Deep learning-based detection of COVID-19 using wearables data

Gireesh K. Bogu[1*], Michael P. Snyder[1*]

1. Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA.

*Corresponding authors, gireesh.bogu@stanford.edu, mpsnyder@stanford.edu

COVID-19 is an infectious disease caused by SARS-CoV-2 and primarily diagnosed based on laboratory tests and usually after the symptom onset[1]. However, SARS-CoV-2 replication starts before the symptom onset and is released into the environment and causes further viral transmission[2,3]. Here, we have developed a deep learning approach based on Long Short-Term Memory Networks-based autoencoder - called LAAD to predict the COVID-19 in each user by using their wearable data. We detected the abnormal resting heart rate during the time of viral infection (7 days before the symptom onset and 21 days after) in 96% (24 out of 25 cases) of the patients with laboratory-confirmed COVID-19. We estimated that 84% (21 cases) of the COVID-19 patients were pre-symptomatic and 12% (3 cases) were post-symptomatic with an average precision score of 0.81, recall of 0.59, and F-beta score of 0.79. In COVID-19 positive patients, abnormal RHR starts 5.26 days (median = 6.90) before the symptom onset (6.20 days (median = 6.96) in pre-symptomatic and 1.36 days (median = 1.75) later in post-symptomatic cases). COVID-19 positive patients have longer abnormal resting heart rate (median = 7.5 days) during the infectious period compared to non-COVID-19 illness (median = 5.5 days) and healthy individuals (median = 3.9 days). These findings suggest that deep learning neural networks and wearables data can be used to detect COVID-19 early and it may help to prevent pre-symptomatic transmission.

COVID-19 is a contagious respiratory disease caused by the novel coronavirus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [1]. As of 6 December 2020, over 66.7 million people infected world-wide and 14.6 million of them are from the U.S alone, and still the number of cases are in rise[4]. COVID-19 testing, shelter-in-place and social distancing have been shown to be effective but not enough to slow down the virus transmission. Viral load (number of viral particles) and duration of viral shedding (where the virus replicates inside your body and is released into the environment) are important determinants for COVID-19 transmission[2,3]. Based on the current literature, the viral shedding of SARS-CoV-2 may begin 5 to 6 days before the appearance of first symptoms and decline after 14.6 days to 17.2 days following the symptom onset[2,3]. This means that people can be infectious during the pre-symptomatic stage, in the days before they develop symptoms. Whereas, viral load peaks in the first week after symptoms onset making it a highly infectious period[2]. Therefore, tracking symptoms alone from patients cannot help pre-symptomatic viral transmission.

Tracking data from wearable devices has shown to be promising in detecting COVID-19[5–9]. Wearable devices contain sensors that collect different types of data like heart rate, steps, and sleep and can be used to track viral infections early using statistical methods like cumulative sum (CUSUM), RHR-Diff, and HROS-AD in our previous work[5]. However, wearable sensor data measurements are often affected by external factors or variables that are not captured by sensors like environmental conditions (high temperature and altitude etc.), and this may lead to inherently unpredictable time-series data[10]. Detecting anomalies in such scenarios become challenging using standard approaches based on statistical measures that rely on stationarity (is one whose properties do not depend on the time at which the series is observed) of the data or a pre-specified time-window to detect changes in the underlying distribution or prediction errors[10].

We propose a deep learning framework, Long Short Term Memory Networks (LSTM) based Autoencoder for Anomaly Detection (LAAD), that learns temporal dependencies from the input data (baseline or training data) to reconstruct 'normal' time-series output, and calculates a threshold based on the reconstruction error from the baseline and uses it to detect anomalies in the test data[10] (**Fig. 1**). LAAD contains an encoder that learns a vector representation of the input time-series data and a decoder that uses this representation to reconstruct the data. For the implementation of both the encoder and the decoder, we use LSTM cells. LSTMs can learn dependencies from one instance to another while solving the vanishing-gradient problem that affects standard recurrent neural networks[10]. This capability is due to a more complex cell architecture that accurately maintains a memory of important correlations. LAAD's ability to learn higher-level temporal patterns without prior knowledge of the pattern duration and its ability to learn from 'normal' data makes it robust to both predictable and unpredictable time-series data[10].

We collected wearable data from 107 individuals including 25 COVID-19, 11 non-COVID-19 illness with self-reported symptoms, and 71 healthy individuals with no symptoms or illness (**Table 1**) during the COVID-19 period to see if there is any anomalous signal during the COVID-19 infectious period (7 days before and 21 days after the symptom onset). Before applying the deep learning framework, we inferred resting heart rate from heart rate (RHR) and steps and then aggregated to 1-hour resolution (**Methods**). We divide the wearable data into a baseline or training data by taking the first 10 days of the data without overlapping 20 days before COVID-19 pre-symptomatic period and the test data by taking days from 20 days before the pre-symptomatic period and after (**Methods)**. We transform univariate RHR data into sub-sequences by combining time steps contiguous data values using data windowing method (**Methods**).

Deep learning models typically require large datasets for training. Otherwise the model would most likely overfit. In our case, we only have a limited number of days available for training. Augmented data can cover unexplored input space, prevent overfitting, and improve the generalization ability of a deep learning model[11–13]. To do this, we have applied different data augmentation techniques[12,13] on our limited training data by transforming the first 10 days of each user's data into around 80 days of augmented data (**Methods**).

Next, we built an LSTM-based autoencoder model, and this model takes input with a specific shape (batch size, sequence length, number of features) and returns an output of the same shape

(**Fig.1, Methods**). We split augmented training data into 95% train and 5% validation and use them to train the LAAD model using pre-selected tuned hyper-parameters (hidden layers = 6, number of neurons = 128, batch size = 64, number of epochs = 1200 with early stopping callbacks, learning rate = 0.0001, optimizer = adam) (**Methods, Supplementary Fig. 4**). Using a reconstruction loss, mean square error (MSE), we build a threshold and if the MSE for a test sample is greater than this threshold, we will label this sample as an anomaly (**Methods, Supplementary Fig. 5**). The higher the loss the worst the model will be where it fails to reconstruct the given input.

Using LAAD, in COVID-19 positive patients, we detected abnormal RHR signal in 21 individuals during pre-symptomatic (early) and 3 individuals during post-symptomatic periods (late) and failed to detect in one individual (missed) (**Fig. 2a, Supplementary Fig. 6, Supplementary Table 2**). For example, in a COVID-19 positive individual, ASFODQR, anomaly score representing abnormal RHR starts before symptom onset and increases during the COVID-19 infectious period compared to the baseline, non-infectious or recovery periods (**Fig. 2b**). On average abnormal RHR starts 5.26 days (median = 6.90) before the symptom onset, 6.20 days (median = 6.96) in pre-symptomatic and 1.36 days (median = 1.75) later in post-symptomatic cases (**Fig. 2d, Supplementary Table 2,5**). Out of 23 pre-symptomatic cases 12 had strong and 9 had weak abnormal RHR during pre-symptom period, and 3 post-symptomatic cases had strong abnormal RHR during post-symptomatic period (**Fig. 2a**, **Methods, Supplementary Table 5**). This indicates that probably not all COVID-19 positive individuals have similar levels of abnormal RHR, and also suggests that LAAD can successfully detect the cases where the signal is not too strong.

To calculate the performance of the model, we further divided test data into test-normal by taking the 10 days before COVID-19 pre-symptomatic period and test-anomaly by taking the days during pre-symptomatic and post-symptomatic periods (**Methods**). On average, the anomaly detection model had a precision of 0.81 (0.49 – 1), a recall of 0.59 (0.15 - 0.95), and a F-beta score of 0.79 (0.48 – 0.97) (**Supplementary Table 8**). Precision, P, was defined as the ratio between true positives, TP (the number of samples belonging to the anomaly class that are correctly classified) and the sum between TP and false positives, FP, where FP represents those normal samples that are incorrectly classified as anomalous, and recall, R was defined as the ratio between the TP and the sum between TPs and false negatives FN, which are the anomalous samples incorrectly classified as normal, and finally, the F-beta score is defined as the weighted mean of P and R (**Methods**). It allows a model to be evaluated taking both the precision and recall into account using a single score.

Next, we investigated 12 non-COVID-19 individuals and found abnormal RHR in 9 individuals during pre-symptomatic and failed to detect in 2 individuals (**Supplementary Fig. 1, Supplementary Table 3,6**). On average, the anomaly detection model had a precision of 0.64 (0.14 – 0.92), a recall of 0.63 (0.01 - 1), and an F-beta score of 0.63 (0.13 – 0.92) (**Supplementary Table 9**). We further investigated 71 healthy individuals by using randomly assigning a symptom date for test data. (**Methods**) and detected abnormal RHR signal in 53 individuals during pre-symptomatic and 16 individuals during post-symptomatic periods and failed to detect in 2 individuals (**Supplementary Fig. 2, Supplementary Table 4,7**). On average, the anomaly detection model had a precision of 0.71 (0.1 - 1), a sensitivity or recall of

0.41 (0.02 - 1), and a F-beta score of 0.70 (0.10 - 0.98), (**Supplementary Table 10**). Overall, LAAD had higher F-beta scores for COVID-19 cases than other groups (**Fig. 2c**).

Interestingly, COVID-19 cases have longer hours of abnormal RHR (180 hours or 7.5 days) during the infectious period compared to non-COVID-19 (133 hours, 5.5 days) and healthy (93 hours, 3.9 days) cases suggesting that COVID-19 infection lasts longer than other infections (**Fig. 2e, Supplementary Fig. 3, Supplementary Table 11, Methods**). In total, 88% (22 of 25 cases) of COVID-19 cases had more than 3 days of abnormal RHR signal compared to 81.9% (9 of 11 cases) of non-COVID19 and 57.14% (40 of 70 cases) of healthy cases (**Supplementary Fig. 3**). Abnormal RHR level during the infectious period was elevated by 4.33 days and lowered by 3.63 days compared to the baseline (**Fig. 2e, Supplementary Fig. 3, Supplementary Table 12, Methods**). Whereas, in non-COVID-19 cases and healthy it was elevated by 3.16 days, 3.96 days and lowered by 4.96 days and 3.69 days respectively (**Fig. 2f**) suggesting that levels of abnormal RHR can predict COVID-19 but cannot distinguish from non-COVID19 illness and healthy cases.

Our study has several limitations. First, all symptom onset dates were self-reported by the patients usually after confirmation of COVID-19. Since the data was divided into train and test data and metrics calculations were based on the self-reported symptom onset date, any errors in self-reporting can introduce bias in our results and model performance. Second, we divided the symptomatic period into pre-symptomatic and post-symptomatic periods using the consensus of recent studies on the viral infectiousness period. However, the length of these periods could vary from one patient to another and could introduce bias in our results. Third, none of the healthy patients had laboratory-confirmed COVID-19 tests and therefore it is possible that some of them could be asymptomatic in nature. Indeed, we found 14 cases where abnormal RHR was detected for 7.5 days (180 hours) more during the infectious period similar to COVID-19 patients (**Supplementary Table 7**). Fourth, on average only around 3 months of data was collected per user and deriving training, validation and test data from such limited data could be a limiting factor for the model performance. Fifth, we did not test any COVID19 negative cases and it limits the potential of our study. Sixth, only 25 COVID-19 cases were used in the analysis and adding more samples would help us better understand the wearable data functionality in detecting COVID-19. Seventh, all the data used in the study were collected from Fitbit smartwatch heart rate and activity data. Adding other data types like temperature, saturated oxygen, sleep, and data from different devices like Apple watch and Garmin watch could potentially improve COVID-19 detection. Eighth, choosing a correct baseline with enough data size to train the model has a major impact on predictions. For example, A7EM0B6 with only 4 days of training data has severe underfitting of the model (**Supplementary Fig. 4**). Ninth, sleep data was inconsistent and was not the same resolution as heart rate and activity data, and this limited the possibility of including sleep in our model. Finally, this deep learning framework uses retrospective data and is designed as a proof-of-concept. It is yet to be tested in a real-time fashion using pre-annotated labels from users.

Our work suggests that wearable sensor data could be used as a marker for early prediction of COVID-19. A detailed real-time wearable study on the COVID-19 patients with symptoms annotated by users, will help us understand more about tracking and modeling, and detecting outbreaks of SARS-CoV-2.

Methods

Datasets

We used publicly shared data from phase-1 study[5]. We selected 25 COVID-positive, 11 non-COVID-19 sickness, and 71 healthy individuals with enough data to train, test and evaluate metrics. It also contains metadata with self-reported symptom dates and diagnosis dates. We used a *sample* function from *pandas* library (https://pandas.pydata.org) to randomly assign a symptom date for each healthy individual using their test data time intervals with a fixed random seed. For two COVID positive datasets without symptom dates, we used diagnosis dates as symptom dates (**Table**).

LAAD framework

Pre-processing. For each user, we selected the heart rate data and merged the step data with the same exact time-stamp as heart rate. Merged data was aggregated to 1-minute resolution using a *resample* function from *pandas* library. Next, the resting heart rate was calculated as heart rate where steps are zero for 12 minutes ahead of a given time point. Next, we used moving-averages (mean = 400 hours) to smooth the time-series data with *rolling* function from *pandas*, and further aggregated to 1 hour by taking the mean.

Data splitting and normalization. We divide the wearable data into a baseline or training data by taking the days prior to COVID-19 pre-symptomatic period (days before -20 days from the symptom date) and the test data by taking days during pre-symptomatic and after (-20 days from the symptom date and later). Training data is further divided into training by taking the first 10 days (days that did not overlap test data) and split further into 95% training and 5% validation data after applying data augmentation (**Methods**). Train data and test data were standardized separately and transformed into a 2D array format before feeding them into the LAAD framework. To calculate performance of the model, we further divided test data in to test-normal by taking the days before COVID-19 pre-symptomatic period (days before -20 to -10) and test-anomaly by taking the days during pre- and post-symptomatic periods (-7 to + 21). Finally, data was standardized by removing the mean and scaling to unit variance with the help of a standard *scaler* from s*klearn* library (https://scikit-learn.org).

Data windowing. We split the RHR sequence and grouped using a fixed-length window (W) of size 8. The value of W defines how many time-lags are processed by the LAAD that classifies the input as an anomaly or not (**ref**). Further we reshape the data format to tensor with (number of samples, number of time steps per sample, number of features).

Data augmentation. We applied seven general time series data augmentation techniques found from literature. ***Scaling*** - changes the magnitude of the data in a window by multiplying by a scalar. The scalar was determined by a Gaussian distribution with mean 1 and standard deviation 0.1. ***Rotations*** - applies arbitrary rotations to the existing data. ***Permutation*** - randomly perturbs the temporal location of within-window events. To perturb the location of the data in a single window, we first slice the data into N same length segments, with N ranging from 1 to 5, and randomly permute the segments to create a new window. ***Magnitude-warping*** – changes the magnitude of each time series is multiplied by a curve created by a cubic spline with four knots at random magnitudes with a mean 1 and standard deviation 0.2. ***Time-warping*** - perturb the

temporal location using a random smooth warping curve generated by a cubic spline with four knots at random magnitudes with a mean 1 and standard deviation 0.2. ***Window-warping*** - selects a random window of 10% of the original data and either speeds it up by 2 or slows it down by 0:5. ***Window-slicing*** - a window of 90% of the original time series is chosen at random.

LSTM-based Autoencoder. We use a sequence-to-sequence autoencoder since our data consists of time-series sequences (**Fig. 1**). The objective is to reconstruct the RHR data using an encoded representation of the input time-series sequences. The autoencoder consists of an encoder and decoder[10]. An LSTM encoder learns a fixed-length vector representation of the input RHR time-series data and the LSTM decoder uses this representation to reconstruct the RHR time-series using the current hidden state and the value predicted from the previous time-step. For the implementation of both encoder and the decoder, we used LSTM layers, which consider temporal dependencies from one sequence to another. We train the autoencoder using the baseline data with normal RHR temporal sequences and reconstruct it with a low reconstruction error and use it to detect anomalies in test data.
For LAAD, we used 4 LSTM layers, one RepeatVector layer, one TimeDense layer and 128 hidden neurons for implementing both encoder and decoder. The reconstruction error is calculated as a MSE (Mean Squared Error) and the ADAM algorithm is used to optimize the learning process. We split the training dataset and use 5% as a validation set for evaluation and monitor validation loss and MSE. We set a value (using two standard deviations away from the mean) of reconstruction error from the baseline data as a threshold and annotate any value in the test data that is greater than this threshold as an anomaly. We used *Talos* (https://github.com/autonomio/talos) to evaluate the algorithm performance by measuring the reconstruction error with a different set of hyper-parameters and consider the best model that gave the lowest error. Early stopping callback was used to avoid overfitting by stopping the training at the right time. The anomaly identification is achieved at inference time using the test set.

Anomaly distance, signal strength
Anomaly distance was calculated by subtracting the date and time of the first anomalous event during COVID-19 infectious period with the symptom date. Anomaly signal strength was calculated by dividing the number of anomalous events in a pre-symptomatic window (7 days before the symptom onset) with the loss calculated by mean square error and further divided by the length of the window (7 days) and multiplied with 100. In post-symptomatic cases, 21 days after the symptom onset and 21 days window length have been used. Cases that have more than 6 anomaly signal strength scores were grouped as strong and the later as weak.

Number of abnormal RHR hours, delta abnormal RHR
For each user, a number of abnormal RHR hours were counted during COVID-19 infectious period. For each user, delta RHR was calculated by subtracting the RHR of the anomalies in test data (COVID-19 infectious period) from baseline/training data (first 10 days). Delta RHR further grouped into positive if the RHR is elevated and negative if the RHR lowered.

Performance evaluation
True positives (TP) are the number of anomalous hours that are correctly identified as anomalous. False positives (FP) are the number of normal hours that are incorrectly identified as

anomalous. True negatives (TN) are the number of normal hours that are correctly identified as normal. False negatives (FN) are the number of anomalous hours that are incorrectly identified as normal. We calculated the performance metrics as follows precision recall where precision is defined as the ratio between true positives and the sum between true positives and false positives (Precision = TP / (TP + FP) and recall (also known as sensitivity or true-positive rate) is defined as the ratio between the true positives and the sum between true positives and false negatives (Recall = TP / (TP + FN). Precision measures the proportion of anomalous hours that are relevant and recall measures how many hours are anomalous. Further we calculate F-beta score, a weighted mean of both precision and recall. The F-beta score is a generalization of the F-score that adds a configuration parameter called beta. A default beta value is 1.0, which is the same as the F-score. We used a beta value, such as 0.1, that gives more weight to precision and less to recall, assuming false positives are more important to minimize, but false negatives are still important (F-beta = $((1 + beta^2) \times Precision \times Recall) / (beta^2 \times Precision + Recall)$).

Visualization

All the plots were generated using the following libraries - Matplotlib https://matplotlib.org/; Seaborn https://seaborn.pydata.org/; ggplot https://ggplot2.tidyverse.org/.

Data availability

https://storage.googleapis.com/gbsc-gcp-project-ipop_public/COVID-19/COVID-19-Wearables.zip

Code availability

Code is based on TensorFlow (http://tensorflow.org/), and is open sourced for non-commercial purposes, and available at https://github.com/gireeshkbogu/LAAD.

References

1. Hodgson, S. H. *et al.* What defines an efficacious COVID-19 vaccine? A review of the challenges assessing the clinical efficacy of vaccines against SARS-CoV-2. *Lancet Infect. Dis.* (2020) doi:10.1016/S1473-3099(20)30773-8.
2. He, X. *et al.* Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat. Med.* **26**, 672–675 (2020).
3. Cevik, M. *et al.* SARS-CoV-2, SARS-CoV, and MERS-CoV viral load dynamics, duration of viral shedding, and infectiousness: a systematic review and meta-analysis. *The Lancet Microbe* (2020) doi:10.1016/S2666-5247(20)30172-5.
4. Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* **20**, 533–534 (2020).
5. Mishra, T. *et al.* Pre-symptomatic detection of COVID-19 from smartwatch data. *Nature Biomedical Engineering* **4**, 1208–1220 (2020).
6. Quer, G. *et al.* Wearable sensor data and self-reported symptoms for COVID-19 detection. *Nat. Med.* 1–5 (2020).
7. Natarajan, A., Su, H.-W. & Heneghan, C. Assessment of physiological signs associated with COVID-19 measured using wearable devices. *npj Digital Medicine* **3**, 156 (2020).
8. Shapiro, A. *et al.* Characterizing COVID-19 and Influenza Illnesses in the Real World via Person-Generated Health Data. *Patterns* 100188 (2020) doi:10.1016/j.patter.2020.100188.
9. Smarr, B. L. *et al.* Feasibility of continuous fever monitoring using wearable devices. *Sci. Rep.* **10**, 21640 (2020).
10. Malhotra, P. *et al.* LSTM-based Encoder-Decoder for Multi-sensor Anomaly Detection. *arXiv [cs.AI]* (2016).

11. Le Guennec, A., Malinowski, S. & Tavenard, R. Data augmentation for time series classification using convolutional neural networks. in (2016).
12. Um, T. T. *et al.* Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks. *Proceedings of the 19th ACM International Conference on Multimodal Interaction - ICMI 2017* (2017) doi:10.1145/3136755.3136817.
13. Iwana, B. K. & Uchida, S. Time Series Data Augmentation for Neural Networks by Time Warping with a Discriminative Teacher. *arXiv [cs.LG]* (2020).

Table 1 and Figures

Group	Number of individuals	Data size (days)	Train (days)	Test (days)	$Test_n$ (days)	$Test_a$ (days)
COV19-positive	25	82.6	26.6	56	10	25.4
Non-COVID-19	11	80.2	24.6	55.6	10	28
Healthy	71	69.6	28.8	40.8	10	20.6

Table 1. Data summary. In total, 107 datasets and 3 groups. Data was split between training and test ($Test_n$ – test-normal size, $Test_a$ – test-anomaly size) using symptom onset day as a reference (**Methods**) and the values shown were the average number of days per group.

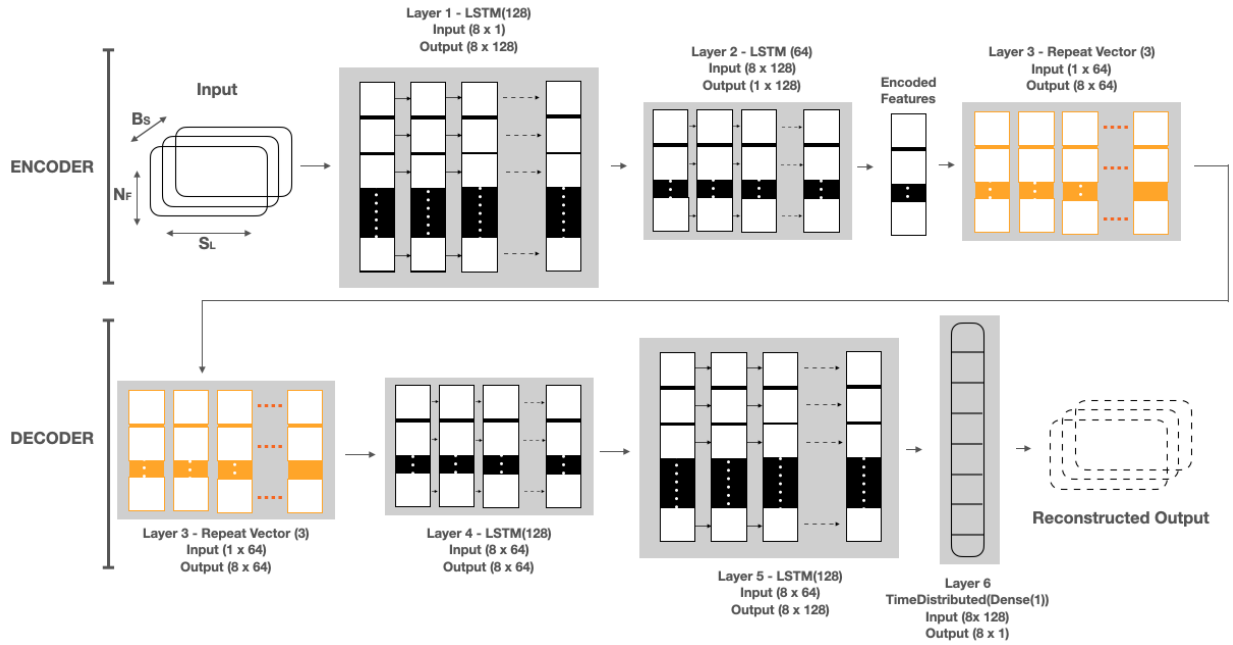


Figure 1. LAAD framework (LSTM Encoder-Decoder inference steps for input to reconstruct the output). LAAD takes standardized RHR data of shape (B_S, S_L, N_F), where B_S is batch size, N_F is number of features and S_L is sequence length or time steps and passes it to the first layer. The input data has 8 timesteps and one feature. First layer has as many LSTM cells as the S_L and makes each cell per timestep emit a signal to a second layer. Layer 1, LSTM(128), reads the input data and outputs 128 features with 8 timesteps. Second layer has half the size of LSTM cells than the previous and only the last cell emits an output. Layer 2, LSTM(64), takes the 8x128 input from Layer 1 and reduces the feature size to 64. The output of this layer is the encoded feature vector of the input data. Third layer uses a Repeat Vector that replicates the feature vector 3 times and gets a 2D array for the fourth layer (1st LSTM layer in Decoder) and acts as a bridge between encoder and decoder. The decoder layers unfold the encoding by stacking LSTM layers in the. reverse order of the encoder. Layer 4, LSTM (64), and Layer 5, LSTM (128), are the mirror images of Layer 2 and Layer 1, respectively. Layer 6, TimeDistributed (Dense(1)), is added in the end to get the reconstructed output, where "1" is the number of features in the input data.

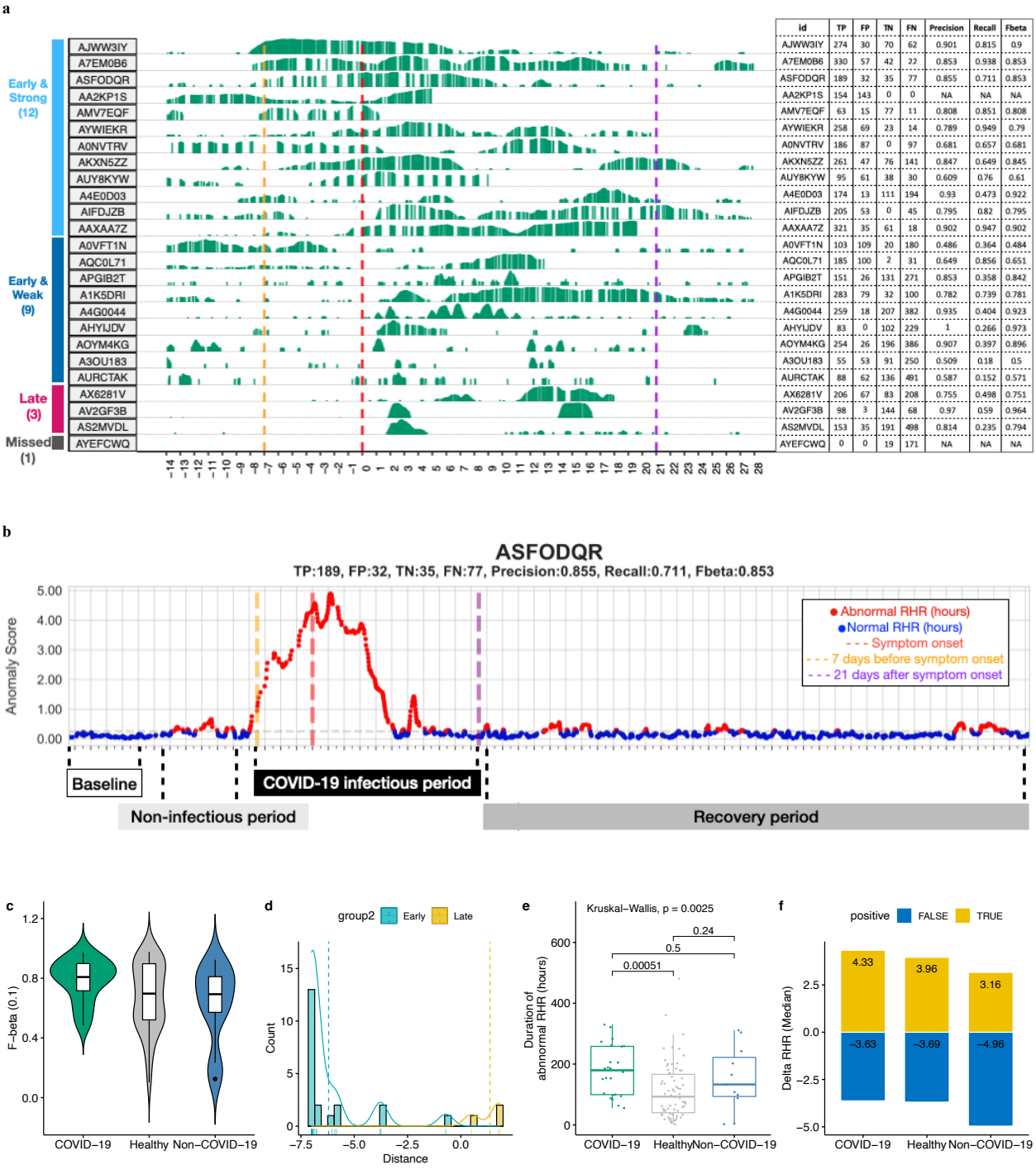


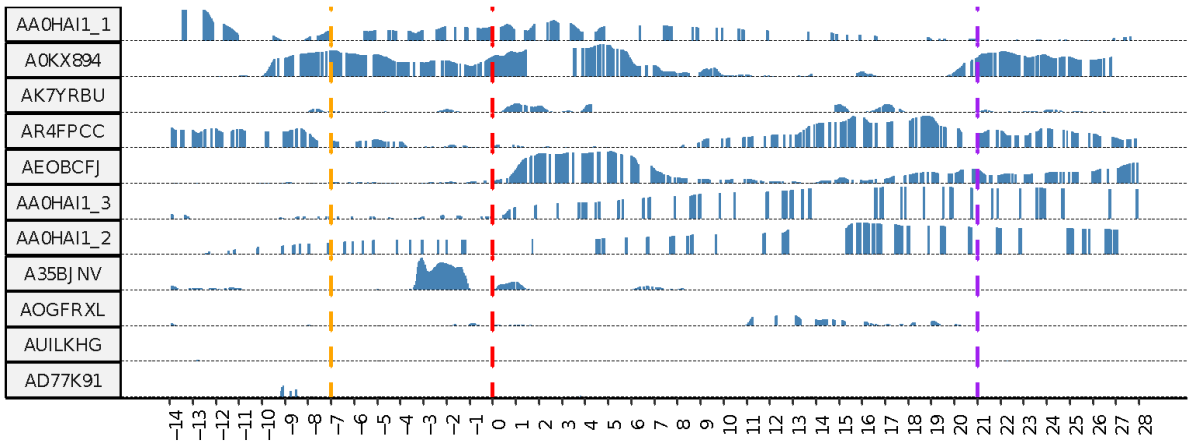
Figure 2. LAAD predictions, Evaluation metrics, Summary of detection timing, Comparison of abnormal RHR between COVID-19, healthy and non-COVID-19 groups. a, On the left, bar plots showing anomaly scores obtained from reconstruction error or mean square error (MSE) and timings of infection detection from LAAD model with respect to the different periods of SARS-CoV-2 infection for COVID-19 positive participants. Based on the time of detection, participants were grouped into early and late groups. Based on anomaly signal strength, these groups were further sub grouped into strong and weak. Participants were ordered

based on early and strong to weak and late grouping order. One participant, who LAAD failed to detect anomaly during infectious period was annotated as “missed”. The x axis shows days before, during and after the infection and the y axis shows loss values from LAAD (green bars). On the right, table showing performance metrics (TP - true positives, FP - false positives, TN – true negatives, FN - false negatives, precision, recall and F-beta score) for each participant. **b**, Scatter plot showing the distribution of anomaly score (MSE loss) of COVID-19 positive individual (ASFODQR) highlighting the baseline (first 10 days), non-infectious (10 days prior to infectious period, -20 to -10) , infectious (-7 to -21) and recovery periods (days after infectious period). Abnormal RHR was shown in red and normal RHR in blue color. Self-reported symptom onset date was shown as a red dotted line, a pre-symptomatic window as a gold dotted line and a post-symptomatic window as a purple dotted line. **c**, Density plot showing the distribution of detection timing during infectious period (gold = pre-symptomatic, grey = post-symptomatic) in COVID-19 participants. **d**, Bar plots showing the number of individuals who had more than 3 days of abnormal RHR signals during infectious period (green = COVID-19, grey = healthy, blue = non-COVID-19). **e**, Boxplots showing the distribution of the number of abnormal hours during infectious period in COVID-19, healthy and non-COVID-19 groups. P-values were shown on top of the boxplot. A Wilcoxon test was used to calculate the p-value between the groups. Kruskal-Wallis test was used to calculate global p-value. **f**, Bar plots showing delta RHR during infectious period in COVID-19, healthy and non-COVID-19 groups. Delta RHR of elevated RHR regions were shown in gold and lowered RHR regions shown in blue.

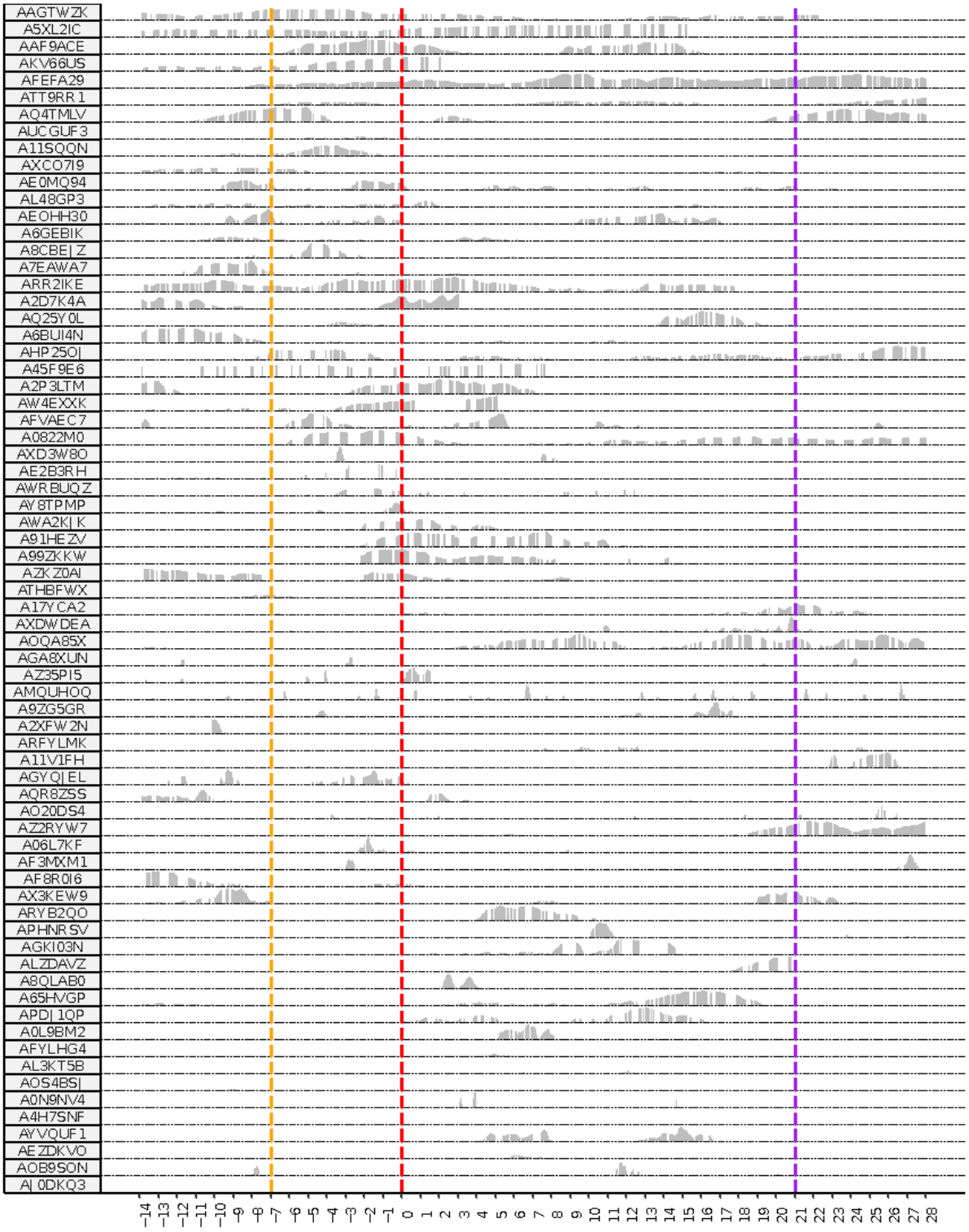
Supplemental Tables

Supplemental Table 1. Symptom and diagnosis dates of COVID-19, non-COVID-19 and randomly chosen symptom dates of healthy participants.
Supplemental Table 2. Anomalies predicted by LAAD in COVID-19 patients.
Supplemental Table 3. Anomalies predicted by LAAD in non-COVID-19 participants.
Supplemental Table 4. Anomalies predicted by LAAD in healthy participants.
Supplemental Table 5. COVID-19 patients grouped into early, late depending on time of detection and number of anomalies (signal), strong and weak groups depending on signal strength during infectious period.
Supplemental Table 6. Non-COVID-19 participants grouped into early, late depending on time of detection and number of anomalies (signal), strong and weak groups depending on signal strength during infectious period.
Supplemental Table 7. Healthy participants grouped into early, late depending on time of detection and number of anomalies (signal), strong and weak groups depending on signal strength during infectious periods.
Supplemental Table 8. Evaluation metrics (precision, recall, F-beta score) in COVID-19 patients.
Supplemental Table 9. Evaluation metrics (precision, recall, F-beta score) in non-COVID-19 participants.
Supplemental Table 10. Evaluation metrics (precision, recall, F-beta score) in healthy participants.
Supplemental Table 11. Duration of the abnormal RHR during infectious period in COVID-19 patients, non-COVID-19 and healthy participants. 14 users listed as possible asymptomatic based on more than 180 hours of abnormal RHR during infectious period.
Supplemental Table 12. Delta RHR (median difference between abnormal RHR from infectious period and RHR from baseline) in COVID-19 patients, non-COVID-19 and healthy participants. TRUE in the positive column indicated increased RHR and FALSE indicates lowered RHR.

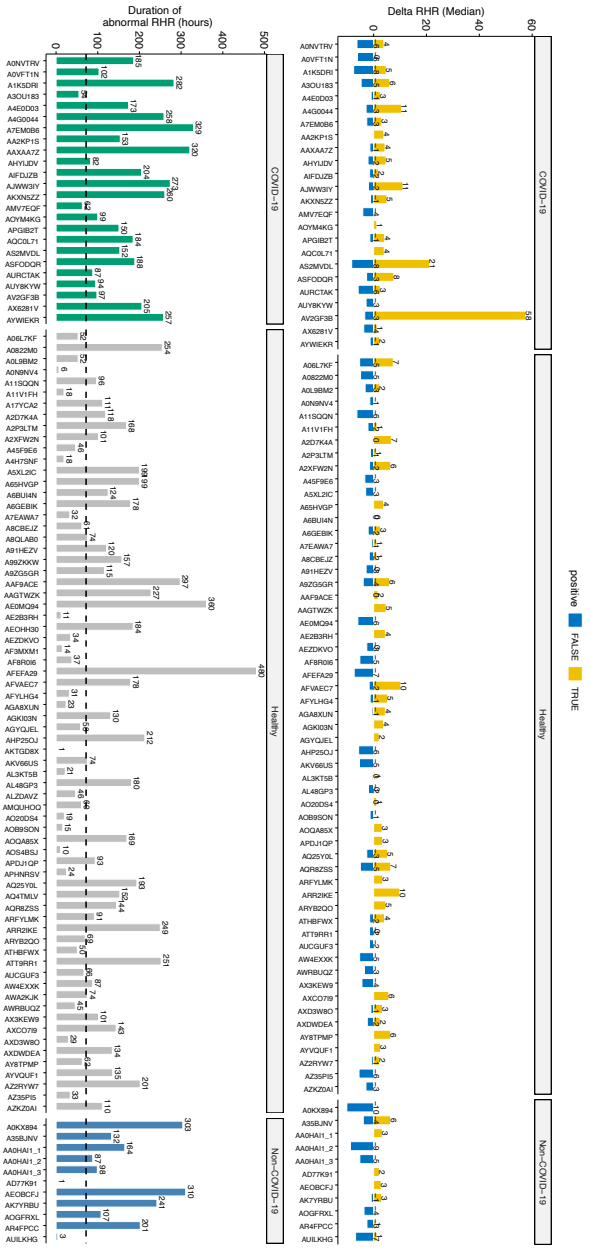
Supplemental Figures



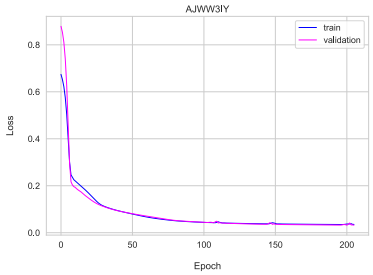
Supplemental Figure 1. LAAD predictions for non-COVID-19 group.

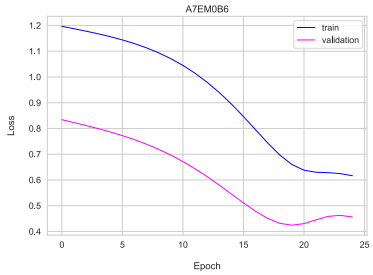


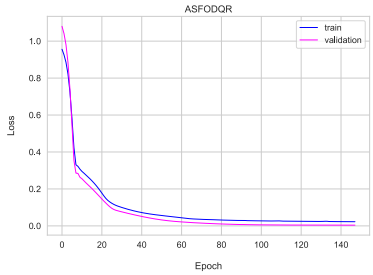
Supplemental Figure 2. LAAD predictions for a healthy group with randomly chosen symptom onset date.

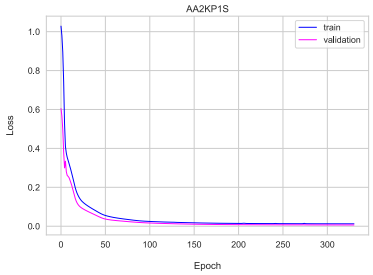


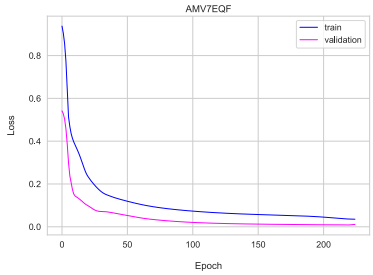
Supplemental Figure 3. Number of abnormal RHR hours (left) and median RHR differences (right) from the baseline in each user separated by COVID-19, non-COVID-19 and healthy groups. On the left, COVID-19 is shown in green, healthy in grey and non-COVID-19 in blue color. On the right, increased (positive direction) abnormal RHR shown in gold and decreased shown in blue (negative direction).

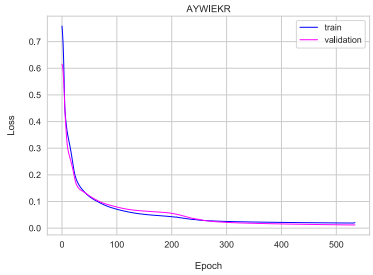

AJWW3IY
train
validation
Loss
Epoch

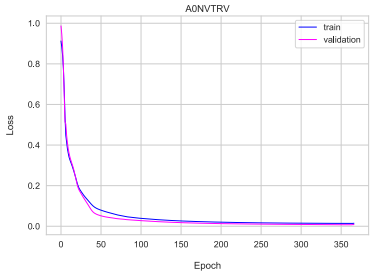

A7EM0B6
train
validation
Loss
Epoch

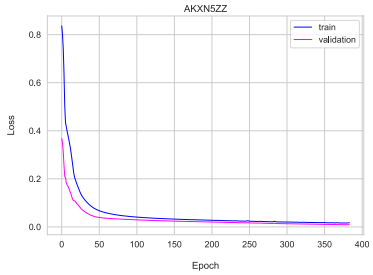

ASFODQR
train
validation
Loss
Epoch

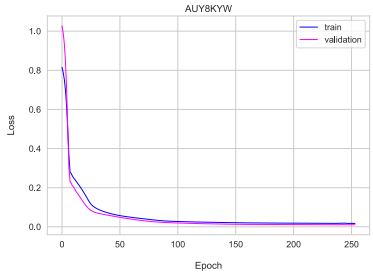

AA2KP1S
train
validation
Loss
Epoch

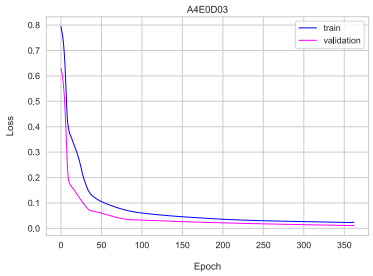

AMV7EQF
train
validation
Loss
Epoch

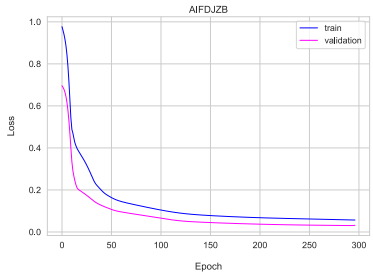

AYWIEKR
train
validation
Loss
Epoch

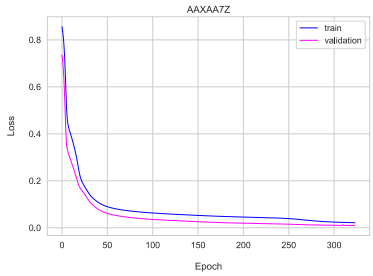

A0NVTRV
train
validation
Loss
Epoch

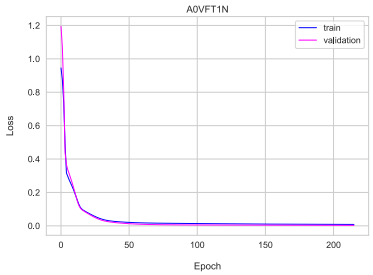

AKXN5ZZ
train
validation
Loss
Epoch

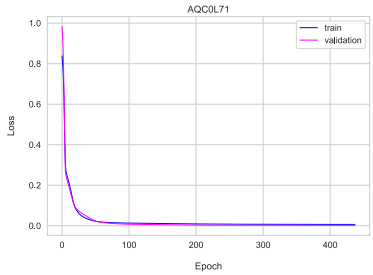

AUY8KYW
train
validation
Loss
Epoch

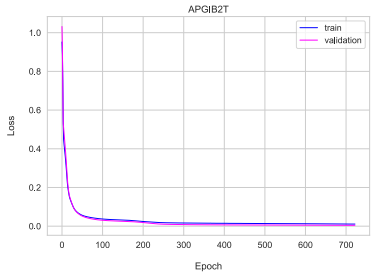

A4E0D03
train
validation
Loss
Epoch

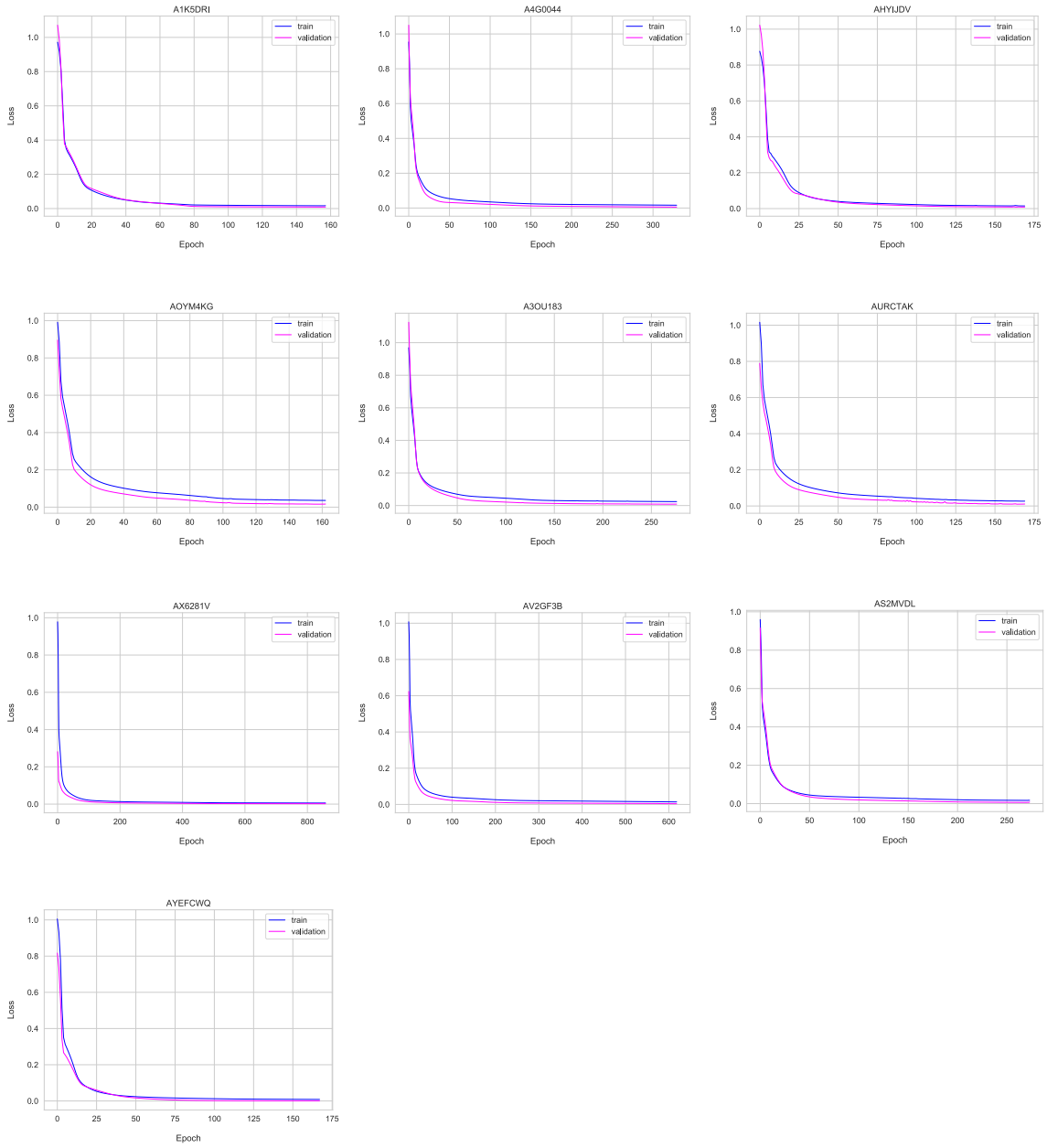

AIFDJZB
train
validation
Loss
Epoch


AAXAA7Z
train
validation
Loss
Epoch

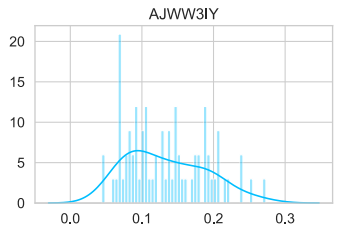

A0VFT1N
train
validation
Loss
Epoch

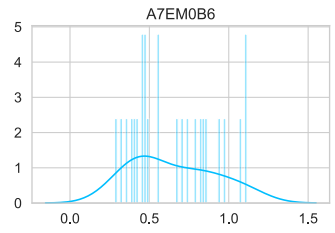

AQC0L71
train
validation
Loss
Epoch

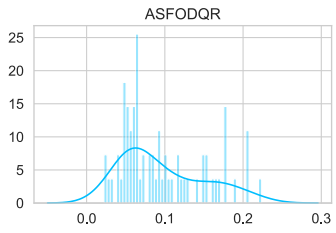

APGIB2T
train
validation
Loss
Epoch

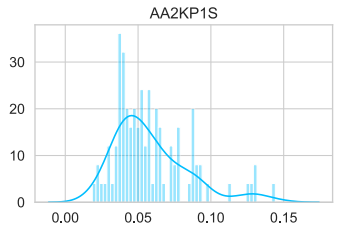


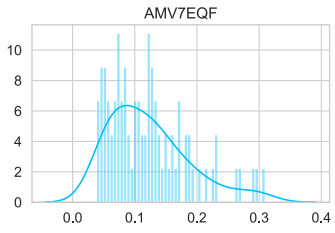
Supplemental Figure 4. Training (blue) and validation (purple) losses. First 10 days of training data were augmented into around 80 days of training dataset. Used 95% of this augmented data as training and 5% as validation and trained the model, and plotted reconstruction loss over several epochs using early call backs. A7EM0B6 had only 4 days of training data before augmentation and showed severe underfitting of the model.

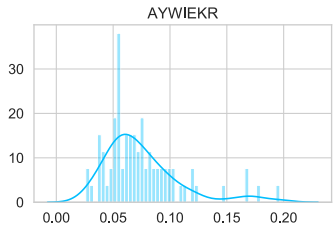

AJWW3IY

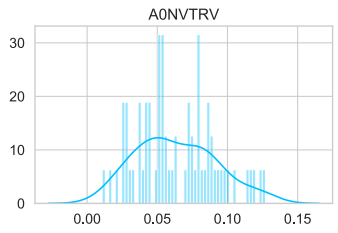

A7EM0B6

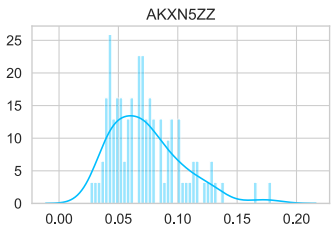

ASFODQR

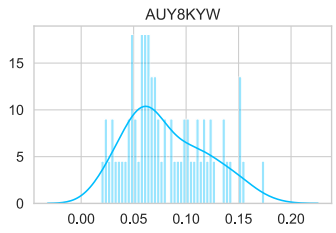

AA2KP1S

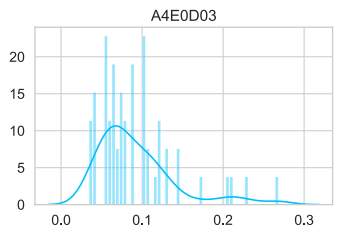

AMV7EQF

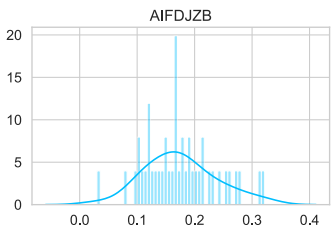

AYWIEKR

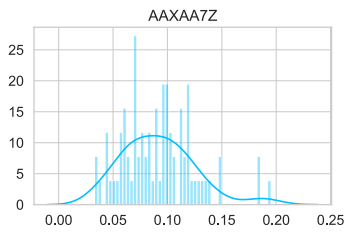

A0NVTRV

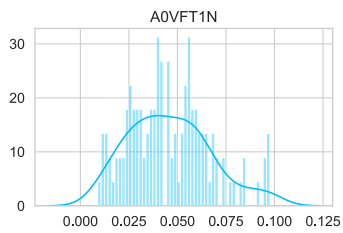

AKXN5ZZ

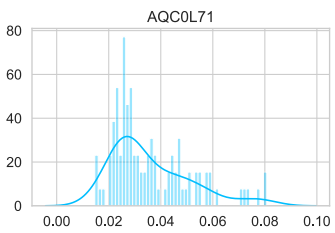

AUY8KYW

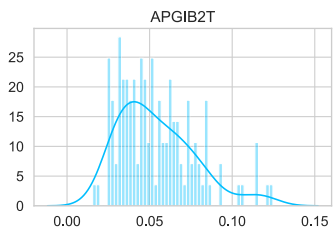

A4E0D03

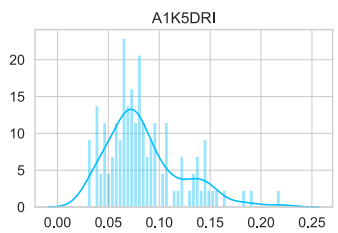

AIFDJZB

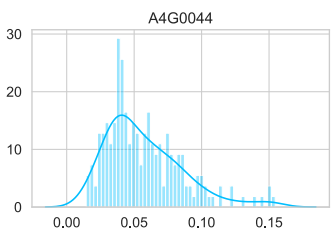

AAXAA7Z

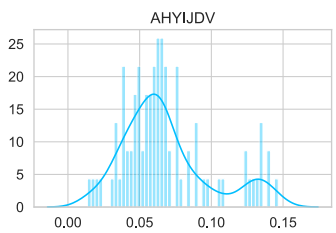

A0VFT1N

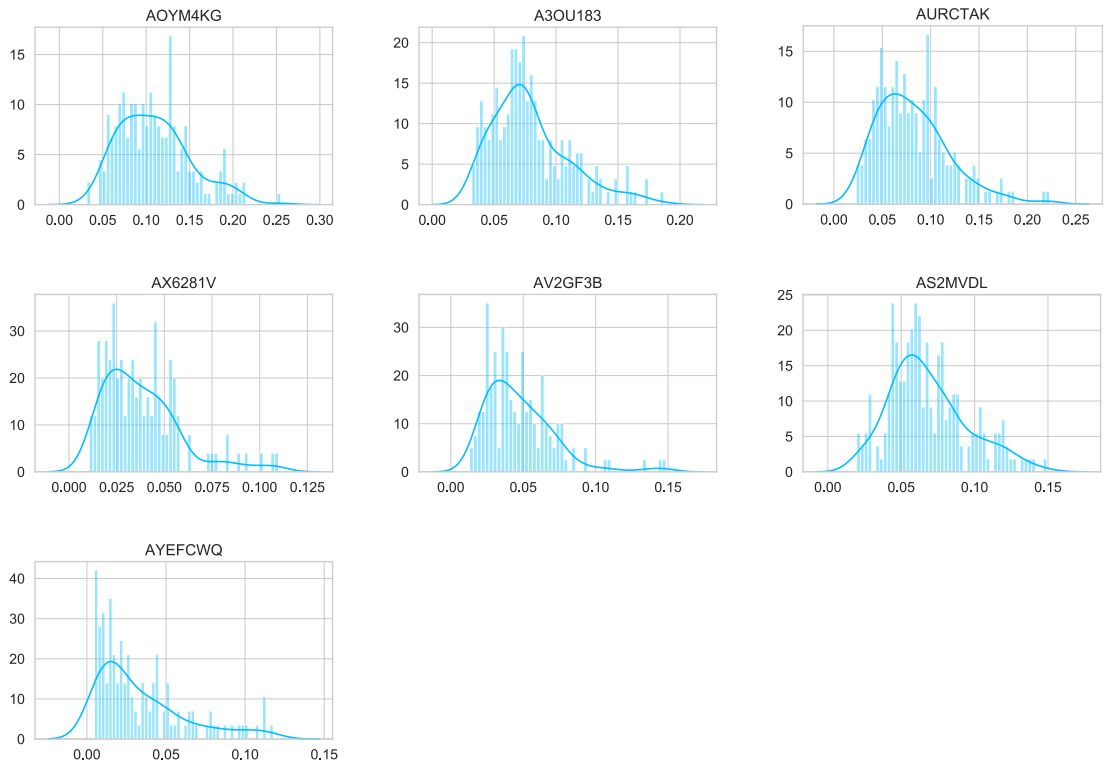

AQC0L71


APGIB2T


A1K5DRI


A4G0044

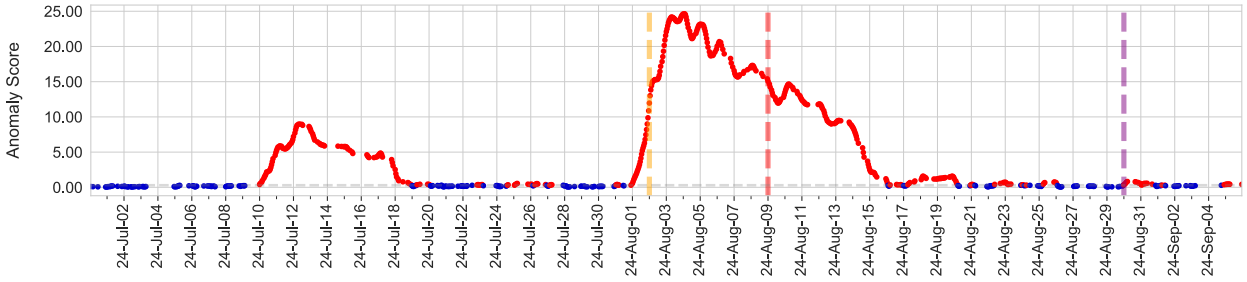

AHYIJDV



Supplemental Figure 5. Train loss (Reconstruction error or Mean Square Root Error) used to select anomaly threshold.

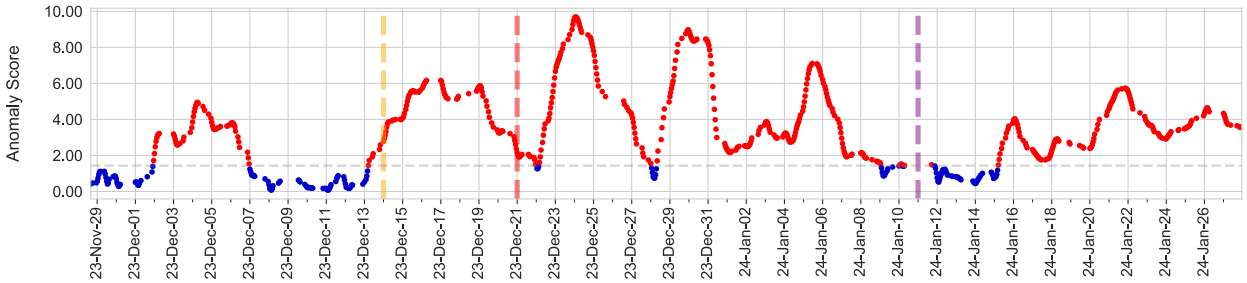
AJWW3IY

TP:274, FP:30, TN:70, FN:62, Precision:0.901, Recall:0.815, Fbeta:0.9



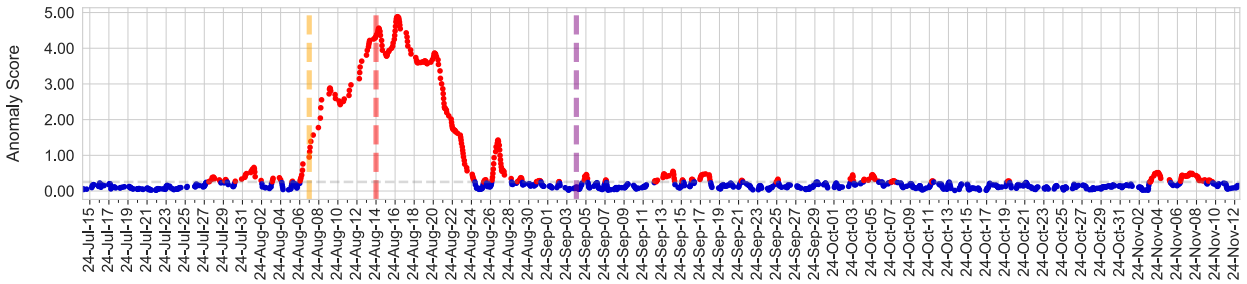
A7EM0B6

TP:330, FP:57, TN:42, FN:22, Precision:0.853, Recall:0.938, Fbeta:0.853



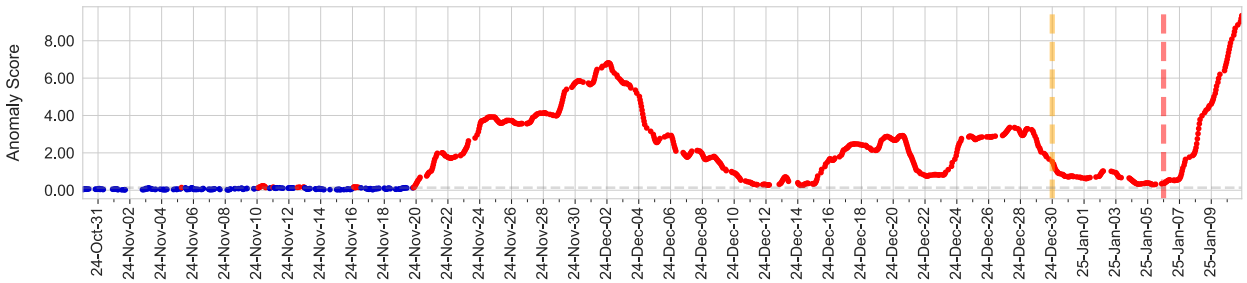
ASFODQR

TP:189, FP:32, TN:35, FN:77, Precision:0.855, Recall:0.711, Fbeta:0.853



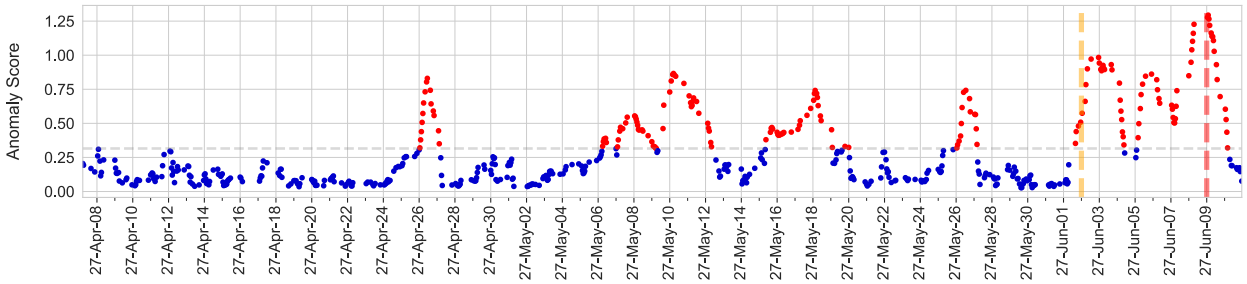
AA2KP1S

TP:154, FP:143, TN:0, FN:0



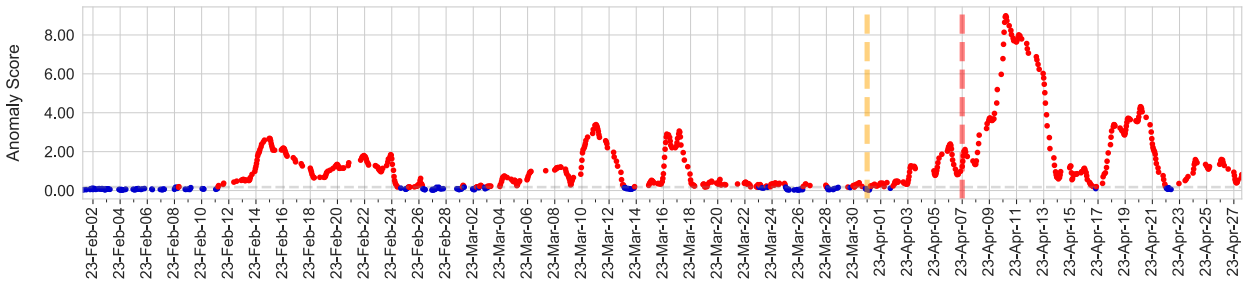
AMV7EQF

TP:63, FP:15, TN:77, FN:11, Precision:0.808, Recall:0.851, Fbeta:0.808



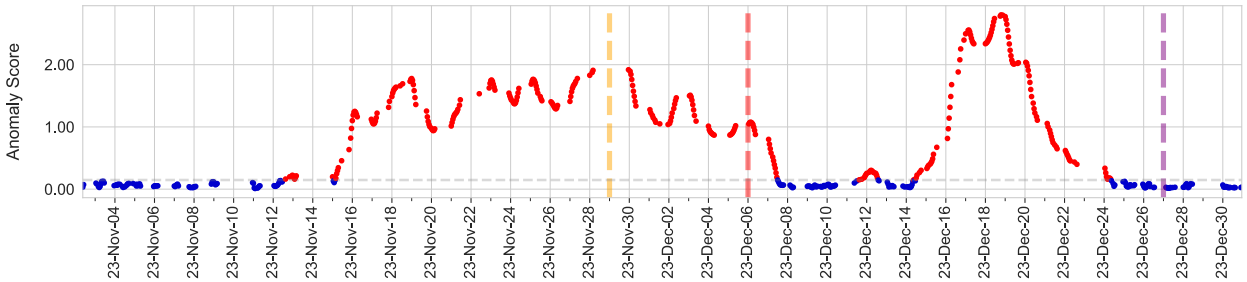
AYWIEKR

TP:258, FP:69, TN:23, FN:14, Precision:0.789, Recall:0.949, Fbeta:0.79



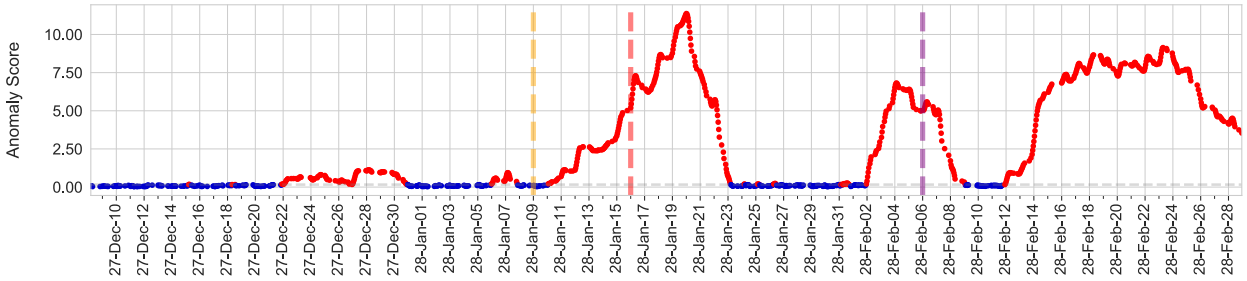
A0NVTRV

TP:186, FP:87, TN:0, FN:97, Precision:0.681, Recall:0.657, Fbeta:0.681



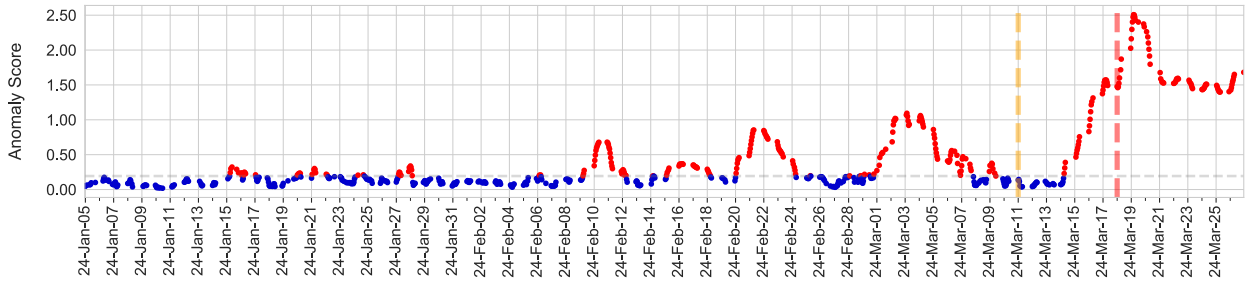
AKXN5ZZ

TP:261, FP:47, TN:76, FN:141, Precision:0.847, Recall:0.649, Fbeta:0.845



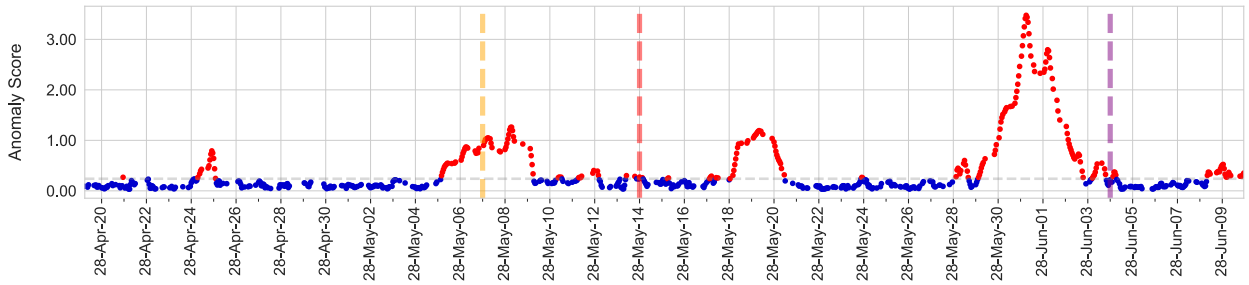
AUY8KYW

TP:95, FP:61, TN:38, FN:30, Precision:0.609, Recall:0.760, Fbeta:0.61



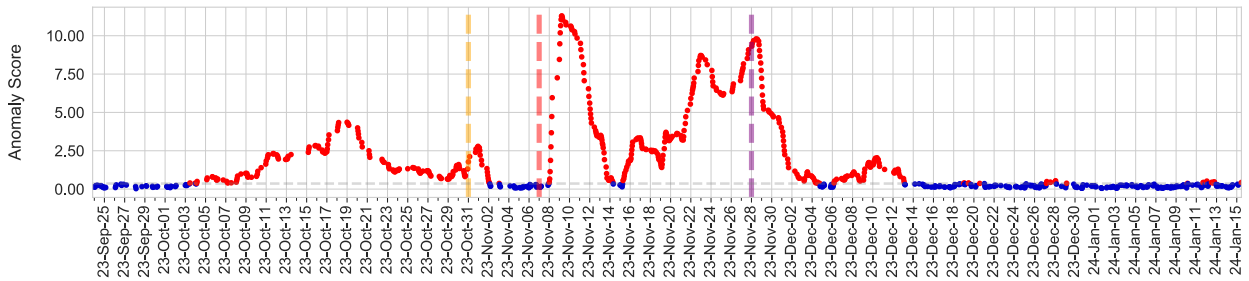
A4E0D03

TP:174, FP:13, TN:111, FN:194, Precision:0.930, Recall:0.473, Fbeta:0.922



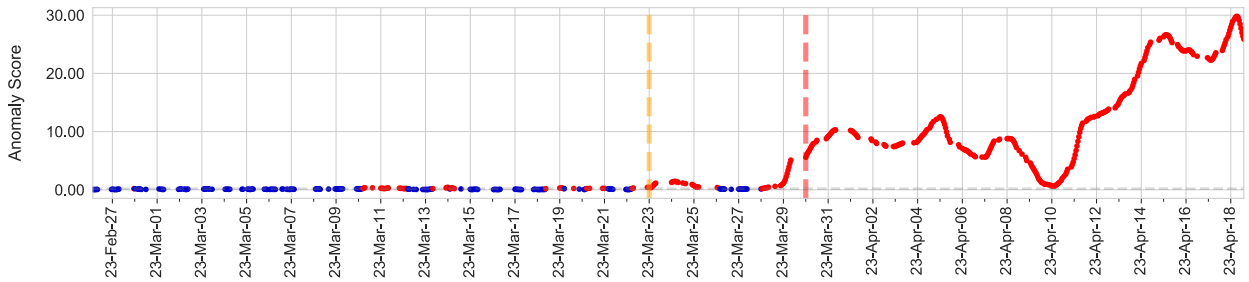
AIFDJZB

TP:205, FP:53, TN:0, FN:45, Precision:0.795, Recall:0.820, Fbeta:0.795



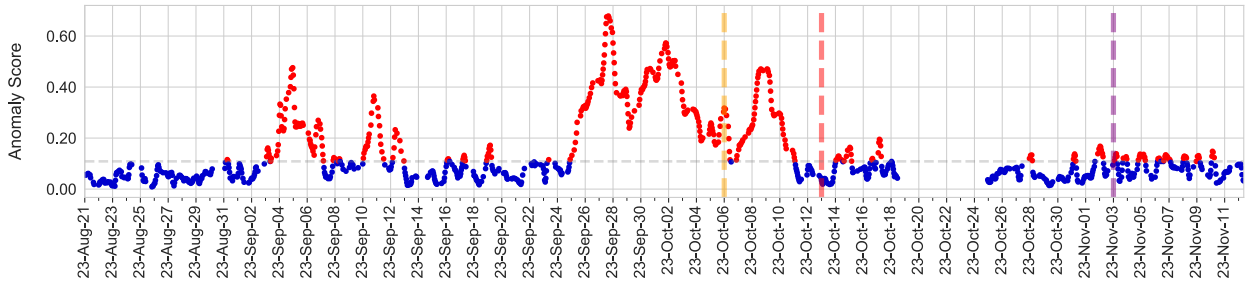
AAXAA7Z

TP:321, FP:35, TN:61, FN:18, Precision:0.902, Recall:0.947, Fbeta:0.902



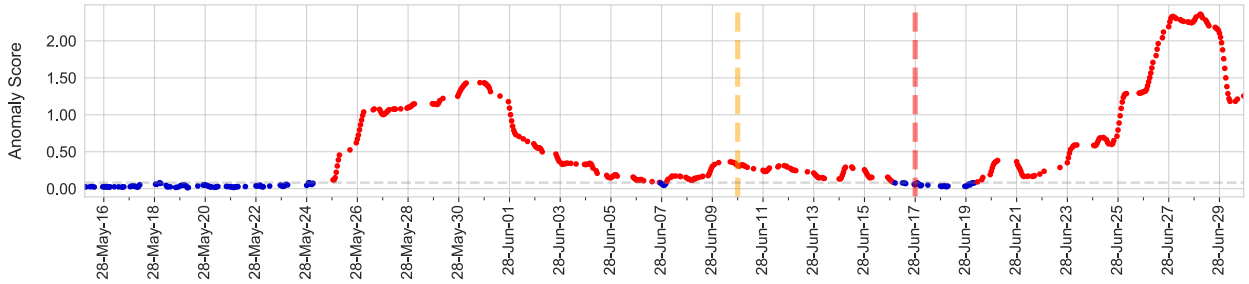
A0VFT1N

TP:103, FP:109, TN:20, FN:180, Precision:0.486, Recall:0.364, Fbeta:0.484



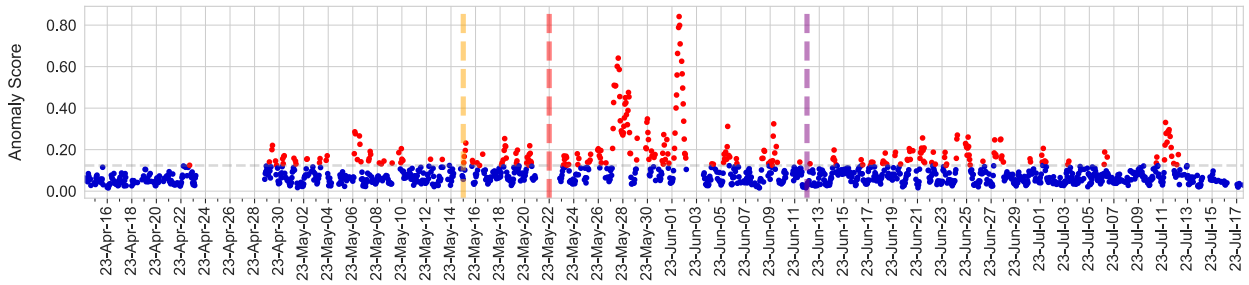
AQC0L71

TP:185, FP:100, TN:2, FN:31, Precision:0.649, Recall:0.856, Fbeta:0.651



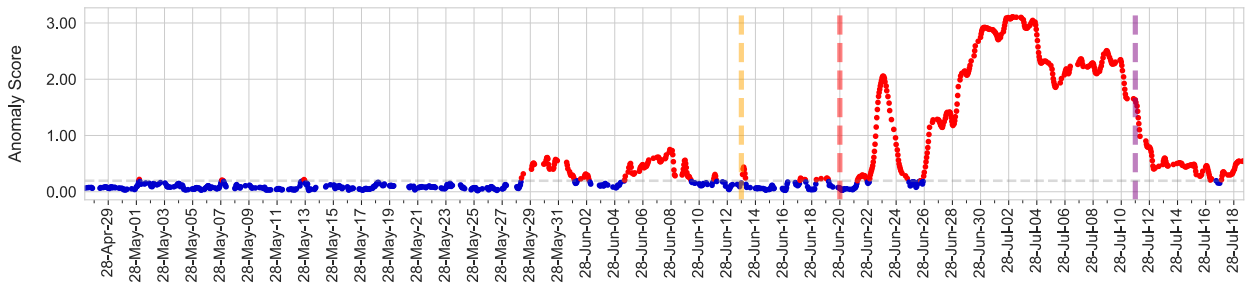
APGIB2T

TP:151, FP:26, TN:131, FN:271, Precision:0.853, Recall:0.358, Fbeta:0.842



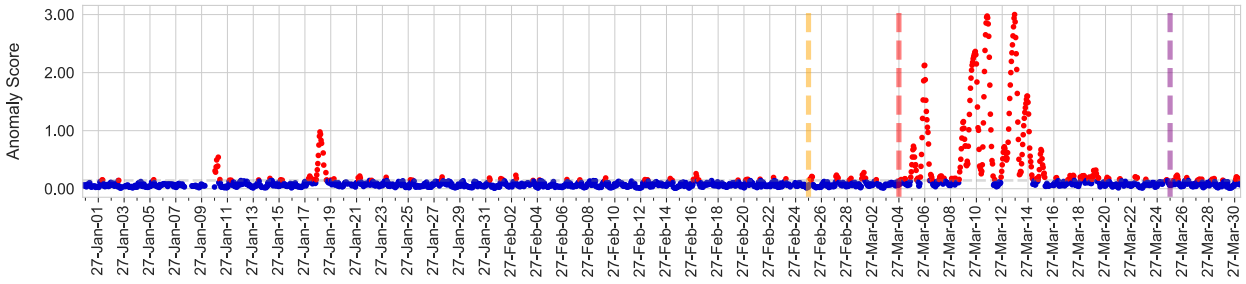
A1K5DRI

TP:283, FP:79, TN:32, FN:100, Precision:0.782, Recall:0.739, Fbeta:0.781



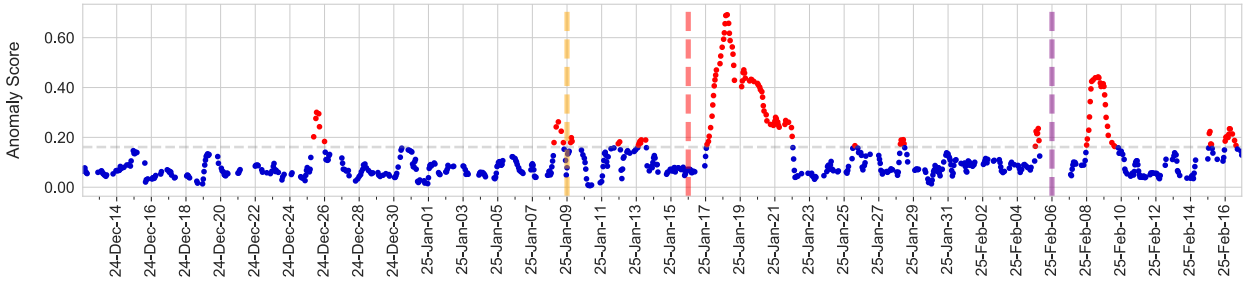
A4G0044

TP:259, FP:18, TN:207, FN:382, Precision:0.935, Recall:0.404, Fbeta:0.923



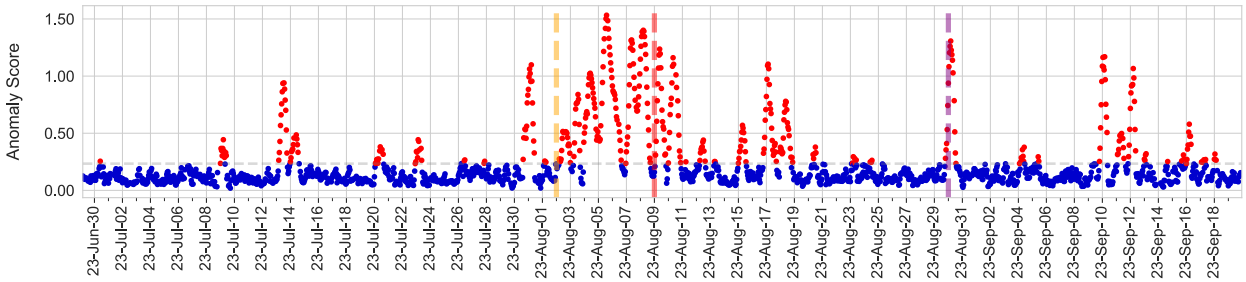
AHYIJDV

TP:83, FP:0, TN:102, FN:229, Precision:1.000, Recall:0.266, Fbeta:0.973



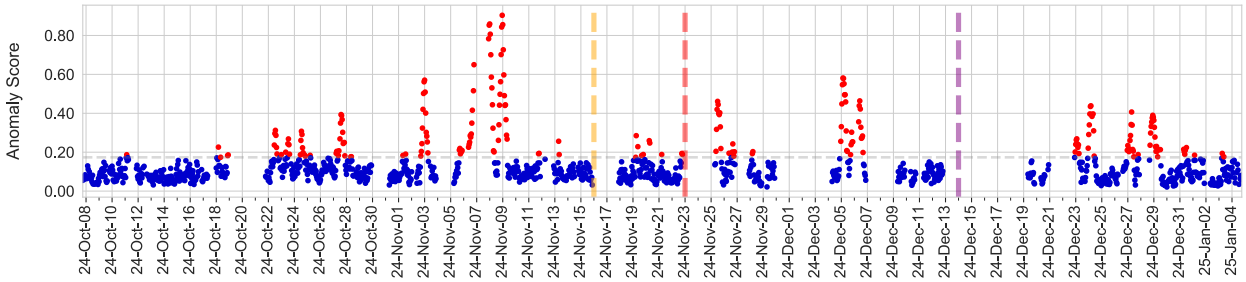
AOYM4KG

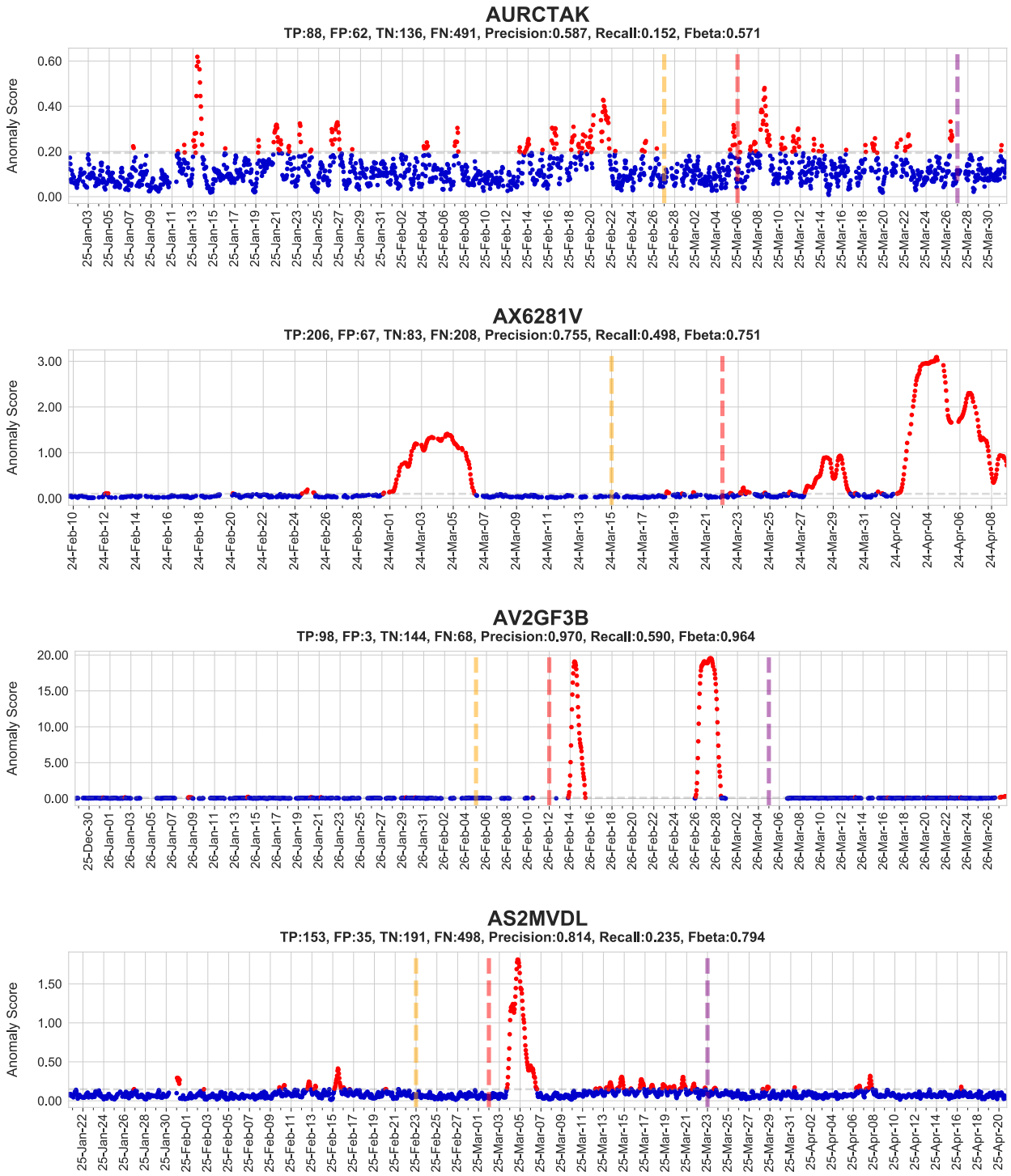
TP:254, FP:26, TN:196, FN:386, Precision:0.907, Recall:0.397, Fbeta:0.896



A3OU183

TP:55, FP:53, TN:91, FN:250, Precision:0.509, Recall:0.180, Fbeta:0.5





Supplemental Figure 6. LAAD predictions in COVID-19 patients (normal – blue, anomaly – red). Red dotted line is the symptom onset date, gold line represents pre-symptomatic window (7 days before symptom onset) and purple line represents post-symptomatic window (21 days after symptom onset).