Topol, E. J. (2021). The proportion of SARS-CoV-2 infections that are asymptomatic. *Annals of Internal Medicine*, M20–6976. http://dx.doi.org/10.7326/M20-6976, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7839426/.

Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Quer, G., et al. (2021). Wearable sensor data and self-reported symptoms for COVID-19 detection. *Nature Medicine*, *27*(1), 73–77. http://dx.doi.org/10.1038/s41591-020-1123-x.

Singhal, T. (2020). A review of Coronavirus Disease-2019 (COVID-19). *The Indian Journal of Pediatrics*, *87*(4), 281–286. http://dx.doi.org/10.1007/s12098-020-03263-6.

Sudre, C. H., et al. (2021). Symptom clusters in COVID-19: A potential clinical prediction tool from the COVID symptom study app. *Science Advances*, *7*(12).

Theel, E. S., Slev, P., Wheeler, S., Couturier, M. R., Wong, S. J., & Kadkhoda, K. (2020). The role of antibody testing for SARS-CoV-2: Is there one? In A. J. McAdam (Ed.), *Journal of Clinical Microbiology*, *58*(8), http://dx.doi.org/10.1128/JCM.00797-20, arXiv:https://jcm.asm.org/content/58/8/e00797-20.full.pdf.

Tostmann, A., Bradley, J., Bousema, T., Yiek, W.-K., Holwerda, M., Bleeker-Rovers, C., et al. (2020). Strong associations and moderate predictive value of early symptoms for SARS-CoV-2 test positivity among healthcare workers, the Netherlands, March 2020. *Eurosurveillance*, *25*(16).

Wang, S., Fu, L., Yao, J., & Li, Y. (2018). The application of deep learning in biomedical informatics. In *2018 international conference on robots intelligent system* (pp. 391–394).

Weissleder, R., Lee, H., Ko, J., & Pittet, M. J. (2020). COVID-19 diagnostics in context. *Science Translational Medicine*, *12*(546), eabc1931. http://dx.doi.org/10.1126/scitranslmed.abc1931.

Worldometer (2022). Coronavirus outbreak. From: https://www.worldometers.info/coronavirus/ Accessed: May 5, 2022.

Yu, H.-F., et al. (2011). Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning*, *85*(1–2), 41–75. http://dx.doi.org/10.1007/s10994-010-5221-8.

Yuki, K., Fujiogi, M., & Koutsogiannaki, S. (2020). COVID-19 pathophysiology: A review. *Clinical Immunology*, *215*, Article 108427. http://dx.doi.org/10.1016/j.clim.2020.108427.

Zoabi, Y., et al. (2021). Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *Npj Digital Medicine*, *4*(1), 3. http://dx.doi.org/10.1038/s41746-020-00372-6.