Classification of Patient Recovery from COVID-19 Symptoms using Consumer Wearables and Machine Learning

Jared Leitner, Alexander Behnke, Po-Han Chiang, Student Member, IEEE, Michele Ritter, MD, Marlene Millen, MD, Sujit Dey, Fellow, IEEE

***Abstract* – Current remote monitoring of COVID-19 patients relies on manual symptom reporting, which is highly dependent on patient compliance. In this research, we present a machine learning (ML)-based remote monitoring method to estimate patient recovery from COVID-19 symptoms using automatically collected wearable device data, instead of relying on manually collected symptom data. We deploy our remote monitoring system, namely eCOVID, in two COVID-19 telemedicine clinics. Our system utilizes a Garmin wearable and symptom tracker mobile app for data collection. The data consists of vitals, lifestyle, and symptom information which is fused into an online report for clinicians to review. Symptom data collected via our mobile app is used to label the recovery status of each patient on a daily basis. We propose a ML-based binary patient recovery classifier which uses wearable data to estimate whether or not a patient has recovered from COVID-19 symptoms. We evaluate our method using leave-one-subject-out (LOSO) cross-validation, and find that Random Forest (RF) is the top performing model. Our method achieves an F1-score of 0.88 when applying our RF-based model personalization technique using weighted bootstrap aggregation. Our results demonstrate that ML-assisted remote monitoring using automatically collected wearable data can supplement or be used in place of manual daily symptom tracking which relies on patient compliance.**

Index Terms – Machine Learning, Wearables, Remote Patient Monitoring, COVID-19

I. Introduction

Around the world, healthcare systems have been overwhelmed by the high numbers of COVID-19 cases, which has surpassed 437 million as of March 2, 2022 according to the World Health Organization (WHO) [1]. In the US, there were approximately 4.5 million COVID-19 hospitalizations between August 1, 2020 and February 28, 2022, according to the Center for Disease Control and Prevention (CDC) [2]. While this is a daunting number of hospitalizations, there have been approximately 80 million cases in the US [3], meaning the large majority of cases involve ambulatory patients being treated from home. This is an unprecedented number of patients needing care in their home and many are not being monitored in any way by medical personnel.

In order to combat this pandemic and provide more optimal care at scale, hospitals are changing the way in which healthcare is delivered. At the center of this changing landscape is a shift towards remote, continuous, and automated delivery of healthcare. This shift can lead to significant improvement in and scalability of at-home patient care for COVID-19, while at the same time enabling significant savings in human and equipment resources. Current remote monitoring for COVID-19 patients relies on manual symptom reporting, which is highly dependent on patient compliance. In this study, we demonstrate that data automatically collected from wearable devices together with machine learning (ML)-assisted diagnosis can enhance the efficiency and increase the scalability of remote monitoring for COVID-19 patients.

Wearable devices are one of the enabling technologies making this shift in healthcare delivery possible [4-7]. Consumer wearables, such as Apple Watch, Fitbit, and Samsung Galaxy Watch, remotely collect a great amount of lifestyle and vitals data in high granularity and continuity. There is great opportunity for ML to assist in remote monitoring due to the large amount of data that is collected. Since it is not possible for doctors to manually review all remotely collected data [8], ML has the potential to provide automated insights into the health status of patients and significantly increase the scalability of remote patient care. This is especially helpful during a pandemic, where in-person interaction and monitoring may pose risks to healthcare workers and other patients. In addition, ML-assisted monitoring can provide patients with insights regarding their own progression, helping to keep them engaged and informed about their health.

Current research on using wearables and machine learning to combat COVID-19 is primarily focused on early detection of infection. The authors in [12-16] have demonstrated that it is possible to detect deviations in health data before significant symptoms arise. Using Fitbit devices, the researchers in [12] found that 26 out of 32 (81%) infected patients in their cohort had alterations in their heart rate, number of daily steps, or time asleep before becoming symptomatic. The authors in [15] used respiration rate, heart rate, and heart rate variability data collected from their wearable devices and proposed a deep learning method to estimate infection before the onset of symptoms. Early detection will enable individuals to quarantine earlier, helping reduce the spread of the virus. These studies demonstrate that

Jared Leitner, Alexander Behnke, and Po-Han Chiang are graduate students with the Electrical and Computer Engineering Department within the Jacobs School of Engineering at the University of California, San Diego. Dr. Sujit Dey is a professor in the Electrical and Computer Engineering Department and head of the Mobile Systems Design Laboratory at UCSD. Michele Ritter, MD is an infectious disease specialist and director of the COVID-19 telemedicine clinic at UCSD Health. Marlene Millen, MD is a primary care physician and chief medical information officer of ambulatory care at UCSD Health.
(Corresponding author: Jared Leitner, e-mail: jjleitne@eng.ucsd.edu)

wearable device data can provide actionable insights into the conditions of patients.

In this research, we propose a novel approach to estimate patient recovery from COVID-19 symptoms using automatically collected device data and machine learning. We partnered with the UCSD Health and Neighborhood Healthcare COVID-19 telemedicine clinics in order to carry out this research. Our remote monitoring system utilizes a Garmin wearable and symptom tracker mobile app for data collection and fuses this data into an online report for clinicians to review. We propose a novel labelling logic for patient recovery from COVID-19 symptoms using the symptom tracker data. The labelling logic was developed in collaboration with UCSD Health doctors and the details are defined in Sec. III (B). Using this data, we train a patient recovery classifier which uses wearable data to estimate whether or not a patient has recovered from COVID-19 symptoms. We evaluate our method according to leave-one-subject-out (LOSO) CV to replicate the clinically relevant use case scenario in which a newly infected patient will not have data for model training. We compare the performance of different ML models and find that Random Forest (RF) is the top performing model. We propose a RF-based personalization technique in order to improve model performance. This technique utilizes the RF's weighted bootstrap aggregation algorithm in order to tune the model to each patient. The details are presented in Sec. III (D). Finally, we conduct Shapley Value analysis to inspect which device features have the greatest impact on classification. This analysis provides an interpretation of what the model has learned, which is especially important for medical applications. Our contributions are as follows:

- We deploy a remote patient monitoring system in two COVID-19 telemedicine clinics. The system consists of a wearable device, symptom tracker mobile app, and online dashboard which collects and analyzes vitals, lifestyle, and symptoms data. The estimated recovery status of each patient using our ML approach is displayed on the dashboard for clinicians to review.
- We propose a patient recovery classifier which uses wearable data to estimate whether or not a patient has recovered from COVID-19 symptoms. This ML tool can provide doctors with automated insights into the recovery status of their infected patients and bypass the need for manual daily symptom tracking.
- We carry out LOSO CV to mirror the clinically relevant use-case scenario and propose a RF-based personalization technique that improves model performance by tuning the model to each patient via weighted bootstrap aggregation.

The rest of the paper is organized as follows. In Section II, we investigate related works that utilize machine learning for COVID-19 diagnosis. In Section III, our remote monitoring system and data acquisition are presented. We then detail the proposed labelling logic and RF-based personalization technique for patient recovery classification. In Section IV, the performance of our proposed ML method is evaluated. In addition, we carry out top feature analysis based on Shapley Values and provide a discussion on research challenges. Finally, we conclude the paper in Section V.

II. Related Work

In this section, we present related research which is grouped into the follow categories: COVID-19 symptom tracking, early diagnosis of COVID-19, and recovery detection from COVID-19. Table I summarizes the comparison of related works.

A. COVID-19 Symptom Tracking

The researchers in [9] utilize a smartphone-based app to collect symptom data from patients. In the app, patients also recorded when they had tested either negative or positive for COVID-19 infection. They propose a logistic regression model that combines the reported symptoms in order to predict COVID-19 infection. A combination of loss of smell and taste, fatigue, persistent cough and loss of appetite resulted in the best model, which achieved a sensitivity and specificity of 0.65 and 0.78, respectively. The authors in [10] also used a mobile app for

Table I. Comparison of related works.

Reference	Objective	Device Features (# features)	Method	ML Model Interpretation
[9]	COVID-19 symptom tracking	No	Logistic Regression	No
[10]	COVID-19 symptom tracking	No	Logistic Regression	Yes
[11]	COVID-19 symptom tracking	No	Gradient-Boosting Machine	Yes
[12]	Early diagnosis of COVID-19	Yes (2)	Gaussian Anomaly Detection	No
[13]	Early diagnosis of COVID-19	Yes (1)	Deterministic State Machine	No
[14]	Early diagnosis of COVID-19	Yes (3)	Logistic Regression	No
[15]	Early diagnosis of COVID-19	Yes (4)	Convolutional Neural Network	No
[16]	Early diagnosis of COVID-19	Yes (16)	Gradient-Boosting Machine	Yes
[17]	Recovery detection from COVID-19	No	Support Vector Machine	No
[18]	Recovery detection from COVID-19	No	Decision Tree	No
Ours	Recovery detection from COVID-19	Yes (28)	Random Forest	Yes

collecting symptoms data and COVID-19 test results. They trained a logistic regression model to predict COVID-19 infection based on self-reported symptoms, and calculated the odds ratio for each symptom in order to understand which symptoms were the strongest predictors. Chills, fever, loss of smell, nausea, and shortness of breath were the top five strongest predictors of COVID-19 infection. Participants in their cohort with a positive test result experienced 5.6 symptoms on average. In [11], the researchers trained a gradient-boosting machine to predict COVID-19 infection based on 8 features: cough, fever, sore throat, shortness of breath, headache, age, sex, and known contact with an individual confirmed to have COVID-19. Their approach achieved a sensitivity and specificity of 0.86 and 0.79, respectively. Fever and cough were the top 2 features with the greatest impact on the model's prediction. These past works demonstrate that self-reported symptoms can be effectively used to predict COVID-19 infection. However, these approaches rely on patient compliance with manual symptom tracking. In contrast, wearable devices can passively collect data that is relevant to COVID-19 infection. In addition, wearable device data can be predictive of COVID-19 infection prior to symptom onset.

B. Early Diagnosis of COVID-19

The authors in [12] use data collected from wearable devices for the early detection of COVID-19 infection. They propose an anomaly detection technique based on two parameters: 1. Resting heart rate (RHR), 2. Heart rate over steps (HROS). HROS was calculated by dividing heart rate by steps data at each hourly interval. They report that significant deviations in these parameters relative to the individual baseline can indicate COVID-19 infection. They utilize Gaussian density estimation to classify anomalies in the dataset. Their results show that 63% of COVID-19 cases in their cohort could have been detected before symptom onset. The researchers in [13] also utilize deviations from RHR to classify a patient as infected. They propose a deterministic finite state machine which triggers an alert when a patient's overnight RHR increases above the median of previous overnight RHRs by an empirically determined threshold. Their system generated alerts for 80% of the infected individuals prior to symptoms, however, many of the alert-generating events were not associated with COVID-19 and instead attributed to other events, such as poor sleep, stress, alcohol consumption, intense exercise or travel. While these studies demonstrate that deviations in physiological and activity data measured by wearable devices can be used for early detection of COVID-19, they only utilize a subset of possible device features (RHR and steps) and do not investigate ML-based approaches which are well suited to handle larger feature sets. Furthermore, they do not investigate whether wearable device data can be used to monitor patient recovery from COVID-19.

The researchers in [14] trained a logistic regression model to differentiate COVID-19 positive vs. negative cases in symptomatic individuals based on symptoms and wearable device data. Baseline device data was calculated as the median of the data from 21 to 7 days before the onset of symptoms. They show an increase in model performance when including device data (RHR, sleep duration and step count) in addition to symptoms data as part of the feature set. The authors in [15] trained a convolutional neural network to predict illness given health metrics for that day and the preceding 4 days. These metrics included the mean respiration rate (RR) during sleep, mean heart rate during sleep, the root mean square of successive differences (RMSSD) of the nocturnal RR series and the Shannon entropy of the nocturnal RR series. They organize each data sample into 5x4 matrix and resize each matrix into a 28x28 image as the input to the network. Their method achieved a sensitivity and specificity of 51% and 90%, respectively. In [16], the researchers presented a gradient-boosting model based on decision trees to detect COVID-19 infection. Their approach achieved a sensitivity and specificity of 71% and 67%, respectively, when only using device features as input to the model. They grouped the device features into activity, sleep and heart rate categories, and found that activity related features had the greatest impact on the model's prediction, followed by sleep and heart rate-related features. These works demonstrate the ability of ML models to learn meaningful relationships between wearable device features and the onset of COVID-19 infection.

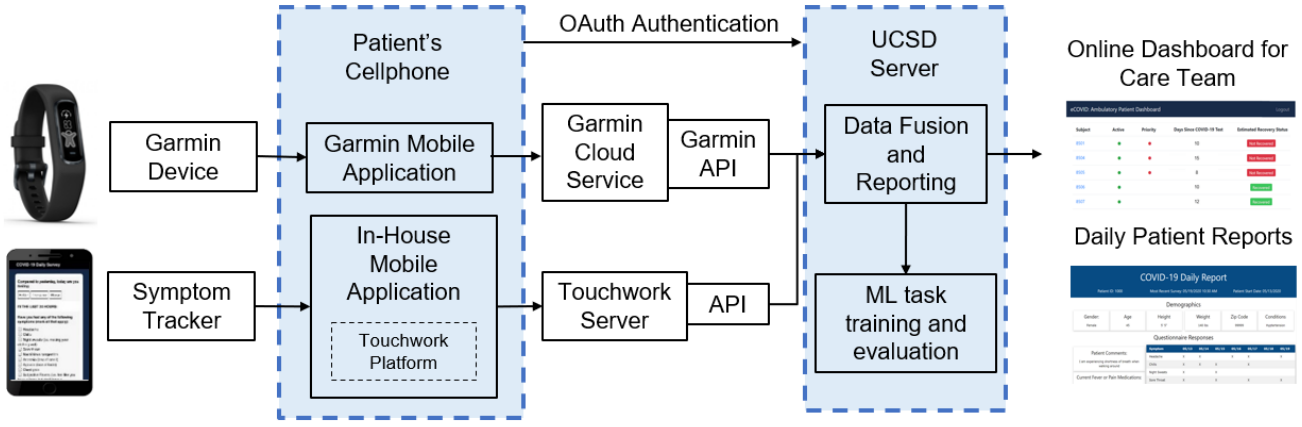

Figure 1. eCOVID remote monitoring and reporting system architecture.

Table III. Daily Questions in Symptom Tracker App.

Question	Answer Options
1. Compared to yesterday, today are you feeling:	a. *Better* b. *Worse* c. *Same*
2. Have you had any of the following symptoms (mark all that apply):	*Headache, Chills, Night sweats, Sore throat, Nasal/sinus congestion, Anosmia (loss of smell), Ageusia (loss of taste), Chest pain, Subjective fevers*
3. How would you rate your fatigue?	a. *0 (no fatigue)* b. *1 (mild fatigue – able to do normal activities)* c. *2 (prefer to lay and sit around, but still doing some activities)* d. *3 (mostly laying/sitting around, but still independently able to prepare meals and take medications)* e. *4 (mostly laying around - need help with preparing meals, but able to dress yourself, take medications and use bathroom independently)* f. *5 (staying in bed or chair all day, need assistance to make meals, ambulate, take medications, get dressed, and/or use bathroom)*
4. How would you rate your cough?	a. *0 (no cough)* b. *1 (minimal - clearing throat, less than 10 times a day)* c. *2 (coughing intermittently throughout the day - over 10 times/day)* d. *3 (Coughing frequently throughout the day, but not preventing sleep or interfering with activities)* e. *4 (Coughing so frequently that it makes it difficult to sleep, and/or is affecting usual activities)* f. *5 (Coughing so severe that it is causing shortness of breath, vomiting, inability to sleep)*
5. How would you rate any shortness of breath?	a. *0 (no shortness of breath)* b. *1 (minimal shortness of breath - only during coughing episodes)* c. *2 (shortness of breath with significant exertion - after climbing flight of stairs, walking long distances)* d. *3 (shortness of breath with usual daily activities - getting dressed, showering, preparing meals)* e. *4 (shortness of breath with minimal activity - moving from bed to chair, going to bathroom)* f. *5 (shortness of breath at rest - just while sitting or lying)*
6. Are you able to drink and eat?	a. *Yes - normal appetite* b. *Somewhat - decreased appetite (50-75% of what I normally eat/drink)* c. *Little (less than 25-50% of what I normally eat/drink)* d. *Minimal (<25% of what I normally eat/drink)*
7. What fever/pain medications have you taken?	a. *Acetaminophen (Tylenol)* b. *NSAIDS (Ibuprofen, Motrin, Advil, Naprosyn)* c. *Not Applicable (N/A)*
8. What cough/breathing medications have you taken?	a. *Steroid inhaler (Advair, Pulmicort, Flovent, Budesonide, Qvar, Symbicort, Beclomethasone)* b. *Pill steroid (Prednisone, Methylprednisolone, Dexamethasone)* c. *Rescue inhaler (Albuterol, ProAir, Ventolin, Proventil)* d. *Not Applicable (N/A)*

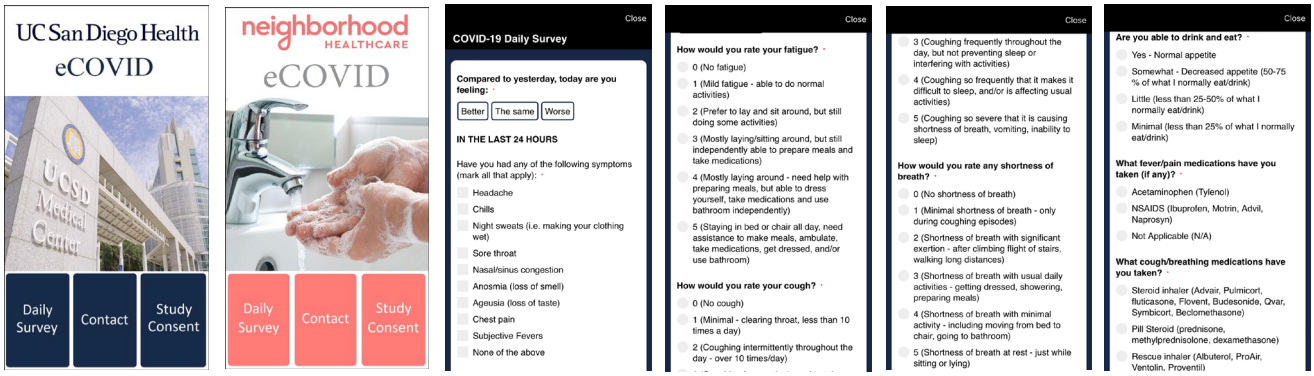

Figure 2. eCOVID symptom tracker app for UCSD Health and Neighborhood Healthcare.

C. *Recovery Detection from COVID-19*

The research presented in [9-16] focused on predicting COVID-19 infection using self-reported symptoms or wearable device data. In contrast to these works, the objective of our research is to estimate recovery from COVID-19 symptoms using wearable device data. The researchers in [17, 18] present different approaches to estimate recovery from COVID-19 infection based on symptoms and demographic data. The authors train a support vector machine [17] and decision tree classifier [18] to estimate patient recovery based on symptoms, demographic, and travel-related features. In [17], the authors found that most of the patients who could not recover experienced a fever, cough and fatigue. In [18], the authors extended their model to predict the number of days needed to

recover from infection. Their model predicted a minimum of 5 days and a maximum of 35 days for COVID-19 patients to recover. Both approaches presented in [17, 18] rely on symptoms data and do not investigate the use of wearable device data for patient recovery estimation. We did not find any previous research that investigates whether wearable device data can be used to estimate patient recovery from COVID-19. This aligns with the observations of the authors in [19] who provide a review on the rise of wearables during the COVID-19 pandemic. None of the works presented in their review are focused on estimating patient recovery from COVID-19 symptoms. This motivates us to develop our own labeling logic for patient recovery in direct consultation with UCSD Health COVID-19 telemedicine doctors. In addition, the dataset we collect consists of a rich feature set spanning activity, sleep, stress, heart rate and SpO_2 data. Our paper provides novel insights into which lifestyle and physiological signals are associated with patient recovery from COVID-19 symptoms.

III. METHOD

In this section, we first detail our study cohort and the proposed remote patient monitoring and reporting system. We then present the ML task of patient recovery classification and discuss its application. Finally, we describe the data preprocessing, the RF model, and our proposed personalization technique.

A. Clinical Study Cohort and eCOVID System

Our IRB-approved clinical study (protocol #181405) was in collaboration with UC San Diego Health and Neighborhood Healthcare, with patient enrollment, onboarding and management conducted by the Altman Clinical & Translational Research Institute at UC San Diego. The study was conducted starting in May 2020. Patients who tested positive for COVID-19 at each location were referred to our study coordinator. Eligible patients were required to be over 18 years old and stable for monitoring in an ambulatory setting, as determined by healthcare personnel at the point of care when testing was initially ordered. The characteristics of the included cohort are shown in Table II. Subjects digitally consented using our symptom tracker mobile app, and those who did were provided a Garmin Vivosmart4 wearable device [20] to collect their lifestyle and vitals data for the study duration of up to 3 months. One of the deciding factors in using this device for this study is its ability to measure blood oxygen saturation (SpO_2). Based on the findings of [21] and our discussion with UCSD Health doctors, SpO_2 is a critical metric in determining the condition of a COVID-19 infected patient. Figure 1 displays the overall architecture of our remote monitoring system, namely eCOVID. The system consists of a symptom tracker mobile app, developed using the Touchwork platform and displayed in Figure 2, and the Garmin device. The daily questions in the symptom tracker app were developed in collaboration with doctors at the UCSD Health COVID-19 telemedicine clinic and are detailed in Table III. The vitals and lifestyle data collected by the Vivosmart4 wearable are detailed in Sec. III (C). Data was collected remotely through the application programming interface (API) provided by Garmin [22].

Table II. Cohort Statistics (n = 30).

	UCSD Health	Neighborhood Health
Total	23	7
# Men	11	3
# Women	12	4
Age (years, mean ± SD)	44.5 ± 13.1	31.6 ± 13.5

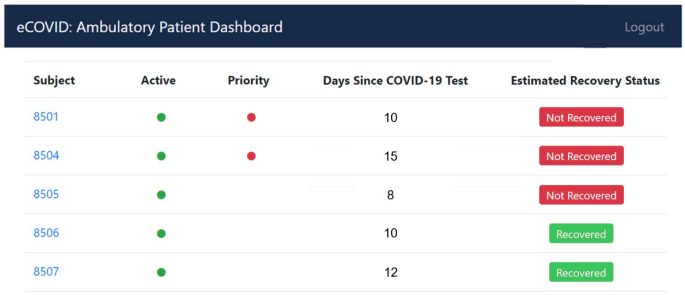

Figure 3. eCOVID dashboard displaying ambulatory COVID-19 patients.

The eCOVID system fused the symptoms and wearable data into a daily report for each patient, which was displayed on our online dashboard for clinicians to review. Both UCSD Health and Neighborhood Healthcare had separate online portals to view their patient data, as displayed in Figure 3. Healthcare workers were able to change the "Active" and "Priority" status in order to indicate which patients needed more timely attention.

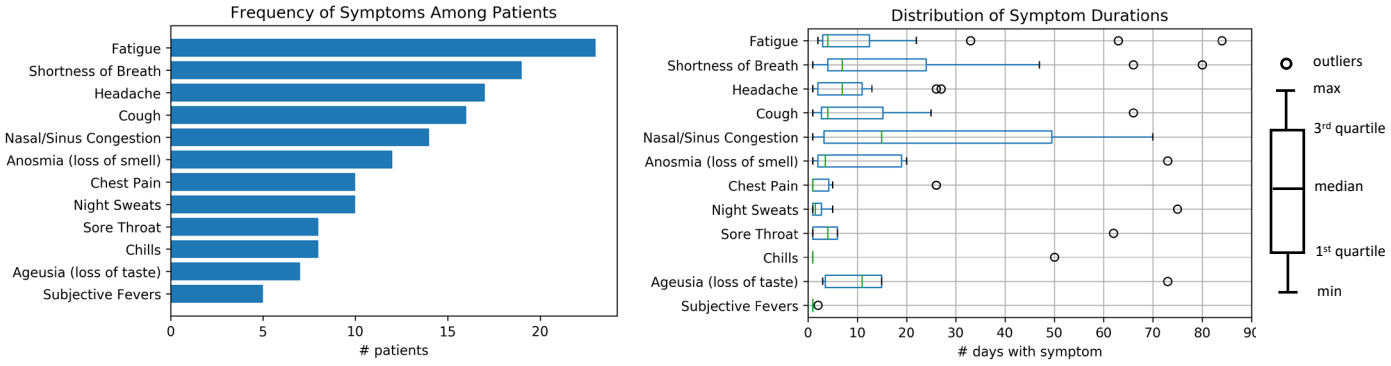

Figure 4. The left plot displays the number of patients who reported at least 1 day of the symptom. The right plot displays the distribution of the number of days each symptom was reported per patient. Only patients who reported the symptom are included in this distribution.

The "Active" and "Priority" statuses are a part of the system design and included to enhance the usability of the dashboard. These statuses are not used for the ML experiment. A detailed report for each patient could be viewed by selecting the patient's ID number. The report included demographic information, symptom tracker app responses, and wearable device data. In addition, the estimated recovery status for each patient using our ML approach was displayed on the dashboard. This demonstrates how our proposed ML approach to patient recovery classification can be effectively incorporated into clinician workflows.

Figure 4 details the distribution of symptoms among patients and describes how long each symptom lasted. For fatigue, shortness of breath and cough, we marked the symptom as present if the patient reported a severity score of 2 or greater. The bar graph in Figure 4 displays the number of patients that experienced each symptom. Fatigue, shortness of breath and headache were the 3 most common symptoms with 23 (77%), 19 (63%) and 17 (57%) patients reporting these symptoms, respectively. Chills, ageusia and subjective fevers were the 3 least common symptoms with 8 (27%), 7 (23%) and 5 (17%) patients reporting these symptoms, respectively. The box plot in Figure 4 details how long each symptom was reported by patients. Only patients who reported the symptom are included in this analysis. Based on the median number of days, nasal/sinus congestion lingered the longest with a median of 15 days followed by ageusia with a median of 11 days. Although ageusia was only reported by 7 patients, the symptom lingered for a longer time compared to other symptoms. Subjective fevers, chills and chest pain were reported for the shortest period of time each having a median of 1 day.

Patients completed the daily symptom tracker an average of 73% of days enrolled in the study. They wore the Garmin device an average of 90% of days enrolled in the study. This indicates that patient compliance with wearing the device was 17% greater than compliance with answering the daily symptom tracker. This statistic demonstrates the higher efficiency of wearable device data for remote monitoring and helps motivate our proposed ML task for patient recovery classification based on automatedly collected device data, as opposed to relying on manually entered symptom data.

B. Patient Recovery Classification

The objective of this ML task is to classify whether a patient has recovered from COVID-19 symptoms based on their device data. This binary classification model can provide healthcare workers with automated insights into the recovery status of their infected patients and bypass the need for manual daily symptom tracking which relies on patient compliance. To the best of our knowledge, there is no clear definition for full recovery from

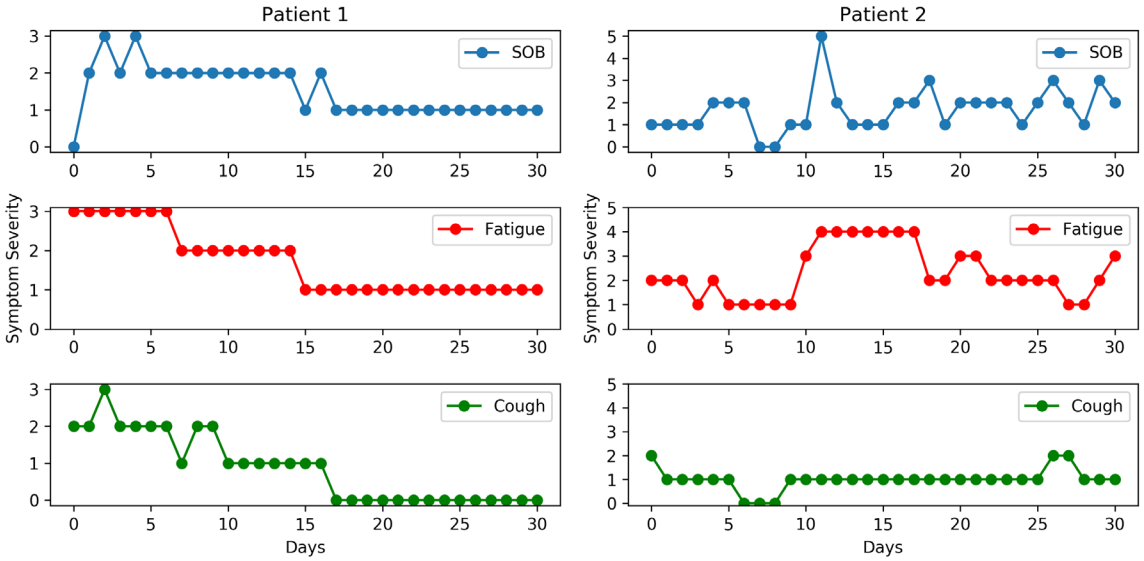

Figure 5. Symptom severity progression for two ambulatory COVID-19 patients. Patient 2's symptom severities decrease by day 7 and then sharply increase again after day 10. The shortness of breath (SOB), fatigue, and cough severities correspond to questions 3-5 of the symptom tracker.

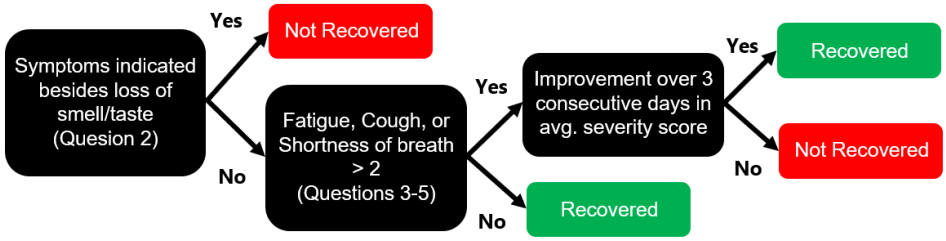

Figure 6. Labeling logic for patient recovery classification based on symptom tracker questionnaire responses.

Table IV. Statistics for label count per patient.

	Mean	Std.	Max	Min	Median
Not Recovered	24	29	85	0	16
Recovered	21	26	76	0	16

COVID-19. The US CDC recommends removal of isolation for COVID-19 infection when a patient's symptoms have significantly improved, they have been afebrile for at least 24 hours in the absence of fever-reducing medications, and it has been at least 10 days since symptom onset [23]. However, it is now understood some patients can suffer from ongoing symptoms from COVID-19 for weeks and even months [24]. Unlike symptom severity which can be identified by patients themselves, recovery is a gradual, subtle and implicit process. In this task, we classify whether a patient has recovered from the COVID-19 symptoms collected by our symptom tracker app. Most patients experienced a steady decline in symptom severities, however, some patients initially appeared to recover and then had symptoms re-appear. Figure 5 displays the symptom severity progression for the first 30 days for two different COVID-19 patients in terms of shortness of breath (SOB), fatigue, and cough. Patient 1 is an example of a patient who experienced a steady recovery. Patient 2, however, demonstrates a complicated symptom progression. The symptom severities for this patient declined by day 7 and then sharply worsened after day 10, especially for SOB and fatigue. All three symptoms linger for this patient for over a month.

A binary label is generated on a daily basis for each patient: recovered (0) or not recovered (1). The labelling logic for patient recovery was developed in collaboration with UCSD Health doctors and is displayed in Figure 6. If symptoms are present besides loss of taste/smell (Question 2), label as not recovered (1). We do not consider loss of smell/taste because these symptoms have been shown to linger after a patient has recovered from COVID-19 [25]. If no symptoms are marked for Question 2 and fatigue/cough/shortness of breath severity is ≤ 2 (Questions 3-5), label as recovered (0). If fatigue/cough/shortness of breath severity is > 2 but there is an improvement over 3 consecutive days in severity scores, label as recovered (0). In order to accommodate for complex cases such as Patient 2 in Figure 5, in which there may be a day labeled as recovered (0) between days labeled as not recovered (1), we apply the following logic. If a patient is labeled as recovered (0) for 7 consecutive days, all the following labels are also marked as recovered (0). Otherwise, the recovered (0) days shorter than 7 days are reverted back to non-recovered (1) days. This ensures there are no "recovered" days between "not recovered" days and vice versa. The statistics of the symptom tracker labels are shown in Table IV. The average number of "not recovered" and "recovered" samples per patient is 24 and 21, respectively. The median number of "not recovered" and "recovered" samples per patient is 16 for both. This difference in mean and median is the result of outlier patients who have a high amount of one label. There are 10 patients for which 90% of their labels are either "not recovered" or "recovered". Patients with few "not recovered" labels may be a result of being asymptomatic or a delay in joining the study after being infected and testing positive. Patients with few "recovered" labels remained symptomatic for the study duration. These labels are used for the patient recovery classification task. Note that the recovery classification technique proposed here can be used with any other labeling logic developed by other health care providers.

C. Device Data and Preprocessing

The Garmin vivosmart4 includes a heart rate monitor, accelerometer, ambient light sensor, and blood oxygen saturation (SpO_2) monitor. The device uses these sensors in order to calculate various health parameters, including lifestyle and vitals information. The device data is presented in Table V and their descriptions are based on the Garmin API documentation [22]. Lifestyle features include activity (steps, distance, floors, active time, etc.), stress (average stress, max stress, stress duration, etc.), sleep timing (duration, bed time, up time), and sleep stages (deep, light, REM, awake). Stress-related features are derived based on heart rate variability [22]. The variable length of time in between each heartbeat is regulated by the body's autonomic nervous system. The less variability between beats equals higher stress levels, whereas the increase in variability indicates less stress. As mentioned in the introduction, the researchers in [12] found that COVID-19 affected the number of daily steps and time asleep for patients in their study. This result motivates us to include all lifestyle features when training our patient recovery classification model. In addition to lifestyle factors, the vivosmart4 measures vitals data including heart rate and SpO_2. The device is capable of manual SpO_2 spot checks during the day and 4 hours of continuous measurement during sleep. Since the symptoms data and patient recovery classification labels are generated on a daily basis, we aggregate the device data features for each day. The Garmin Health API provides summarized activity, sleep, stress and heart rate features on a daily basis. The features in Table V marked with a * require additional processing after receiving the data from Garmin. These include BedTime, UpTime, $MaxSpO_2$, $MinSpO_2$, and $MeanSpO_2$. The BedTime and UpTime features are encoded as the number of seconds before or after midnight (e.g., 11:30 PM bed time is encoded as -1800 seconds, 8:00 AM wake time is encoded as 28800 seconds). Since only the continuous SpO_2 data is available through the Garmin API, we transform the SpO_2 time series each day into the $MaxSpO_2$, $MinSpO_2$, and $MeanSpO_2$ features displayed in Table V. Note that a subset of the features is marked with ^ in Table V indicating they are available in the dataset from [12] which we discuss in Sec. IV (B). Once the device data is aggregated for each day, we match it with the corresponding patient recovery label to form patient-day samples. Each patient-day sample consists of the recovery label and the summarized lifestyle and vitals features for one patient's day in the study. Note that symptoms data are not directly used as part of the training data, but rather to generate the daily patient recovery labels.

Figure 7 displays a heatmap of the correlation between the aggregated daily lifestyle/vitals features and symptoms data for our study cohort. We use Spearman correlation because the symptom variables are not continuous. Spearman evaluates the monotonic relationship between two continuous or ordinal

Table V. Description of Garmin device features that our approach uses. Features marked with * require additional processing after receiving the data from Garmin. Features marked with ^ are available in the dataset from [12] which we discuss in Sec. IV (B).

Features	Description
Steps ^	Count of steps recorded during the monitoring period.
Distance	Distance traveled in meters during the monitoring period.
ActiveTime	Portion of the monitoring period (in seconds) in which the device wearer was considered Active. This relies on heuristics internal to the device.
ModerateIntensityDuration	Cumulative duration of activities of moderate intensity, lasting at least 600 seconds at a time. Moderate intensity is defined as activity with MET value range 3-6.
VigorousIntensityDuration	Cumulative duration of activities of vigorous intensity, lasting at least 600 seconds at a time. Vigorous intensity is defined as activity with MET value > 6.
FloorsClimbed	Number of floors climbed during the monitoring period.
AverageStressLevel, MaxStressLevel, StressDuration, RestStressDuration, ActivityStressDuration, LowStressDuration, MediumStressDuration, HighStressDuration	Stress levels are generated on the device with values ranging from 1 to 100. Scores between 1 and 25 are considered "rest" (i.e., not stressful), 26-50 as "low" stress, 51-75 "medium" stress, and 76-100 as "high" stress. These numbers are derived based on heart rate variability (HRV) and will adjust to the wearer of the device based on the user's natural biometric norms.
SleepDuration ^	Length of the sleep period in seconds.
BedTime *^, UpTime *^	Time the user went to bed and woke up. These are encoded as the number of seconds before or after midnight (e.g., 11:30 PM bed time is encoded as -1800 seconds, 8:00 AM wake time is encoded as 28800 seconds).
DeepSleepDuration ^, LightSleepDuration ^, REMSleepDuration ^, AwakeDuration ^	Time in seconds the user spent in deep/light/REM/awake sleep stage during the sleep period.
MinHeartRate ^, MaxHeartRate ^, MeanHeartRate ^	Minimum/Maximum/Mean of heart rate values captured during the monitoring period, in beats per minute.
RestingHeartRate ^	Average heart rate at rest during the monitoring period, in beats per minute.
MinSpO_2*, MaxSpO_2*, MeanSpO_2*	Minimum/Maximum/Mean of SpO2 values captured during the monitoring period, in percentage. These are calculated from 4 hours of continuous measurement during sleep.

variables [26]. The color of each heatmap square describes the magnitude and directionality of the correlation. Darker red squares correspond to a stronger positive correlation while darker blue squares correspond to a stronger negative correlation. Table VI displays the top 10 most significant correlations between symptoms and device features and in Figure 7 we circle notable correlations in yellow. These include distance and steps vs. fatigue and shortness of breath (SOB) severity, and deep and REM sleep vs. cough and fatigue severity. The correlations for distance vs. SOB and fatigue are -0.38 and -0.37, respectively. The correlations for steps vs. SOB and fatigue are -0.32 and -0.33, respectively. It is sensible that distance and steps are negatively correlated with cough and SOB severity. A patient is less likely to be active if their symptom severities are higher. Deep and REM sleep duration are positively and negatively correlated, respectively, with cough, fatigue and SOB severity. The most significant correlation is deep sleep vs. cough, which has a correlation of 0.47. REM sleep is most correlated with fatigue, with a correlation of -0.34. According the American Academy of Sleep Medicine, as the immune system fights infection, the amount of time spent in REM sleep is decreased while deep sleep is increased [27]. This is because it is during deep sleep that many reparative bodily processes occur. This validates the directionality of the correlations between REM/deep sleep and symptom severities. While the individual correlations between other lifestyle/vitals features and symptoms are not as prominent, the heatmap in Figure 7 indicates that a combination of these features can provide useful information about symptom severity when training the ML model. Overall, these correlation observations help motivate our ML approach to patient recovery classification based on device data.

D. Random Forest and Personalization

We train multiple ML classifiers in order to determine which is most effective at modelling the patient recovery task, as described in Sec. IV (A). As indicated in Table VII, the Random Forest (RF) model results in the best performance during LOSO CV. In this section, we discuss the operation of the RF model and our personalization technique.

RF is an ensemble model that aggregates a collection of decision trees in order to reduce overfitting and the resulting high variance in prediction [28]. To do this, RF utilizes bootstrap aggregation (bagging) and feature bagging. RF produces bootstrap datasets that are randomly and independently drawn with replacement from the training dataset. Each bootstrap dataset has the same size as the original training set and is used to train a decision tree. Bootstrap aggregation in RF averages the prediction of all decision trees which greatly reduces the variance compared to a single decision tree. Moreover, since individual trees generated in the bagging process are identically distributed, the expected prediction of RF is the same as the

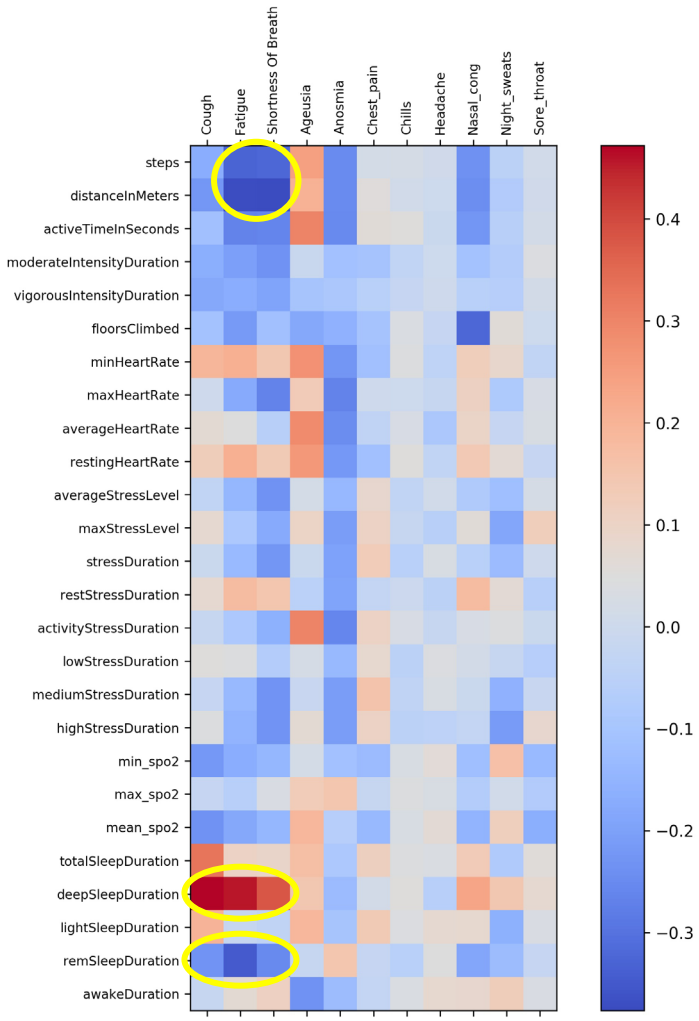

Figure 7. Spearman correlation between lifestyle/vitals and symptoms. Notable correlations are circled in yellow.

expected prediction of individual trees. Combining the above facts, RF has a lower variance than individual trees, while its bias remains the same [29]. RF further reduces the correlation between its member decision trees by introducing feature bagging, which randomly selects a subset of features when constructing each tree. In addition, RF is known to perform well even when using redundant or irrelevant features. Since we utilize multiple lifestyle and vitals features for model training, it is possible that some features do not provide useful information. Since RF is more robust to noisy features as compared to the other models [30], redundant or irrelevant features will not greatly impact performance.

Multiple studies that focus on ML for health applications have shown that model personalization is a key step in improving performance due to the physiological differences between patients [31-34]. In this study, we observe that vitals and lifestyle factors vary among patients and propose a RF-based personalization technique to tune the model to each patient. Our technique involves including the first k days of labeled data from the test patient in the training set. In the traditional RF bootstrapping process, each training sample has uniform weight,

Table VI. Top 10 correlations between symptoms and device features.

Symptom	Device Feature	Spearman Correlation
Cough	DeepSleepDuration	0.47
Fatigue	DeepSleepDuration	0.46
SOB	DeepSleepDuration	0.38
SOB	DistanceInMeters	-0.38
Fatigue	DistanceInMeters	-0.37
Fatigue	REMSleepDuration	-0.34
Cough	TotalSleepDuration	0.33
Fatigue	Steps	-0.33
SOB	Steps	-0.32
Nasal Congestion	FloorsClimbed	-0.32

which means each data sample is resampled with the same probability. To emphasize the test patient's calibration samples during model training, we assign a greater weight to these k samples using the Weighted Bootstrapping algorithm [35]. In order to implement this algorithm, a vector of sample weights $\boldsymbol{W} = \boldsymbol{w_1}, \boldsymbol{w_2}, \ldots, \boldsymbol{w_N}$ is maintained where N is the total number of training samples. Weights $\boldsymbol{w_1}, \ldots, \boldsymbol{w_k}$ correspond to the k personalization samples from the test patient and are given larger values. Weights $\boldsymbol{w_{k+1}}, \ldots, \boldsymbol{w_N}$ correspond to the data samples from the remaining patients used for training and are assigned lower values. The operation of the Weighted Bootstrapping algorithm is as follows [35]: In step 1, a new bootstrap dataset for one decision tree is initialized. In step 2, the weights in $\boldsymbol{W}$ are mapped into the interval $\left[\boldsymbol{0},\ \sum_{j=1}^{N} \boldsymbol{w_j}\right]$ with subintervals $\boldsymbol{I_1}, \boldsymbol{I_2}, \ldots\ \boldsymbol{I_N}$. The length of each subinterval is proportional to the value of its weight. In steps 3 to 7, each data sample is drawn using subintervals $\boldsymbol{I_1}, \boldsymbol{I_2}, \ldots\ \boldsymbol{I_N}$ and the uniform distribution function. The process repeats N times such that the size of all bootstrap datasets equals that of the original dataset. Consequently, the samples with higher weights are more likely to appear in each bootstrap dataset. In Sec. IV (B), we compare

Weighted Bootstrapping Algorithm

Input: Training dataset $\boldsymbol{X} = \{x_i, i = 1,2 \ldots N\}$, a sequence of N examples, and weights of the N examples $\boldsymbol{W} = \{w_i, i = 1,2 \ldots N\}$, $w_i \in [0,1], \forall i$

1: Create a new dataset $\boldsymbol{X'}$ with the same size as $\boldsymbol{X}$

2: Partition the interval $\left[0, \sum_{j=1}^{N} w_j\right]$ into N subintervals $I_1 = (0, w_1)$, $I_2 = (w_1, w_1 + w_2)$, $\ldots I_N = \left(\sum_{j=1}^{N-1} w_j, \sum_{j=1}^{N} w_j\right)$

3: **for** $i = 1$ **to** N **do**

4: Simulate $u \sim U\left(0, \sum_{j=1}^{N} w_j\right)$, U is uniform distribution function where the probability density of $U(a, b)$ is $f(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b \\ 0, & \text{otherwise} \end{cases}$

5: Identify the interval $I_{j^*}, j^* \in \{1,2 \ldots N\}$ such that $u \in I_{j^*}$

6: Add sample x_{j^*} to $\boldsymbol{X'}$

7: **end for**

Output: $\boldsymbol{X'}$

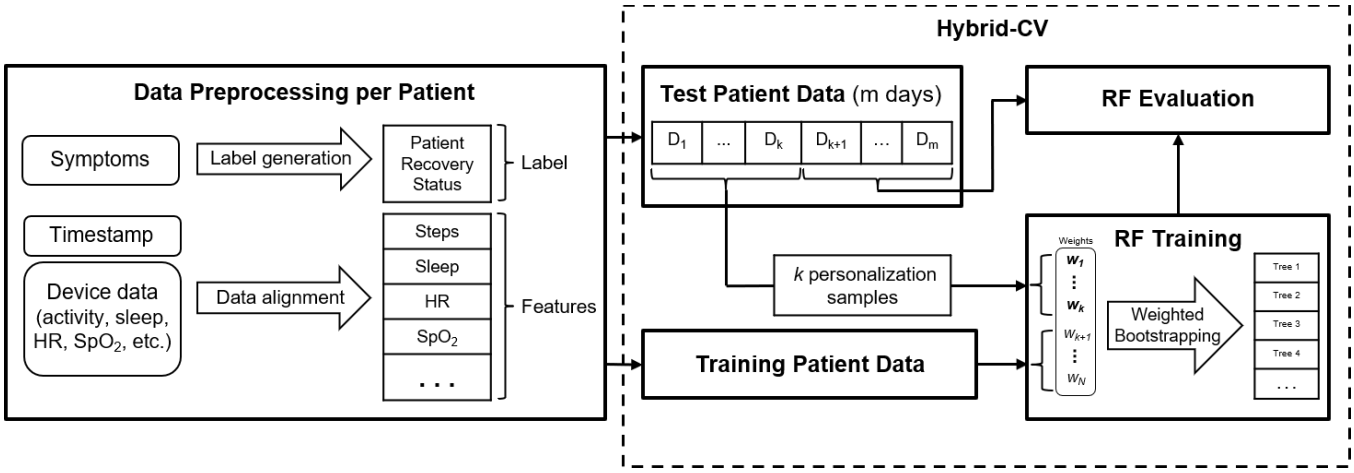

Figure 8. Block diagram of our proposed RF personalization approach. After data preprocessing, the first *k* samples from the test patient are included in the training set during Hybrid-CV. These samples are assigned larger weights, which are bolded in the figure, during weighted bootstrap aggregation. After training, the model is evaluated on the remaining test patient data samples.

the performance for different values of *k* and different values of $\boldsymbol{w_1, \ldots, w_k}$. Figure 8 displays a block diagram of our proposed RF personalization technique. After preprocessing each patient's data, Hybrid-CV is carried out in which the training and test sets are split on a per patient basis and the first *k* days of test patient data are added to the training set as personalization samples, as shown in Figure 8. These *k* samples are assigned greater weights, which are bolded in the figure, during weighted bootstrapping. After training, the model is evaluated on the remaining, future data samples of the test patient.

IV. Results and Discussion

In this section, we describe the experiment settings and present patient recovery classification results. We discuss the effects of our RF model personalization technique on performance and carry out feature analysis using Shapley Values in order to interpret what the model has learned. Finally, we provide a discussion on the challenges encountered during this study.

A. Experiment Setting

We implement and evaluate our machine learning models using the Scikit-learn library in the python environment on an Intel i5 3.2GHz quad-core and 16GB RAM computer. Accuracy, sensitivity, specificity, and F1-score are calculated and used as our evaluation metrics for the patient recovery classification task. For this task, a negative and positive sample correspond to a "recovered" and "not recovered" patient-day sample, respectively. Accuracy returns an overall measure of how much the model is correctly predicting on the entire set of test data. Sensitivity and specificity measure the true positive and true negative rate, respectively. F1 score is calculated as the harmonic mean of precision and recall (sensitivity) and is used to find the best trade-off between the two quantities [36]. As a result, we use F1 score for deciding the top performing model.

We carry out LOSO CV to mirror the clinically relevant use-case scenario of diagnosis for newly infected subjects [43]. LOSO CV separates the data into train and test sets on a per patient basis in order to simulate the practical application. This data split ensures that data from the same patient does not appear in both the training and testing sets. We use LOSO CV to compare the performance of different ML models. We then carry out Hybrid-CV, in which a specified number of samples from the test patient are included in the training set. These personalization samples are not included in the test set to ensure there is no overlap between train and test sets at the sample level. We compare how performance is affected by applying varying levels of personalization using our RF-based personalization technique described in Sec. III (D). Since the number of samples for each patient is different based on their participation in the study, the training and testing sets will vary in size for both CV experiments. Instead of averaging the results over each data split, we save the model predictions for each data split and calculate metrics over all predictions. This ensures that each patient-day contributes equal weight to the final result.

In the LOSO CV experiment, we compare RF with the following ML models: logistic regression (LR) [37], *k*-nearest neighbors (KNN) [38], support vector machine (SVM) [39], artificial neural network (ANN) [40], and long short-term memory (LSTM) neural network [41]. Model hyperparameter tuning is performed with each training set using a randomized search over a predefined hyperparameter grid for each model. Since LSTM models take sequential data as input, we organize the lifestyle and vitals features into sequential data samples using a window length of 7 days and a step size of 1 day. A step size of 1 day is used to extract the maximum number of samples. As a result, each input sample has a dimension of ($7, N_{features}$) where $N_{features}$ represents the number of lifestyle and vitals features. The patient recovery label for the last day of each window is assigned to each input sample. We train the LSTM as a many-to-one model, as opposed to a many-to-many model, since the application of this method is only concerned with estimating whether the patient is recovered or not for the current day. In addition, training the LSTM to estimate one label at a time matches the process for the other ML models, resulting in a fairer

comparison. We carry out two LSTM experiments using 16 and 32 hidden units for the LSTM layer followed by a fully connected layer with 1 output unit. For these experiments, we train the models using the Adam optimizer [42] and a dropout rate of 50% to reduce overfitting. For the LSTM layers, we use a sigmoid activation function for the input, forget and output gates, and a hyperbolic tangent (tanh) activation function for the cell state and hidden state. The fully connected layers use a sigmoid activation function and we use binary cross entropy loss as the loss function. We experimented with different numbers of training epochs and batch sizes and found that 25 epochs and a batch size of 32 resulted in the best performance.

B. Patient Recovery Classification Results

Accuracy, sensitivity, specificity, and F1-score for each ML model during LOSO CV are presented in Table VII. The LSTM-32 model achieves the highest accuracy and sensitivity, both equal to 0.64, while the RF model achieves the highest specificity and F1 score equal to 0.78 and 0.66, respectively. As described in the experiment setting, we use F1 score for deciding the top performing model since this metric calculates the tradeoff between precision and sensitivity. Since RF achieves the highest F1 score, we conclude that RF is the best performing model for patient recovery classification. We attribute the RF's top performance to its ability to reduce the variance in prediction via the bagging process and its robustness to redundant or irrelevant features. The LSTM-32 model is the second-best performing model, indicating that meaningful temporal information exists in the data for estimating recovery from COVID-19. Since RF is the top performer, we use this model in the next experiment to understand how the number of personalization samples impacts RF performance.

Next, we discuss the results of the Hybrid-CV experiment. As mentioned in the experiment settings, LOSO CV separates the data into train and test sets on a per patient basis. Since physiology and lifestyle differ between patients, we apply varying levels of personalization during the Hybrid-CV experiment. We implement our RF-based personalization technique by including the first 1-5 days of test patient data in the training set. These personalization samples are assigned a larger weight so that they are sampled more frequently during the bootstrap aggregation step. Table VIII displays the results for different amounts of personalization. Evidently, the classification results are worse when no personalization is applied. The accuracy, sensitivity, specificity and F1-score are 0.59, 0.52, 0.78, and 0.66, respectively, when no personalization is applied. As personalization samples are included in the training set, accuracy, sensitivity and F1-score increase, while specificity decreases. When using 5 personalization samples, the accuracy, sensitivity, specificity and F1-score are 0.82, 0.89, 0.63, and 0.88, respectively. Since the personalization samples for each patient correspond to their first 1-5 days in the study, these samples are primarily labeled 1 or "not recovered". This means that as more personalization samples are included in the training set, the model is able to increasingly learn the infected baseline of the patient based on their vitals and lifestyle data. This causes the sensitivity to increase since the model will be able to increasingly correctly classify a patient who has not recovered. This corresponds to increasing true positives (classifying a patient as not recovered when they are indeed not recovered) while minimizing false negatives (classifying a patient as recovered when they are not recovered). As the sensitivity increases, the specificity decreases. Since the model is increasingly tuned to classify a patient as not recovered, this will result in more false positives and a lower specificity. For this ML task, false positives are more acceptable than false negatives. Classifying a patient as not recovered when they actually are recovered is less harmful than classifying a patient as recovered when they are not recovered. Overall, adding personalization samples increases the model performance. When applying this personalization technique to a new patient, the first few days will involve data collection without any classifications from the ML model. After this initial data collection, the personalized model will provide estimations with improved accuracy, sensitivity and F1-score. The results demonstrate the potential for ML-assisted remote patient monitoring to

Table VII. Comparison of ML model performance for LOSO CV.

Model	Acc	Sens	Spec	F1
LR	0.60	0.61	0.52	0.61
ANN	0.59	0.62	0.62	0.63
SVM	0.54	0.61	0.59	0.62
KNN	0.55	0.51	0.68	0.60
LSTM-16	0.63	0.56	0.71	0.61
LSTM-32	**0.64**	**0.64**	0.60	0.64
RF	0.59	0.52	**0.78**	**0.66**

Table VIII. Hybrid-CV results using different levels of personalization

Personalization Samples	Acc	Sens	Spec	F1
0	0.59	0.52	0.78	0.66
1	0.63	0.59	0.75	0.70
2	0.67	0.66	0.71	0.75
3	0.72	0.73	0.68	0.79
4	0.80	0.86	0.64	0.86
5	0.82	0.89	0.63	0.88

Table IX. Performance comparison when applying different RF bootstrap aggregation weights to 5 personalization samples.

Bootstrap Aggregation Weights	Acc	Sens	Spec	F1
1	0.70	0.69	0.73	0.77
10	0.82	0.89	0.63	0.88
100	0.81	0.88	0.62	0.87

Table X. Evaluation of proposed method on open dataset from [12].

	Acc	Sens	Spec	F1
W/O Personalization	0.49	0.33	0.73	0.44
W/ Personalization (5 samples)	**0.61**	**0.55**	**0.67**	**0.61**

supplement traditional manual monitoring tools, like daily manual symptom tracking.

The results presented in Table VIII are generated by setting the bootstrap aggregation weights for the personalization samples to 10. This means these samples are 10 times more likely to be sampled during the RF weighted bagging process. In Table IX, we compare how classification performance is affected by applying different bagging weights to 5 personalization samples. We set the weights to 1, 10 and 100. Using a bagging weight of 1 means the personalization samples have the same probability of being sampled as the training data from other patients. Evidently, a bagging weight of 1 produces worse performance with an accuracy, sensitivity, specificity and F1-score of 0.7, 0.69, 0.73, and 0.77, respectively. In this case, the personalization samples are not emphasized and the model is not effectively calibrated. Increasing the bagging weight from 10 to 100 does not improve model performance. This indicates that at a certain weight, the personalization samples are sampled frequently enough during bagging to effectively calibrate the model. Further increasing the bagging weight does not provide additional utility in model personalization.

In order to extend the evaluation of our proposed method, we applied our approach to the dataset collected in [12]. This dataset includes sleep, heart rate and steps data collected from a wearable device, and the date of first symptoms and date of recovery which are manually recorded by each patient. Since this dataset does not include SpO2, stress or activity (besides steps) data, the number of features is significantly less than our own dataset (12 vs. 28). In Table V, features marked with ^ are available in the dataset in [12]. We labelled all days between the start of symptoms and recovery dates as "not recovered" and all days after the recovery date as "recovered". We then combined these labels with the corresponding device features to create the dataset in the same manner as our experiment setting. After these data processing steps, 15 patients had sufficient data to be included in this experiment. Table X displays the results when applying our method to this dataset. We train a Random Forest model with and without personalization and calculate the accuracy, sensitivity, specificity and F1-score. We use 5 samples when applying our personalization technique and observe that the performance significantly improves compared to the non-personalized results. With personalization, our approach achieves an accuracy, sensitivity, specificity and F1-score of 0.61, 0.55, 0.67, and 0.61, respectively. Evidently, the performance metrics are not as good for this dataset. This may be due to the limited feature set and inaccurate recovery dates recorded by patients. We observe similar patterns in the results compared with our own dataset which include that there is a performance enhancement when applying our personalization technique. Overall, these consistent observations between our dataset and the dataset in [12] indicate that our proposed approach is not only applicable to our dataset, but can potentially be applied to different datasets collected in clinical practice.

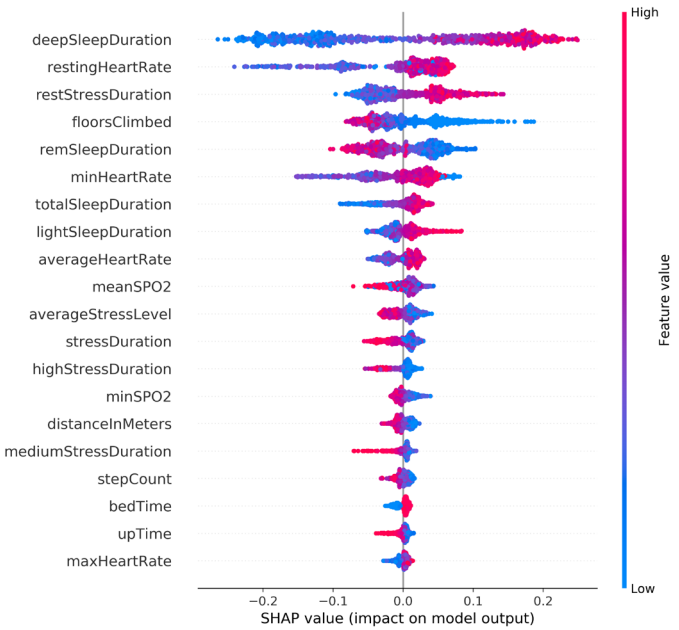

Figure 9. Summary of Shapley top features where each point corresponds to a data sample. The x-axis represents a feature's impact on model output. Positive SHAP values push the model to output 1 or "not recovered".

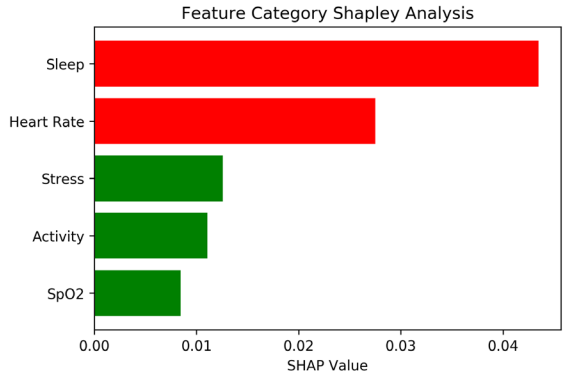

Figure 10. Impact of feature categories on model output. Features are grouped into 5 categories and a categorical SHAP score is calculated. Red or green bars indicate that an increase in the category's feature values pushed the model to output "not recovered" or "recovered", respectively.

C. Model Interpretability via Shapley Value Analysis

Next, we utilize Shapley Values [44, 45] in order to determine which lifestyle and vitals features have the most significant effect on model classification for our dataset. Shapley Value analysis is a model-agnostic interpretation method derived from game theory. Given a set of feature values and a trained machine learning model, the estimated Shapley value indicates how each feature contributes to the model's classification. We use the tree SHAP (SHapley Additive exPlanations) framework [46, 47], which is optimized for tree-based models, to interpret the output of the RF model for patient recovery classification. Figure 9 displays the Shapley results where the features are ranked from

the top to bottom based on their impact on the model's output. Each point on the plot corresponds to an individual data sample and represents the contribution from the feature listed on the Y-axis to the RF's classification. The placement on the X-axis represents the amount of positive/negative contribution to the classification. Positive contribution corresponds to pushing the model to estimate that a patient is not recovered. The color of each point represents the actual value of the feature (red is high while blue is low). The top two features based on Shapley analysis include deep sleep duration and resting heart rate. Higher values of deep sleep duration (colored in red) contribute to a positive, or not recovered, classification. This observation aligns with the correlation analysis presented in Sec. III (C). As mentioned earlier, deep sleep increases when a patient is sick since this is when many reparative bodily processes occur. Increased resting heart rate also contributes to a positive classification by the RF model. This relationship makes sense since resting heart rate will decrease as a patient recovers. Additional observations include that a lower number of floors climbed contributes to a positive classification while an increased mean SpO_2 contributes to a negative, or recovered, classification. Both of these relationships are sensible, as a patient who has not recovered will be less active and a patient who has recovered will have a higher SpO_2.

In addition to analyzing the impact of individual features, we grouped the features into 5 categories (Activity, Sleep, Stress, Heart Rate and SpO2) and investigated their impact on model output. A SHAP score for each category was calculated as the average of the absolute SHAP values for the features in that category. Figure 10 displays the ranking of feature categories based on their categorical SHAP score. We also examined whether, on average, an increase in the feature values for each category pushed the model to estimate "recovered" or "not recovered". In Figure 10, a red colored bar indicates that an increase in the category's feature values pushed the model to output "not recovered". A green colored bar indicates that an increase in the category's feature values pushed the model to output "recovered". Evidently, the sleep category had the most significant impact on model output. An increase in feature values in the sleep and heart rate categories pushed the model to estimate "not recovered" (red bars) while an increase in feature values in the stress, activity and SpO2 categories pushed the model to estimate "recovered" (green bars). Overall, the individual feature and feature category Shapley analysis demonstrates that our model can learn clinically relevant relationships between device data and the status of patients. The interpretability of a ML model is necessary for humans to understand what the model has learned, especially in medical applications.

D. Limitations and Research Challenges Encountered

In this section, we discuss limitations to our proposed approach and challenges faced while implementing this study. One limitation in our approach is that patients were only enrolled and provided devices for data collection after testing positive for COVID-19. It is likely that some patients started experiencing symptoms before going for a COVID-19 test. This meant we were not able to collect symptoms and wearable data during the initial days of the infection. In order to ensure that data can be collected before and during the onset of COVID-19 infection, participation could be made available to a larger number of patients that already own a wearable device. After testing positive for COVID-19, a patient could immediately enroll and begin sharing both past and current data. Another limitation to our approach is that the RF model does not process data sequentially while the progress of COVID-19 is sequential. In this work, we experimented with LSTM, a popular temporal model, however, found its performance to be worse than RF. Training an LSTM requires significantly more data since neural networks are highly prone to overfitting when the underlying dataset size is small [48, 49]. In order to fully utilize temporal relationships in the data, we plan to further investigate sequence modeling with additional data in our future work. This will include implementing many-to-many sequence models using different time windows to learn temporal progression along with the label. In addition, a larger dataset can enable the use of additional features such as patient demographic information. The model may learn relationships between recovery from COVID-19 and demographic data such as age, gender and ethnicity.

Concern over privacy was an issue encountered during recruitment for this study. As mentioned in Sec. III (A), we recruited patients from both the UCSD Health and Neighborhood Healthcare (NH) COVID-19 telemedicine clinics. NH is a community clinic that primarily provides care to underserved populations. In order to increase accessibility to our study, we developed a Spanish version of our symptom tracker app with assistance from NH. Overall, we experienced more difficulty recruiting from NH. One reoccurring reason why NH patients did not want to partake in our study included a concern over privacy. Certain patients expressed discomfort over wearing the device 24/7 due to concerns of being tracked. Our recruitment personnel would highlight that the device does not collect any location data, however, certain patients still declined participation. The above challenge encountered during our study showed that privacy concerns and lack of trust in wearables may further limit access and use of digital technologies by underserved populations, contributing to an increased digital divide in healthcare. As healthcare begins to rely more on digital technologies, these concerns must be addressed in order to ensure equal access to high quality healthcare [50].

V. CONCLUSION

In this paper, we propose an intelligent remote monitoring platform, namely eCOVID, for enhanced COVID-19 ambulatory care. Based on data collected from our study with the UCSD Health and Neighborhood Healthcare COVID-19 telemedicine clinics in San Diego County, we demonstrate correlations between automatically collected wearable data and manually entered symptom data. We propose a novel ML approach to estimate whether or not a patient has recovered from COVID-19 symptoms based on the automatically collected wearable data. Our results demonstrate that ML-assisted remote monitoring using wearable

data can supplement or be used in place of manual daily symptom tracking which relies on patient compliance.

By developing and demonstrating the ability to track patient recovery status remotely, our approach can enable more optimal care of COVID-19 ambulatory patients at scale. Care teams will be able to track patient recovery efficiently through automatically generated and updated dashboards instead of the current practice of manual symptom tracking and phone calls, the latter becoming ineffective when there is a surge in cases. This shift can lead to significant improvement in the efficiency and scalability of ambulatory patient care, while at the same time enabling savings in human and equipment resources. Moreover, the approach can be used for providing scalable and efficient care for future pandemic and epidemic challenges.

ACKNOWLEDGMENT

The research reported in this paper was supported by the Connected Health Program of the UC San Diego Center for Wireless Communications and its member companies, and the Jacobs Family Endowed Chair fund for Professor Sujit Dey. We would like to thank Dr. Rakesh Patel, Jessica Gatien, and Azucena Maldonado from Neighborhood Healthcare, and Jeff Mills and Kathleen Bordeaux from UCSD Health for their assistance in patient enrollment for this study.

DATA AVAILABILITY

The de-identified data used in this study can be downloaded from the study data repository on IEEE DataPort (https://ieee-dataport.org/). This includes the symptom tracker responses and data collected from Garmin devices.

REFERENCES

[1] "WHO Coronavirus Disease (COVID-19) Dashboard." World Health Organization, covid19.who.int/.

[2] "CDC COVID Data Tracker." Centers for Disease Control and Prevention, covid.cdc.gov/covid-data-tracker/.

[3] "COVIDView: A Weekly Surveillance Summary of U.S. COVID-19 Activity." Centers for Disease Control and Prevention, www.cdc.gov/coronavirus/2019-ncov/covid-data/covidview/index.html.

[4] D. M. Roblyer, "Perspective on the increasing role of optical wearables and remote patient monitoring in the COVID-19 era and beyond," in *Journal of Biomedical Optics* 25(10) 102703 (21 October 2020) https://doi.org/10.1117/1.JBO.25.10.102703

[5] A. Mahajan, G. Pottie, and W. Kaiser, "Transformation in Healthcare by Wearable Devices for Diagnostics and Guidance of Treatment," in *ACM Trans. Comput. Healthcare* 1, 1, Article 2 (February 2020), 12 pages. DOI: https://doi.org/10.1145/3361561

[6] M. M. Islam, S. Mahmud, L. J. Muhammad, M. R. Islam, S. Nooruddin and S. I. Ayon, "Wearable technology to assist the patients infected with novel coronavirus (COVID-19)", *Social Netw. Comput. Sci*., vol. 1, no. 6, pp. 320, Nov. 2020.

[7] N. Ji et al., "Recommendation to Use Wearable-based mHealth in Closed-Loop Management of Acute Cardiovascular Disease Patients during the COVID-19 Pandemic," in *IEEE Journal of Biomedical and Health Informatics*, doi: 10.1109/JBHI.2021.3059883.

[8] M. Au-Yong-Oliveira, A. Pesqueira, M. J. Sousa, F. D. Mas, and M. Soliman, "The Potential of Big Data Research in HealthCare for Medical Doctors' Learning," in *Journal of Medical Systems* 45, 13 (2021). https://doi.org/10.1007/s10916-020-01691-7

[9] C. Menni, A. M. Valdes, M.B. Freidin, et al., "Real-time tracking of self-reported symptoms to predict potential COVID-19," *Nat Med*, vol. 26, pp. 1037-1040, 2020.

[10] M. Zens, A, Brammertz, J. Herpich, N. Sudkamp, and M. Hinterseer, "App-based tracking of self-reported covid-19 symptoms: Analysis of questionnaire data," *Journal of Medical Internet Research*, 22(9):e21956, 2020.

[11] Y. Zoabi, S. Deri-Rozov, and N. Shomron, "Machine learning-based prediction of covid-19 diagnosis based on symptoms," *npj digital medicine*, vol. 4, no. 1, pp. 1–5, 2021.

[12] T. Mishra et al., "Pre-symptomatic detection of COVID-19 from smartwatch data", *Nat. Biomed. Eng*., vol. 4, no. 12, pp. 1208-1220, 2020.

[13] Alavi, A., Bogu, G.K., Wang, M. et al. Real-time alerting system for COVID-19 and other stress events using wearable data. Nat Med (2021). https://doi.org/10.1038/s41591-021-01593-2

[14] G. Quer, J.M. Radin, M. Gadaleta, et al., "Wearable sensor data and self-reported symptoms for COVID-19 detection," *Nat Med* (2020). https://doi.org/10.1038/s41591-020-1123-x

[15] A. Natarajan, H.W. Su, and C. Heneghan, "Assessment of physiological signs associated with COVID-19 measured using wearable devices," *npj Digit. Med.* 3, 156 (2020). https://doi.org/10.1038/s41746-020-00363-7

[16] Gadaleta, M., Radin, J.M., Baca-Motes, K. et al. Passive detection of COVID-19 with wearable sensors and explainable machine learning algorithms. npj Digit. Med. 4, 166 (2021). https://doi.org/10.1038/s41746-021-00533-1

[17] A. Salama, A. Darwsih, and A.E. Hassanien, "Artificial Intelligence Approach to Predict the COVID-19 Patient's Recovery," *Digital Transformation and Emerging Technologies for Fighting COVID-19 Pandemic: Innovative Approaches*, vol. 322, pp. 121-133, 2021.

[18] L.J. Muhammad, M.M. Islam, S.S. Usman, and S.I. Ayon, "Predictive Data Mining Models for Novel Coronavirus (COVID-19) Infected Patients' Recovery," *SN COMPUT. SCI.* 1, 206 (2020).

[19] Channa, A., Popescu, N., Skibinska, J., & Burget, R. (2021). The rise of wearable devices during the COVID-19 pandemic: A systematic review. Sensors, 21(17), 5787.

[20] Garmin, and Garmin Ltd. or its subsidiaries. "Garmin Vivosmart® 4: Fitness Activity Tracker: Pulse Ox." Garmin, buy.garmin.com/en-US/US/p/605739.

[21] S. Shah, K. Majmudar, A. Stein, et al., "Novel use of home pulse oximetry monitoring in COVID-19 patients discharged from the emergency department identifies need for hospitalization," *Acad Emerg Med* 2020; 27. doi: https://doi.org/10.1111/acem.10453

[22] "Overview: Health API: Garmin Developers." Overview | Health API | Garmin Developers, developer.garmin.com/health-api/overview/.

[23] "When You Can Be Around Others After You Had or Likely Had COVID-19." *Centers for Disease Control and Prevention*, www.cdc.gov/coronavirus/2019-ncov/if-you-are-sick/end-home-isolation.html.

[24] "Post-COVID Conditions." *Centers for Disease Control and Prevention*, www.cdc.gov/coronavirus/2019-ncov/long-term-effects.html.

[25] S. Ikegami, R. Benirschke, T. Flanagan, et al., "Persistence of SARS-CoV-2 nasopharyngeal swab PCR positivity in COVID-19 convalescent plasma donors," *Transfusion.* 2020; 60: 2962– 2968. https://doi.org/10.1111/trf.16015

[26] Spearman Rank Correlation Coefficient. In: The Concise Encyclopedia of Statistics. Springer, New York, NY. https://doi.org/10.1007/978-0-387-32833-1_379

[27] J. Cline. "Flu Season and Sleep." *Psychology Today*, Sussex Publishers, 31 Dec. 2014.

[28] L. Breiman, "Random Forest," *Machine Learning*, vol. 45, no.1, pp. 5–32, 2001

[29] B. Efron and R. Tibshirani, "An introduction to the bootstrap," *CRC Press*, 1994.

[30] V. F. Rodriguez-Galiano, B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez, "An assessment of the effectiveness of a random forest classifier for land-cover classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 67, pp. 93–104, 2012.

[31] P. Chiang and S. Dey, "Personalized Effect of Health Behavior on Blood Pressure: Machine Learning Based Prediction and Recommendation," in *Proc. of IEEE International Conference on E-health Networking, Application & Services (Healthcom'18),* Ostrava, Czech, 2018

[32] J. Leitner, P. Chiang and S. Dey, "Personalized Blood Pressure Estimation Using Photoplethysmography: A Transfer Learning Approach," in *IEEE Journal of Biomedical and Health Informatics*, doi: 10.1109/JBHI.2021.3085526.

[33] C. L. Stewart, A. Folarin and R. Dobson, "Personalized acute stress classification from physiological signals with neural processes", 2020.
[34] D. Lopez-Martinez and R. Picard, "Multi-task neural networks for personalized pain recognition from physiological signals," *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, San Antonio, TX, 2017, pp. 181-184, doi: 10.1109/ACIIW.2017.8272611.
[35] P. Chiang and S. Dey, "Offline and Online Learning Techniques for Personalized Blood Pressure Prediction and Health Behavior Recommendations," in *IEEE Access*, vol. 7, pp. 130854-130864, 2019, doi: 10.1109/ACCESS.2019.2939218.
[36] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi-class classification: an overview," arXiv preprint arXiv:2008.05756, 2020.
[37] R. E. Wright, "Logistic regression," in L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (p. 217–244). American Psychological Association.
[38] A. Mucherino, P.J. Papajorgji, and P.M. Pardalos, "K-nearest neighbor classification," *Data mining in agriculture*. Springer, New York, NY, 2009. 83-106.
[39] J. A. K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Process. Lett.*, vol. 9, no. 3, pp. 293–300, 1999.
[40] D. Rumelhart, G. Hinton, and R. Williams, "Learning representations by back-propagating errors," *Cogn. Model*., vol. 5, no. 3, p. 1, 1988.
[41] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, pp. 1735-1780, 1997.
[42] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014
[43] S. Saeb, L. Lonini, A. Jayaraman, et al., "The need to approximate the use-case in clinical machine learning," *Gigascience* 2017;6(5):1–9.
[44] L. Shapley, "A value for n-person games." *Contributions to the Theory of Games*, vol. 2, no. 28, pp. 307-317, 1953.
[45] S. Cohen, E. Ruppin and G. Dror, "Feature Selection Based on the Shapley Value," in *International Joint Conferences on Artificial Intelligence*, vol. 5, pp. 665-670, 2005.
[46] S. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *Advances in neural information processing systems*, pp. 4765-4774, 2017.
[47] S. Lundberg, G. Erion, H. Chen, et al., "From local explanations to global understanding with explainable AI for trees," *Nat Mach Intell* 2, pp. 56-67, 2020.
[48] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *The journal of machine learning research*, vol. 15, no. 1 pp. 3133-3181, 2014.
[49] S. Wang, C. Aggarwal, and H. Liu. "Using a random forest to inspire a neural network and improving on it," In Proceedings of the *SIAM international conference on data mining*, Houston, TX, USA, 2017.
[50] A. Ramsetty and C. Adams, "Impact of the digital divide in the age of COVID-19," *Journal of the American Medical Informatics Association*, vol. 27, pp. 1147-1148, 2020.