Journal of Infectious Diseases & Therapy
ISSN: 2332-0877

Research Article Open Access

Point-of-Care No-Specimen Diagnostic Platform using Machine Learning and Raman Spectroscopy: Proof-of-Concept Studies for Both COVID-19 and Blood Glucose

Allen Chefitz[1*], Rohit Singh[2], Thomas Birch[3], Yongwu Yang[1], Arib Hussain[3] and Gabriella Chefitz[4]

[1]123IV Inc., New Rochelle, United States of America
[2]Department of Computational Biology, Massachusetts Institute of Technology, Massachusetts, United States of America
[3]Holy Name Medical Center, Teaneck, United States of America
[4]Department of Medicine, Icahn School of Medicine at Mount Sinai, New York City, United States of America

Abstract

Significance: We present a innovative, specimen-free diagnostic platform that can immediately detect both a metabolite (glucose) or an infection (COVID-19), by non-invasively using Raman spectroscopy and machine learning.

Aim: Current diagnostic testing for infections and glucose monitoring requires specimens, disease specific reagents, processing and increases environmental waste. We propose a new hardware-software paradigm by designing and constructing a finger-scanning, hardware device to acquire Raman spectroscopy readouts and, by varying a machine learning algorithm to interpret the data, allows for diverse diagnoses.

Approach: 455 patients were enrolled prospectively in the COVID-19 study. 148 tested positive and 307 tested negative on nasal PCR testing done concurrently with testing using our viral detector. The tests were performed on both outpatients (N=382) and inpatients (N=73) at Holy Name Medical Center in Teaneck, NJ between June, 2021 and August, 2022. Patients' fingers were scanned using an 830 nm Raman System and then, using machine learning, processed to provide an immediate result.

In a separate study between April, 2023 and August, 2023 measurements using the same device and scanning a finger were used to detect blood glucose levels. Using a Dexcom sensor and an Accu-Chek device as references, a cross-validation based regression of 205 observations of blood glucose was performed with a machine learning algorithm.

Results: In a five-fold cross-validation analysis (including asymptomatic patients), a machine learning classifier using the Raman spectra as input achieved a specificity for COVID-19 of 0.837 at a sensitivity of 0.80 and an Area Under Receiver Operating Curve (AUROC) of 0.896. However, when the data were split by time, with training data consisting of observations before 1st July, 2022 and test data consisting of observations after it, the model achieved an AUROC of 0.67, with 0.863 sensitivity at a specificity of 0.517. This decrease in AUROC may be due to substantial domain shift as the virus evolves. A similar five-fold cross validation analysis of Raman glucose detection produces an Area Under Precision-Recall Curve (AUPR) of 0.58.

Conclusion: The combination of Raman spectroscopy, Artificial Intelligence/Machine Learning (AI/ML) and our patient-interface admitting only a patient's finger and using no specimen, offers unprecedented flexibility in introducing new diagnostic tests or adapting existing ones. As the ML algorithm can be iteratively retrained with new data and the software deployed to field devices remotely, it promises to be a valuable tool for detecting rapidly emerging infectious outbreaks, as well as disease specific biomarkers, such as glucose.

Keywords: Raman spectroscopy; Artificial intelligence/Machine learning; Medical diagnostics; Point-of-Care

Introduction

The prevailing approach to diagnostic testing requires taking specimens and relies on assays customized to specific biomolecules. For instance, rapid antigen tests and PCR tests rely on specimens from nasal swabs and hemoglobin or glucose measurements require biochemical tests on blood or urine samples. Each new diagnostic test thus requires a custom-designed assay, with its attendant challenges (e.g., ramping up availability during pandemics, or democratizing access to a diverse population). This is also a problem with infectious diseases where rapid evolution of the pathogen might make existing assays obsolete, wasting limited production resources and slowing down public health response due to delays in designing, developing, manufacturing and distributing new test kits.

The last decade of methodological advances in molecular and cellular biology suggest an alternative paradigm: A hardware platform capable of taking multiple, general-purpose measurements of the system under study, followed by computational analysis to extract the biological covariates of interest. For instance, this is broadly the approach followed in single-cell genomic assays. The power of this approach lies in a careful combination of broad, yet sensitive assays paired with custom-designed and calibrated computational analysis. We hypothesize that a similar approach could also be transformative in clinical diagnostics, with a no specimen platform that could provide accurate readouts for a variety of diagnostic tasks while being accessible in a point-of-care setting. Crucially, adapting to a pathogenic variant or even an entirely different disease would simply be a matter of loading new software on the device (or on the cloud), making it possible to introduce new tests much more efficaciously.

We propose the use of Raman spectroscopy, introducing a novel pulse oximeter-like device to obtain readouts non-invasively (US patent # 11452454; 11304605) and analyze the resulting Raman spectra with machine learning techniques trained to detect an infectious state or a biochemical metabolite. Raman spectroscopy detects precise

*Corresponding author: Allen Chefitz, 123IV Inc., New Rochelle, United States of America, E-mail: ohmori@edu.k.u-tokyo.ac.jp

Received: 13-Dec-2024, Manuscript No. JIDT-24-155179; **Editor assigned:** 16-Dec-2024, PreQC No. JIDT-24-155179 (PQ); **Reviewed:** 30-Dec-2024, QC No. JIDT-24-155179; **Revised:** 06-Jan-2025, Manuscript No. JIDT-24-155179 (R); **Published:** 13-Jan-2025, DOI: 10.4173/2332-0877.24.S4.004.

Citation: Chefitz H, Singh R, Birch T, Yang Y, Hussain A, et al. (2025) Point-of-Care No-Specimen Diagnostic Platform using Machine Learning and Raman Spectroscopy: Proof-of-Concept Studies for Both COVID-19 and Blood Glucose. J Infect Dis Ther S4:004.

vibrational modes of molecules in organic liquids and, as such, provides a physicochemical fingerprint of molecules. The instrumental setup typically includes a light source, such as a laser, a spectrograph to collect and disperse the scattered light by Fourier Transform methods, a detector, such as a Charge-Coupled Device (CCD) and a filter to separate the Raman scattering from other light signals. We chose Raman spectroscopy over other spectroscopic methods, as the Raman technique offers several advantages. Compared with infrared spectroscopy, near-infrared, mid-infrared and far-infrared, Raman has the advantage of being nondestructive, fast to acquire and capable of providing information at the molecular level. Raman also analyzes samples in aqueous solutions, since water produces a weak Raman scattering, especially important in the biochemical field studying the ionization behavior, pH change, or amino acid configuration [1].

We hypothesized that Raman spectroscopy readouts of an individual's blood chemistry, performed non-invasively by scanning one's finger in a form similar to a pulse oximeter, could provide the raw information needed to diagnose a variety of conditions. For instance, it could be directly used to assess the blood metabolic profile of the individual. Moreover, we reasoned that even infections like COVID-19 could be indirectly measured, as the combination of infection-specific antigen and immune response would leave molecular traces captured in the spectrum.

We pair such a hardware platform with a powerful yet flexible ML framework for interpreting the raw data. Previously, the use of support vector machines with Raman spectroscopy has been demonstrated for the classification of breast cancer tissue [2]. In the COVID-19 setting, detection can be posed as a classification problem given the Raman spectrum, while blood glucose measurement can be a regression problem. Essentially, any new assay requires only the training of a new machine learning module, trained by paired measurements of the Raman spectrum and ground-truth data, illustrating the flexibility of the device for various normal and disease states. Here, we demonstrate the feasibility of such an approach by designing, constructing and operating such a device and applying it to predict COVID-19 infection status and blood glucose levels in cross-validation settings. In this feasibility study, we sought to show that a machine learning algorithm can indeed recover the underlying ground-truth from the Raman spectra. Our algorithms were trained on a limited amount of ground-truth training data for both COVID-19 or glucose tests and demonstrate substantial accuracy in predicting COVID-19 infection status or high glucose levels. For instance, on COVID-19, our model achieves a cross-validation sensitivity of 80% with a specificity of 83.7%. In comparison, sensitivities for current COVID-19 antigen tests are reported to range from 63.9%-70.2% [3,4]. With more extensive data collection for training the machine learning algorithms, we expect the accuracy of the platform to further improve.

Our work addresses a pressing gap in the availability of diagnostic frameworks that can serve as convenient, non-invasive point-of-care screening choices and do not require expertise in sample handling. In comparison, testing for the presence of COVID-19 infection and blood glucose levels with currently available tests requires a nasal swab, sputum or blood specimen. While some previous work has demonstrated the feasibility of combining ML with Raman spectroscopy, it has been used with patient blood, sputum, nasal or tissue samples [2,5-7]. In contrast, our platform's innovation includes not requiring any bodily fluid sample, features the novel patient interface form-factor for Raman spectroscopy and AI/ML.

We formulated a hypothesis that COVID-19 could be detected transcutaneously in a finger in a manner similar to pulse oximetry. The feasibility of using Raman spectroscopy to detect the unique spectrum of SARS-CoV-2, as well as other viruses and molecules, using a specimen has previously been demonstrated [8-10]. For instance, a Raman-based classification model to discriminate the signal in saliva of COVID-19 patients has been described, with the study noting peaks that are attributable to the tryptophan and phenylalanine signal and to the C-N and C-C stretching [11]. Aromatic amino acids, including tryptophan, play an important role in the spike protein structure.

One of our key conceptual advances is to use Raman spectroscopy to detect COVID-19 non-invasively and transcutaneously from the blood. This would side-step the need for collecting specimens, unlocking unprecedented ease-of-use. We set out to target the entire mixture of blood components that is typically seen in COVID-19. Our effort is motivated by several reports of the systemic effects and sequelae of COVID-19, involving the heart, liver, clotting system, inflammatory system, intestines, kidney and brain, suggesting that a combination of COVID-19 molecular markers, direct or indirect, must be prevalent in the blood of COVID-19 patients [5,12]. Relatedly, over 900 metabolites in the blood have been inputted into a predictive model for COVID-19, adding to previous studies reporting metabolomics in the blood [13]. Given all the evidence for biomarkers in the blood in patients with COVID-19, we hypothesized that we could target the unique combinations and concentrations of multiple blood-borne molecules *in vivo*, in addition to SARS-CoV-2 RNA, using Raman transcutaneously. Our approach would thus combine the optical engineering of Raman (detecting the signals of multiple targets) with machine learning (classifying and predicting) to arrive at a diagnosis of COVID-19.

Glucose

We sought to extend our diagnostic platform, using the same hardware but with the software re-trained on new data from a different target-glucose. Kang et al., observed glucose with a Raman spectroscope [14]. They predicted glucose in a pig model from the Raman spectra of 911 cm^{-1}, 1060 cm^{-1} and 1125 cm^{-1} using Partial Least Square (PLS) regression analysis, as opposed to a more powerful machine learning algorithm. Enejder et al., describe the non-invasive measurement of glucose in 17 humans following a glucose tolerance test using Raman spectroscopy and PLS analysis [15]. We sought to improve on the predictive power of Raman and address some of the challenges described in those studies, such as background noise and limits of detection, by integrating our machine learning algorithms. The advantages of our glucose study protocol include being outpatient, with the usual variations in glucose levels due to exercise and diet and random. We did not intentionally raise the glucose levels to increase the Raman signal. Our subject is diabetic on pump therapy, thereby more closely representing the intended patient population. We also note that sensors for glucose monitoring can deviate from Fingerstick (FS) tests by 20 points, as we show. Davis et al., compared, in the hospital setting, sensor and FS readings and found a mean absolute relative difference of 12.8% [16]. With more variation in exercise and diet in the outpatient setting, one would expect a larger difference than when compared in the inpatient setting, as we find.

Materials and Methods

Hardware prototype (same device was used for both COVID-19 and glucose testing): The assembly includes a 830 Raman system with a high Signal-to-Noise Ratio (SNR) of (6400:1) with a fiber-optic probe housed in a patient-interface into which an individual inserts a finger to receive the laser excitation (Figures 1-4). The 830 nm laser is selected as an excitation source to reduce fluorescence interference. The system also employs a back illuminated CCD to increase system sensitivity.

The 830 nm laser is transmitted through a 105 micro core fiber to the probe which is collimated and shined on the fiber. The Raman signal is captured and guided through a signal fiber to the Raman spectrograph. The laser powered at 250 mW was used to capture Raman spectra for, sequential 20 second intervals [3]. The components of the device include an 830 nm laser (Innovative Photonics Solutions, Inc.,), BI-CCD camera (Andor), Imaging Spectrograph SRaman 830 W2 Innovations, Inc., and Raman mprobe 830 (W2 Innovations, Inc.,). Both the raw data (including fluorescence) and processed data (pure Raman) were stored.

The device was inspected and the design analyzed for electrical, mechanical and laser safety. All electrical components are shrouded within a metal box and connections to the patient-interface are fiber-optic. For use on patients, a medical grade isolation transformer was used to provide two means of patient protection per the requirements of 60601-1. The device uses a class 3B laser that is housed, including a permanent light blocking shield that conforms around the finger. The viral detector device blocks a direct or reflected beam. Measured laser power was below the Maximum Permissible Exposure (MPE) for Skin Exposure per ANSIz136.1 of 364 mW/cm^2.

The captured data was inputted into a machine learning classifier that had been previously trained to distinguish individuals who have been found to test positive for disease, such as COVID-19, from those who have been found to test negative for that disease. The classifier outputted an appropriate indicator, providing the user with an immediate indication of whether the patient was COVID-19 positive or negative.

Figure 1: Viral Detector-Human interface with attached laser and signal fiberoptic cables. Inserted finger with minimization of ambient light, as employed in 476 patient study. Interface houses the Raman probe U.S. Patent No. 11,452,454, U.S. Patent No. 11,304,605.

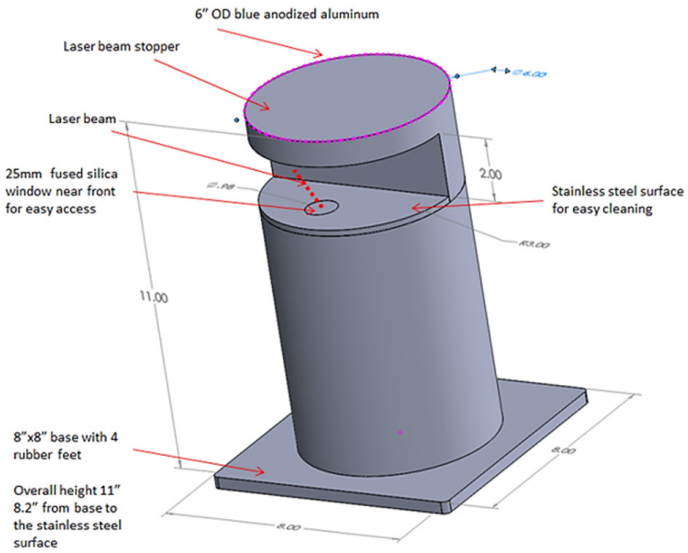

Figure 2: Dimensions of human interface of viral detector used in 476 patient study.

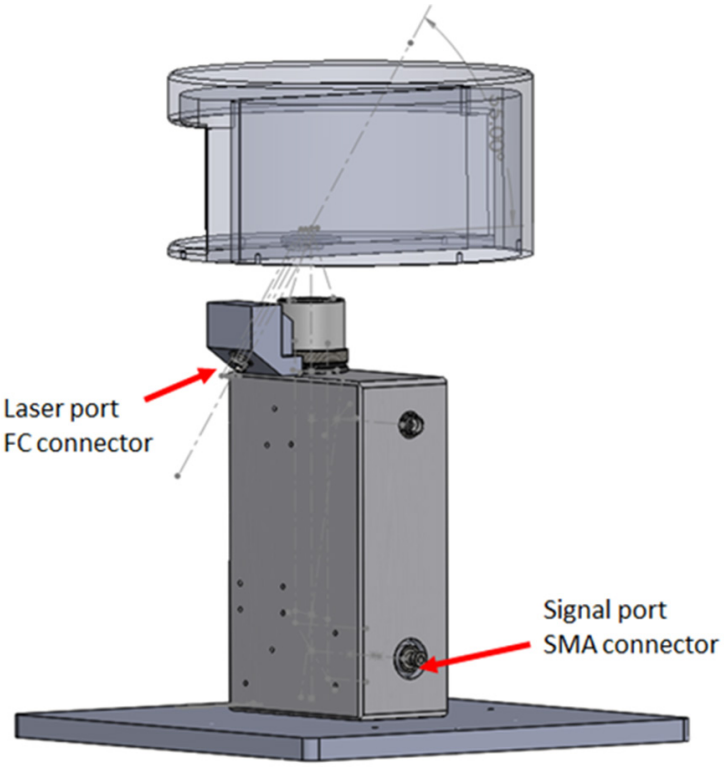

Figure 3: Components of human interface for finger insertion with additional laser safety protections. Raman Probe connections for laser and signal fiberoptic cables. Cables connect to laser source and CCD.

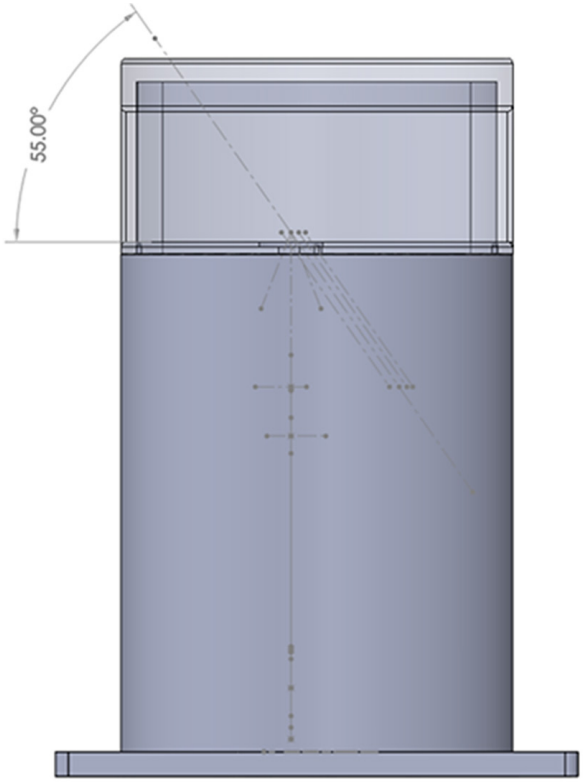

Figure 4: Frontal view depiction of direction of laser and signal with focus on laser safety.

COVID-19 study

Prior to obtaining IRB approval at Holy Name Medical Center in Teaneck, NJ, the device underwent extensive safety testing at Sunrise Labs, Inc., in Bedford, NH. Informed consent was obtained for all patients. We excluded from the study anyone younger than 18 years of age. Patient data were collected, including age, symptoms, vaccination data, gender and viral load (cycle numbers). Since transcutaneous devices have previously been documented to report incorrect results for non-caucasian skin colors, we also collected skin color data [17].

Expected outcome of the research

We set out to demonstrate a correlation between positive and negative COVID-19 tests by PCR with the test results from the experimental transcutaneous viral detector.

Exclusion criteria

Pregnant women and other special populations, such as minors and legally incompetent patients.

Recruitment

While undergoing PCR testing, either at the outpatient lab of Holy Name Medical Center or on the hospital ward, the finger test was done concurrently, as outlined in the protocol. No patient identifiers were entered. The PCR test was then processed at the Holy Name Medical Center lab and results entered into the patient's medical record. PCR test identifying numbers were recorded and stored.

Study duration

The pilot study was conducted June, 2021 to August, 2022.

Sample size

In designing this feasibility study and its statistical power, we anticipated the need for 100 patients comprising a sample of 50 individuals (both men and women) with SARS-CoV-2 by PCR and 50 with negative PCR tests, according to the McNemar χ^2 test. We studied 476 (316- and 160+) patients.

Variables

Available for all patients: The following details available for all the patients.

- Raman spectra
- Inpatient/Outpatient
- PCR results

Available for some patients: The following details available for some patients.

- Symptoms
- Fever
- Cycle threshold numbers
- Skin color (motivated by studies of skin color-effect on pulse oximeter accuracy)
- Flu vaccination status

Secondary outcomes

In addition, there are secondary outcomes that we hope to explore in future work. These include:

- Effect of skin color on Raman measurements using our device
- Ability of our device to detect immune status
- Ability of our device to calculate viral load

Ethics

This study protocol complied with the Institutional Review Board (IRB) and Investigative Committee on Clinical Research (ICCR) at Holy Name medical Center and was conducted according to rules and guidelines of Good Clinical Practice (GCP). Data was handled confidentially and all data was stored for the length of the study and for 15 years afterwards at the site, for further publication. Informed consent, approved by the Center's Ethical Committee, was obtained from the subject by an authorized research team member only.

Glucose study

Between April, 2023 and July, 2023 one patient (signed consent received) with Type 1 IDDM on a Tandem pump with a Dexcom G6 sensor underwent 205 Accu-Chek glucometer finger stick blood glucose testing at home. The tests were performed randomly throughout the day and night. Each time a test was performed, 3 sequential tests-FS and Dexcom readings-were taken separated by 1 minute. Concurrently, 3 finger insertions into the Raman/AI device were done over 3 minutes (Table 1).

Data analysis

The details of specific machine learning are described in Results. Here we describe the preliminary processing and normalization of the Raman spectra. For each observation, the set of Raman scores were treated as a 2000-dim feature vector. Since limits to the device's spectral resolution may cause "bleed-through" in adjacent points of the spectrum, we explored a Gaussian kernel centered at each point to smoothen the observations. However, initial cross-validation studies suggested this was unnecessary. Next, in the training set, we fitted a StandardScaler to standardize each feature to have mean=0, standard deviation=1. To avoid risks of data snooping, the same scaling was applied to the test set rather than fitting a separate scaler on the test.

Results

COVID-19 testing

Sample Raman spectra are shown in Appendix A1. We also show the Receiver Operating Characteristic (ROC) curve, the average spectrum of a patient and t-tests indicating that multiple points of the Raman spectra contain predictive power (Figures 5-7). The sensitivity and specificity results are shown in Table 2.

AUROC 0.896

We employed a Gradient Boosted Tree (GBT) classifier to learn SARS-CoV-2 status from a patient's finger spectroscopy (hyperparameters described later in this paper). We chose a GBT model (500 trees) since such tree-based models have performed well with tabular data analyses. In a preliminary cross-validation analysis, we observed that GBTs performed slightly better than random forests and other approaches. As its prediction, the GBT model reports a numeric score between 0 and 1 (inclusive), with 1 indicating highest confidence of a positive (i.e., infected) case. To binarize these predictions, an appropriate threshold needs to be selected (the AUROC metric aggregates over all such thresholds). In the results we report below, we focused on thresholds corresponding to sensitivity ≥ 0.8 and assessed the corresponding specificity at that level. We performed two analyses:

Data split randomly: Over all 455 observations, we performed a 5-fold cross-validation analysis (i.e., 80% training, 20% test data), with observations shuffled across time. At a sensitivity of 0.80, the GBT model achieved an average specificity of 0.837 (standard error: 0.046). Over all thresholds, it achieved an average AUROC of 0.896 (standard error: .025)

Data split by time: We wondered if the mutation of the virus and the appearance of different strains over time would have an impact on the model's predictive ability. We therefore split the data by time, training on observations before 01st July, 2022 (N=404 with 278 negatives) and testing on later observations (N=51 with 29 negatives). With the same hyperparameters as before, we trained a GBT model. It achieved a sensitivity of 0.818 with a specificity of 0.517, with AUROC=0.684.

The reduction in performance from a random-split to a temporal-split evaluation suggests that the changes in the patterns of Raman spectra over time, possibly induced by mutations in the virus, can have an impact on model performance. We believe that a larger, more extensive study could further resolve these issues. Additionally, we believe that a practical approach would be to update the model by re-training on updated data. Fully anonymized patient observations could be for iterative refinement of the model. The advantage of a machine learning based solution is that such refinement can be made in software and then transmitted back to field devices with minimal delay.

We believe it is unlikely that our model is confounded by some non-COVID condition among the patient population. With 382 outpatients in our study that spanned all 4 seasons, the possibility that an unknown medical condition common to all outpatient COVID-19 positive patients would account for the unique spectrum assigned to COVID-19 and thus confound our results is extremely unlikely.

Another advantage of the GBT is that we could explore feature importances, in order to interpret the model and identify the most important spectra. In particular, we observed that spectra corresponding to the range of 1700-2200 cm^{-1} were important to distinguishing between positive and negative examples.

Thus, our study strongly suggests that with increasing data, the prototype-a-non-invasive, immediate, trans-cutaneous diagnostic device- becomes more powerful.

Glucose testing

Sample Raman spectra are shown in Appendix A2. We also show the precision-recall curve (Figure 8).

The glucose data analysis used the same overall machine learning framework as the COVID-19 study, but was trained afresh, in five-fold cross-validation, on Glucose-specific data. As shown by the no-skill line in Figure 8, the AUPR of 0.58 indicates that the model is indeed able to learn aspects of glucose concentration, though we expect that more training data, beyond the scope of this proof-of-concept study, is needed for stronger predictive performance.

Attribute	Details
Age	63
IDDM history	53 years
Gender	Male
Dexcom	G6
Accu-Chek	Guide
Pump	Tandem Control-IQ v7.6.0.1 SN1065558
Skin color	White

Table 1: Patient characteristics.

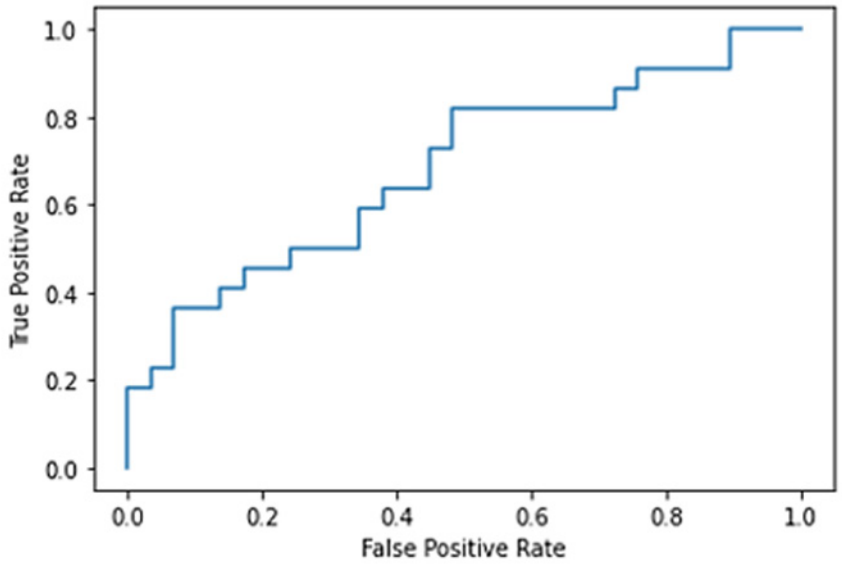

Figure 5: Receiver operating curve-From the machine learning data analysis.

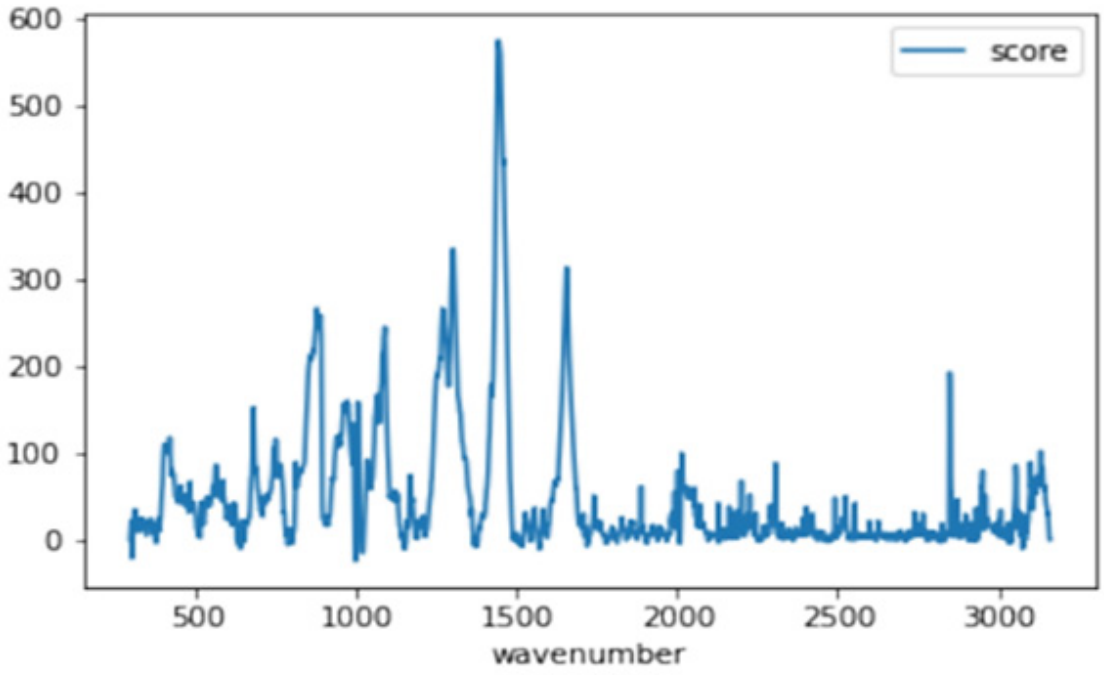

Figure 6: Average Raman spectrum output from the device (average intensity shown across cases and controls in the training dataset).

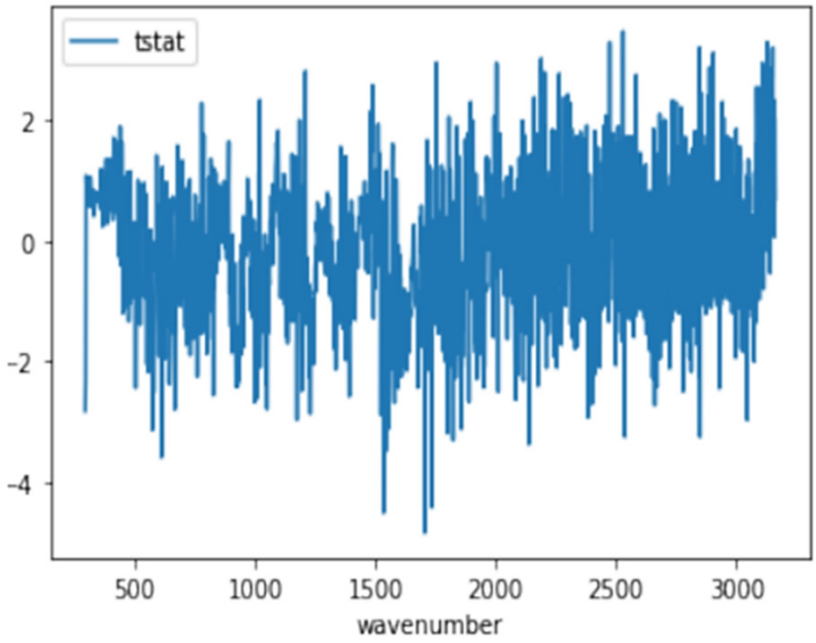

Figure 7: Individual Raman peaks contain information that can distinguish between cases and controls. T-tests on score Z: For each individual, wavenumber values (across the 2000 points) were converted to z-scores to control for patient-specific biases.

	PCR	123CY (Finger test)
COVID+	148	118/148
COVID-	307	257/307
Sensitivity	-	0.80
Specificity	-	0.837

Table 2: Number of patients.

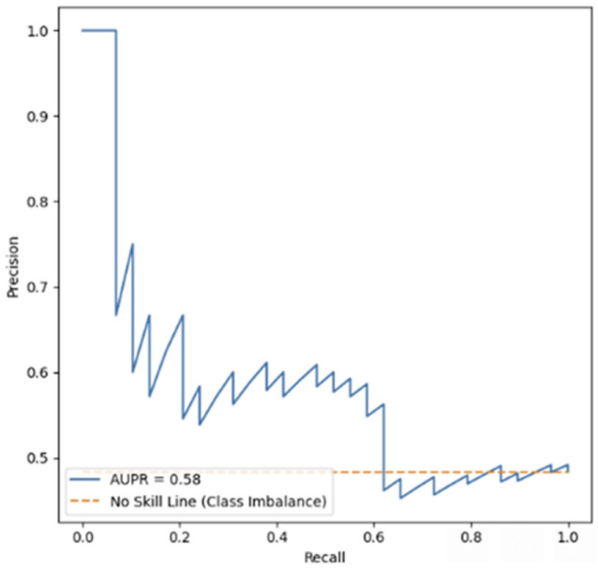

Figure 8: Glucose testing precision-recall curve.

Discussion

We have described a diagnostic platform that employs Raman spectroscopy with a unique patient-interface for finger insertion and, using specific hardware parameters together with machine learning, can detect COVID-19 transcutaneously and non-invasively, as well as glucose. When the device targets the blood transcutaneously, the derived Raman spectra together with machine learning produces a significantly different result between COVID-19$^{(+)}$ and COVID-19$^{(-)}$ patients. We included in our study both inpatients and outpatients, symptomatic and asymptomatic. To demonstrate the robustness of our platform, we extended its application to glucose detection. It is precisely the integration of our AI/ML model with Raman spectroscopy, using the finger insertion patient-interface, which allows the specific target's signal to be detected with as few as 205 glucose readings.

SARS-CoV-2

In conceiving our idea and design for the diagnostic device, we anticipated various challenges that were sufficiently met. Although COVID-19 is caused by a "respiratory" virus, we believed at the outset that COVID-19 could be detected by investigating the blood. The rich vasculature of the aero-digestive tract, upper and lower airways and oropharynx would absorb almost anything inhaled, as occurs with many inhaled substances. Even if the virus itself were not in the blood, or at a concentration too low to reliably detect with our device, the increasing reports of systemic effects of COVID-19 were strong evidence that there were hematological abnormalities to target [5,12,13].

Our study enrolled 476 patients (after quality control, we kept N=455 samples), none of whom experienced any adverse effects from the device. We anticipate that we can improve our results with further modifications of the device's parameters, as well as more data input.

With the advent of Spatially Offset Raman Spectroscopy (SORS) and other deep tissue techniques, more attention has been aimed at biomedical applications [8,9,18]. However, those applications of Raman include a specimen on a slide, other platform, or container, or a direct application of a probe to a specimen [19,20]. Our objective was to safely combine the best Raman parameters of laser power, wavelength, distance, detector and spectrograph processing with the most powerful machine learning algorithm in order to non-invasively "see" beneath the skin into the human vasculature and target the many molecules that represent the intravascular unique signature combinations, both quantitatively and qualitatively, in an infected COVID-19 patient.

Others have demonstrated the feasibility of combining Raman spectroscopy with machine learning. A Multiple Instance Learning (MIL) approach in machine learning was employed for COVID-19 infection in saliva, in which an Area Under the Curve (AUC) of 0.8 and a sensitivity of 79% (males) and 84% (females) and a specificity of 75% (males) and 64% (females) were achieved [7]. A feasibility study that analyzed actual blood tests, focusing on variables "age", "wbc", "crp", "ast" and "lymphocytes" from 102 COVID-19 negative and 177 COVID-19 positive patients, applied different machine learning algorithms to achieve an accuracy of 82%-86%. Using a Random forest classifier, their best results were accuracy 82%, sensitivity 92%, PPV 83%, AUC 84% and specificity 65% [5].

The underlying biochemistry of the expected acute phase reactants, as well as viral composition, has been studied using Raman spectroscopy on a specimen and has thus served as the scientific basis of our study. Please refer to Appendix B for further details.

In our study, we examined non-invasively with a unique patient-interface the entire mixture of blood components, which when viewed as a whole indicate the presence or absence of COVID-19. The text files from the Raman spectra, when processed with machine-learning, detected differences between COVID-19 positives and negatives. The feature importance analysis of the gradient boosted tree classifier reported 1800, 2200 and 2400 cm-1 as important peaks for distinguishing COVID-19. Glycerol, a metabolic product of glycerides and lipids, is known to produce a peak in this area (1000-1500) when examined on a slide. It is possible that a change in lipid metabolism occurs as a result of COVID-19, similar to the alteration in the clotting cascade. Our differences in Raman spectra between positives and negatives would be consistent with the findings of Roberts et al., in their study of the metabolomics of COVID-19 [13]. Similarly, activity at 1700 Raman shift is strongly associated with molecules containing a carbonyl group C=O, such as proteins. Carbonyl groups are found in HIV as well as COVID-19 [21]. Higher shifts in the 2000 range could be associated with cyano groups C=N, also in molecules such as RNA. While more data is needed to definitively ascribe these important peaks

to causal COVID-19 factors, the observations above suggest that ML approach may be detecting viral breakdown products (spike protein and RNA), as well as some other acute phase proteins [21]. It is precisely the combinations of molecules we sought to detect and analyze.

In designing the prospective, observational study, we sought to anticipate and mitigate any potential bias. We compared our diagnostic device with the gold standard RT-PCR, performed in a hospital approved lab. As we classified in our data 73 inpatients as well as 382 outpatients and classified them further into symptomatic and asymptomatic, we believe that the potential for bias is minimized. Further, we have included both inpatients who are negative as well as asymptomatic positive patients in our analysis. Most importantly, because the underlying basic science of Raman spectroscopy has been proven to generate a unique spectral signature for each different organism, such as influenza, Ebola, or SARS-CoV-2, in a specific specimen, we believe the same likely holds true for different organisms examined transcutaneously [6,10].

Glucose and beyond

We also demonstrated that, with our model and platform, blood glucose can be detected in this non-invasive manner. Although we plan to extend our device's applications in larger studies, the benefits of using one patient's data for a feasibility study include improved calibration, less distortion and cleaner data. With more data, the power of the AI/ML will increase further, not only because of the quantitative increase, but also due to the increasing richness of the data. We know that with hypoglycemia, other blood-borne molecules are present, such as glucagon and epinephrine, so that the analysis of the entire mixture of blood analytes will yield valuable input data.

A limitation of our glucose study concerns the detection of hypoglycemia, likely attributed to including only 31 values for a blood glucose under 70 mg/dl, as well as an absolute smaller amount of glucose target. However, because we do detect a distinct Raman shift representing glucose (800 cm^{-1} and 1100 cm^{-1}) with a greater quantity of glucose in the blood, we expect that with more hypoglycemic data, the AI/ML will improve its detection of hypoglycemia.

We are currently working on other biomedical applications for our technology, such as infections, pathology and cancer diagnosis. We aim to solve existing dilemmas involving intraoperative margin status in cancer surgery, infections (bacterial, viral, or fungal) and pathology with devices, such as ours, with further research. The secondary outcomes analysis from our data is ongoing and will be included as part of future research. As we measured viral load in accordance with the PCR cycle number, we anticipate correlating our device's test results with the cycle number and other parameters. Interestingly, among the false negatives for our COVID-19 Raman tests were positive PCR tests whose cycle number was 40. The device can detect COVID-19 among symptomatic patients even more powerfully than the statistics indicate because the study group included both symptomatic as well as asymptomatic individuals in both the PCR negative and positive groups. Quantifying someone's immunity, by looking not only at antibody level but also at other immune markers, is another goal. Determining the presence of viral variants and predicting the nature of the variants will require further analysis [22].

Conclusion

Although a library of Raman spectra already exists for laboratory-examined infectious agents and other disease-associated molecules, and is likely to closely resemble our transcutaneously-derived library, we aim to expand our own collection of Raman spectra for infectious agents and molecules. A similar approach can be applied to other pathogens or variants using our device. By employing probabilistic models and computational biology, we hope to map the physical structure of future infectious agents. Using machine learning, we aim to determine the Raman signature of new pathogens, enabling us to identify a novel infection before it is sequenced in the lab. Building on our glucose study, we plan to extend the application of our diagnostic platform to detect additional molecular markers and chemistries in blood.

Acknowledgements

AC and RS are co-inventors on a patent.

Code and Data Availability

The code and datasets generated and/or analyzed during the current studies are available on GitHub, accessible at https://github.com/123IVrobot/raman_ai., upon acceptance for publication.

Conflict of Interest

AC and RS are co-inventors on a patent.

Data Analysis

The details of specific machine learning are described in Results. Here we describe the preliminary processing and normalization of the Raman spectra. For each observation, the set of Raman scores were treated as a 2000-dim feature vector. Since limits to the device's spectral resolution may cause "bleed-through" in adjacent points of the spectrum, we explored a Gaussian kernel centered at each point to smoothen the observations. However, initial cross-validation studies suggested this was unnecessary. Next, in the training set, we fitted a StandardScaler to standardize each feature to have mean=0, standard deviation=1. To avoid risks of data snooping, the same scaling was applied to the test set rather than fitting a separate scaler on the test.

Author Contribution

A.C. conceived and designed the protocol and wrote the introduction and discussion. R.S. created the artificial intelligence/machine learning model and analyzed and interpreted the study data. T.B. helped with the design of the study protocol and is the principal investigator for the COVID-19 study. Y.Y. optical engineer who built the index diagnostic device and wrote the methods section pertaining to the spectroscopy. A.H. interviewed the study patients, collected their data and tabulated it. G.C. researched the biochemistry and basic science supporting our findings, wrote the abstract and edited the discussion.All authors reviewed the manuscript.

References

1. Balan V, Mihai CT, Cojocaru FD, Uritu CM, Dodi G, et al. (2019) Vibrational spectroscopy fingerprinting in medicine: From molecular to clinical practice. Materials 12:2884.
2. Zhang L, Li C, Peng D, Yi X, He S, et al. (2022) Raman spectroscopy and machine learning for the classification of breast cancers. Spectrochim Acta A Mol Biomol Spectrosc 264:120300.
3. Jegerlehner S, Suter-Riniker F, Jent P, Bittel P, Nagler M, et al. (2021) Diagnostic accuracy of a SARS-CoV-2 rapid antigen test in real-life clinical settings. Int J Infect Dis 109:118-122.
4. Schuit E, Venekamp R, Hooft L, Veldhuijzen IK, Bijllaardt W, et al. (2022) Diagnostic accuracy of COVID-19 rapid antigen tests with unsupervised self-sampling in people with symptoms in the omicron period: Cross sectional study. BMJ 378:e071215.
5. Brinati D, Campagner A, Ferrari D, Locatelli M, Banfi G, et al. (2020) Detection

of COVID-19 infection from routine blood exams with machine learning: A feasibility study. J Med Syst 44:135.

6. Vallejo-Perez MR, Sosa-Herrera JA, Navarro-Contreras HR, Álvarez-Preciado LG, Rodríguez-Vázquez AG, et al. (2021) Raman spectroscopy and machine-learning for early detection of bacterial canker of tomato: The asymptomatic disease condition. Plants (Basel) 10:1542.

7. Ember K, Daoust F, Mahfoud M, Dallaire F, Ahmad EZ, et al. (2022) Saliva-based detection of COVID-19 infection in a real-world setting using reagent-free Raman spectroscopy and machine learning. J Biomed Opt 27:025002.

8. Nicolson F, Kircher MF, Stone N, Matousek P (2021) Spatially offset Raman spectroscopy for biomedical applications. Chem Soc Rev 50:556-568.

9. Andreou C, Kishore SA, Kircher MF (2015) Surface-enhance Raman spectrosopy: A new modality for cancer imaging. J Nucl Med 56:1925-1929.

10. Huang J, Wen J, Zhou M, Ni S, Le W, et al. (2021) On site detection of SARS-CoV-2 antigen by deep learning-based surface-enhanced Raman spectroscopy and its biochemical foundations. Anal Chem 93:9174-9182.

11. Carlomagno C, Bertaziolo D, Gualerzi A, Picciolini S, Banfi PI, et al. (2021) COVID-19 salivary Raman fingerprint: Innovative approach for the detection of current and past SARS-CoV-2 infections. Scientific Reports 11:4943.

12. Aluko OM, Lawal SA, Reuben CS, Jeje SO, Ijomone OM, et al. (2022) Understanding the systemic effects of COVID-19: Possible clues to potential therapeutic approaches. Int J Trop Dis 5:057.

13. Roberts I, Muela MW, Taylor JM, Davison AS, Xu Y, et al. (2021) Untargeted metabolomics of COVID-19 patient serum reveals potential prognostic markers of both severity and outcome. Metabolomics 18:6

14. Kang JW, Park YS, Chang H, Lee W, Singh SP, et al. (2020) Direct observation of glucose fingerprint using *in vivo* Raman spectroscopy. Sci Adv 6:eaay5206.

15. Enejder AMK, Scecina TG, Oh J, Hunter M, Shih W, et al. (2005) Raman spectroscopy for noninvasive glucose measurements. J Biomed Opt 10:031114.

16. Davis GM, Spanakis EK, Migdal AL, Singh LG, Albury B, et al. (2021) Accuracy of Dexcom G6 continuous glucose monitoring in non-critically ill hospitalized patients with diabetes. Diabetes Care 44:1641-1646.

17. Sjoding MW, Dickson RP, Iwashyna TJ, Gay SE, Valley TS, et al. (2020) Racial bias in pulse oximetry measurement. N Engl J Med 383:2477-2478.

18. Mosca S, Conti C, Stone N, Matousek P (2019) Spatially offset Raman spectroscopy. Nat Rev Methods Primers 1:21.

19. Yeh YT, Gulino K, Zhang Y, Sabestien A, Chou TW, et al. (2019) A Rapid and label-free platform for virus capture and identification from clinical samples. Proc Natl Acad Sci U S A 117:895-901.

20. Shipp DW, Rakha EA, Koloydenko AA, Macmillan RD, Ellis IO, et al. (2018) Intra-operative spectroscopic assessment of surgical margins during breast conserving surgery. Breast Cancer Res 20:69.

21. Kolgiri V, Patil VW (2017) Protein carbonyl content: A novel biomarker for aging in HIV/AIDS patients. Braz J Infect Dis 21:35-41.

22. Hie B, Zhong ED, Berger B (2021) Learning the language of viral evolution and escape. Science 371:284-288.