

Challenges in Federated Learning Trained Anomaly Detection applied to Hospital Data without a Baseline

1st Susana Polido

Instituto Universitário de Lisboa (ISCTE-IUL)
Lisbon, Portugal
Susana_Polido@iscte-iul.pt

2nd Otavio Oliveira Napoli

Institute of Computing, UNICAMP
Institute of Computing, UNICAMP
São Paulo, Brazil
otavio.oliveira@ic.unicamp.br

3rd Maurício Breternitz Jr.

Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR
Lisbon, Portugal
Mauricio.Breternitz.Jr@iscte-iul.pt

4th Ana Maria de Almeida

Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR
Lisbon, Portugal
CISUC-Center for Informatics and Systems of the University of Coimbra
Coimbra, Portugal
ana.almeida@iscte-iul.pt

Abstract—During the COVID-19 pandemic, data collected via personal wearable devices was used to create models for the detection of a possible alteration of health status by defining an individual’s healthy baseline data. This work explores the usage of one of those models to enable a Federated Learning (FL) approach aiming to achieve a process applicable to sensing data from hospital-admitted patients. The fact that hospital data does not contain any samples that can confidently be considered “healthy” and thus serve as a baseline makes hospital COVID-19 detection a relevant challenge for anomaly detection techniques. After an adequate data preparation process, we were able to use the individually trained models to build an aggregated model for application to hospital data. Although the FL models obtain worse mean precision and recall scores when compared to the individual models, this experiment brings forth relevant knowledge on the compromises that might be necessary to develop a clinical anomaly detection model to be used in an Intensive Care Unit or monitored patients’ data lacking baseline samples.

Index Terms—Anomaly Detection, Federated Learning, Health data model, Machine Learning

This work was partially supported by Fundação para a Ciência e a Tecnologia, I.P. (FCT) with ISTAR Projects: UIDB/04466/2020 and UIDP/04466/2020 and DSAIPA/AI/0122/2020 Aim Health. The authors would also like to thank Hospital de Santa Maria (Portugal) for their collaboration and, in particular, the researchers José Miguel Dias and Luís Rosário, for their help in accessing and understanding this data.

I. INTRODUCTION

Early disease detection is key for successful infection treatment since it offers the best chance of cure ¹. Early detection of infectious diseases, besides lowering the cost of healthcare, plays a crucial role in all treatment and prevention strategies. It has been overly established that an incubation period follows the exposure event, during which the infected person may be asymptomatic and unaware of the infection. Moreover, early detection is especially relevant in an intensive care unit (ICU) context, where an infected individual at a contagious stage may have a serious impact on the cohort of patients admitted to the same ward.

In terms of machine learning (ML), several techniques have been employed for anomaly detection in different areas of application (such as bank fraud and malware detection among others). Nevertheless, detecting anomalies is based on the domain context and in signal signatures which tend to be unique to that area of application [1]. Anomaly detection (AD) using an individual autoencoder model was used to detect anomalies in infected individuals’ data [2], [3]. Data collected from personal wearable devices, like smartwatches, has shown that it is possible to early detect COVID-19 infections before the individual feels any symptoms of the illness, with models based on deviations from a predefined threshold using a baseline of the resting heart rate (RHR) [4]. The latter has

¹<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5395386/> (Accessed: 2023/09/25)

been further tested using an online real-time individual alert prototype implemented with a state-machine that signals its user to a possible infection of COVID-19 when the overnight RHR is at least 4bpm above the personal baseline for two nights in a row, managing to identify 80% of the infections [5].

As far as we know, no work relates the exploration of an anomaly detector for heart rate signals without establishing a previous baseline signal from which a threshold may be set. Bearing in mind that this detection can be viewed as a binary classification task with a severe imbalance problem where the positive class is the minority one and, thus, arising as a challenging problem as a classification task, the present study explores the possibility of application of one of the previously mentioned models, trained with data from multiple individuals, to new data acquired from different individuals, including data collected using hospital sensing devices in ICU patients. In the latter case, no samples can confidently be considered healthy and no individual baseline exists, preventing the retraining of the model to adjust it to the patient or the training of a new personal model.

On the other end, FL is an approach for using distributed ML where several clients locally train a model with their data and a central server combines the different models' weights into a single model via an aggregation function [6]. Building on the power brought by using varied data to build a robust model, the experiments next reported also include the exploration of training a model via Federated Learning (FL) to be used in situations where no previous healthy baseline is known, like in the context of the intensive care unit.

As a first step towards the development of an FL system for early detection of abnormal symptoms that may signal a possible health anomaly, this work uses the work of Bogu and Snyder [2] and their publicly available code as a starting point. The results obtained by the models trained via FL are compared to those obtained via the personalized models, i.e., models trained on an individual user's data for customized detection of anomalies, and models that are trained by joining the data of each one of the individuals into a unique training dataset (either locally or pooled).

Section II presents a brief description of the data and conducted experiments whose discussion occurs in Section III, ending with conclusions and directions for future work.

All the experiments here reported are performed with the ML library TensorFlow² and the simulation feature of the FL framework Flower³.

II. DATA AND EXPERIMENTATION SET UP

The data used comes from three different sources: data collected and explored by Mishra et al. [4] (named as P1), data collected and explored by Alavi et al. [5] (named as P2), and data collected using hospital sensing devices by the Hospital de Santa Maria in Lisbon, Portugal, during the hospital's normal activities, between the years 2019 and 2022 (hereinafter named as HSM data).

Both P1 and P2 datasets are publicly available [4], [5] and contain files with heart rate and activity data from volunteer participants collected via the volunteers' wearable devices. To extract the RHR, both signal datasets were processed according to the instructions found in [2]: heart rate data is resampled and averaged at one-minute frequency, and activity (that is number of steps taken) is resampled and summed at the same frequency (one-minute). The heart rate measurements used are the ones taken when there has been no known activity during the previous 12 minutes - termed RHR (resting heart rate) measurements. Data goes through a preprocessing phase that includes standardization based on the baseline data, using the StandardScaler method from the scikit-learn API⁴, and the application of a rolling window (RW) of 400-minute samples. The baseline samples also go through seven data augmentation techniques (scaling, rotation, permutation, magnitude-warping, time-warping, window-warping and window-slicing) before the training. Next, the first known data for a COVID-19 positive test of each of the individuals provided in the metadata is used to split the individual's data into a *baseline* period (ranging from 30 to 20 days before the first positive test), a *normal* period (from 20 to 10 days before the first test), and an *anomaly* period ranging from 7 days before the test until 21 days after (Fig. 1).

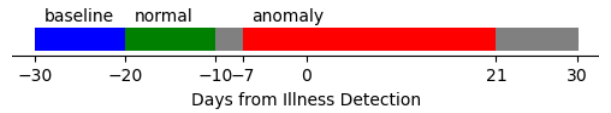


Fig. 1. Labeling of data from individuals with COVID-19. Days marked with grey are inconclusive and are not considered in this study.

Both the baseline and normal labels refer to samples that are considered healthy, with the baseline samples being used to train the models and calculate a threshold for anomaly detection, while the normal ones are used later to test the model [2].

Hospital data is composed of observations from 212 patients. Unlike the data from P1 and P2, HSM hospital data is collected via clinical monitoring devices, being acquired when patients are admitted into the hospital's ICU. Thus, data do not contain any samples that can be considered as a "healthy" baseline, meaning they cannot be used to (re)train the models nor generate a threshold for anomaly detection. Therefore, samples ranging from the patient's earliest date of known COVID-19 infection until 30 days after will be used for the anomaly detection test, that is HSM data used to understand if the models trained with P1 and P2 data are capable of detecting HSM signals as anomalous. The COVID-19 infection date has been determined via inspection of medical reports and observations, lab data tests, or the start of admission reports expressing the date of the first positive COVID-19 test, which was marked as day zero. HSM heart rate data

²<https://www.tensorflow.org/>

³<https://flower.dev>

⁴<https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>

can be considered RHR signals as described by Bogu and Snyder [2] since it is reasonable to assume patients from the ICU are resting in beds, chairs or stretchers. The time between samples varies from a couple of seconds to several minutes. Thus, according to the previously described preprocessing of RHR into one-minute interval samples, the HSM data were resampled at the same frequency and the exact same preprocessing steps were used, with the obvious exception of the labelling the signal samples as baseline/normal/anomaly.

As previously mentioned, the model used by Bogu and Snyder [2] is the only one here explored [2]. The authors present an AD pipeline featuring an autoencoder model that, at its core, consists of two consecutive LSTM layers with input and output shapes of $(8 \times 1) \rightarrow (8 \times 128)$ and $(8 \times 128) \rightarrow (1 \times 64)$, respectively, to extract 64 features from the original input which are replicated by a Repeat Vector layer before going through mirror versions of the first two layers with a TimeDistributed (Dense(1)) that reconstructs the shape of the input as its sixth and final layer. The autoencoder receives eight hourly RHR measurements, each one obtained using the mean of the previous 60-minute sample. Those eight hours of measurements do not necessarily correspond to consecutive hours as long as they are considered measurements in a resting state.

The Mean Absolute Error (MAE) of the RHR signal recreated by the model is compared to the maximum MAE from the baseline samples that acts as the threshold value to decide if the eighth hour of the input is anomalous: if the MSE is higher than the threshold it is considered an anomaly. A scheme of the anomaly detection process can be seen in Fig. 2



Fig. 2. Anomaly Detection Pipeline

The data of each individual is used to train an individual model and calculate a personal threshold to classify their

data or to train an overall model, according to the following described scenarios. Between the two public datasets, a total of 106 individuals are utilized to train models: 29 from P1 and 77 from P2.

To create the models, three different basic methods of training will be used and next compared: (a) Solo training - each individual's data creates its model and threshold; (b) Local training - joining the baseline data of all the different individuals to train a single model; and (c) FL training for 30 rounds, where each individual counts as a different client. These scenarios explore the possibility of having each create their threshold for detection versus the creation of an average threshold (Avg), necessary for scenarios (b) and (c). The differences between these methods can be seen in Figures 3 through 5.



Fig. 3. Training Methods - Solo Training

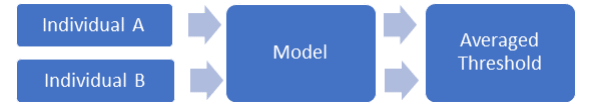


Fig. 4. Training Methods - Local Training with Averaged Threshold (Local Avg)

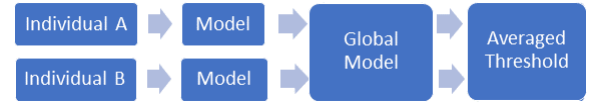


Fig. 5. Training Methods - FL Training with Averaged Threshold (FL Avg)

Finally, as previously mentioned, for the Local and FL models we have decided to create average thresholds (named Local Avg and FL Avg) from the averaged baseline signal of the individuals in the training dataset. These average thresholds are applied to hospital data for testing its expressiveness in terms of usage of AD pre-trained models on new unseen data.

Notice that, due to the lack of baseline in the hospital data, our work also explores the effect of the removal of the standardization step and the decrease of the RW dimension in these data. This is because 20% of the hospital patients don't possess enough samples for the application of an RW (Table I).

III. RESULTS

This section is divided into 3 subsections relating to changes in the preprocessing steps, threshold setting, and the method used for training (Solo, Local or FL), ending with the application of the Local and FL models with an averaged threshold (Local Avg or FL Avg) on the hospital dataset. All the scores next presented are calculated by considering

an hour measurement as an individual observation. The first two subsections report experiments performed using the P1 or P2 trained models separately.

A. Standardization, Learning Rate and Rolling Window

In this subsection, we present the results for different experiments performed using the publicly available models already trained [2] and training the models using either P1 or P2 datasets in isolation. The only training methodology employed in this experiment was *Solo training*, but experimented with a combination of different preprocessing steps. Starting with the application of the publicly available models trained with P1 data, we then trained new models under different data preparation and selection techniques, namely: without using data augmentation, avoiding standardization of the data, increasing the LR from 0.0001 to 0.001, halving the RW (from 400 and 200 samples), and removing the usage of an RW. Bear in mind that the described changes are cumulative. Table I contains a summary of the mean precision and recall scores obtained with the several solo training scenarios.

TABLE I

MEAN PRECISION AND RECALL SCORES OBTAINED USING SOLO TRAINING (ONE PERSONALISED MODEL FOR EACH OBSERVATION), DIFFERENT DATA PREPROCESSING CONFIGURATIONS, AND CHANGING THE LEARNING RATE (LR) PARAMETERS (THE CHANGES ARE CUMULATIVE). THE HIGHER AND LOWER SCORES OBTAINED WITH EACH DATASET PER METRIC ARE ALWAYS MARKED BY + AND -, RESPECTIVELY.

Experiment	P1		P2	
	Precision	Recall	Precision	Recall
Public Model	0.842	0.118(-)		
No augmentation	0.797	0.222	0.730	0.296(+)
No standardization	0.741(-)	0.200	0.778(+)	0.204
Increase LR	0.846	0.224(+)	0.771	0.282
Decrease RW	0.789	0.180	0.734	0.216
Remove RW	0.888(+)	0.123	0.711(-)	0.066(-)

Regarding model's performance with the P1 dataset test, we find that removing both the augmentation step and the standardisation process decreases the precision mean scores in the P1 dataset, as expected. However, this effect can be minimised by increasing the LR parameter which, although achieving the second-best precision value, manages to achieve the best recall. Notably, the highest precision value is achieved when we (cumulatively) remove the use of the RW. However, this action makes the values of recall decrease to the second lowest value.

Note that our experiments over the P1 dataset replicate the tests of Bogu and Snyder but achieve lower scores than those reported by the authors [2]. This is possibly due to the fact that they split the individuals' data based on the date of symptoms onset, while we chose to use the date of the positive COVID-19 test because we intend to study the scenario of infection bearing no evident physical symptoms.

The same experiments on the P2 dataset achieved generally lower precision scores when compared to those obtained over P1, with the second experiment - the one that avoids

augmentation of the training dataset - showing the highest precision. However, the recall suffers a slight decrease along the different experiences. Interestingly, the last experiment now performed the worst of all in both metrics.

Hereinafter, results are obtained using an RW of 200, no standardization and a LR of 0.001.

B. Threshold

For the next set of experiments, four different methods for calculating the threshold are used to determine if a sample is anomalous: the maximum MSA (Max) or using $\bar{X} + k\sigma$ MSA, with $k = 1, 2, 3$. In general, the effect on the metrics presents inverse directions and affects more the recall than the precision values, with the highest change being 0.075 for precision and 0.240 for recall. The mean scores of this variation can be seen in Table II.

TABLE II

MEAN SCORES OBTAINED WITH SOLO TRAINING MODELS, USING AN RW OF 200, NO STANDARDIZATION, AND AN LR OF 0.001, AND VARYING THE THRESHOLD USED TO DETERMINE IF A SAMPLE IS AN ANOMALY.

Threshold	P1		P2	
	Precision	Recall	Precision	Recall
$\bar{X} + \sigma$	0.748(-)	0.396(+)	0.744	0.390(+)
$\bar{X} + 2\sigma$	0.807	0.249	0.757	0.250
$\bar{X} + 3\sigma$	0.823(+)	0.156(-)	0.760(+)	0.165(-)
Max	0.789	0.180	0.734(-)	0.216

In what concerns experiments with the P2 dataset, replacing the Max by one of the other thresholds in terms of precision results in higher values than the second-worst found in Table I. Generally, the recall also presents higher values, especially for $\bar{X} + k\sigma$, where we achieve the best recall for both datasets considering all the experiments.

For the goal of this work (abnormal symptomatology alert for early disease detection), we gave preference to decreasing the number of anomalous samples that are not detected, i.e., to increase recall. Thus, the remaining experiments are conducted setting the anomaly detection threshold using the $\bar{X} + \sigma$ of the MAE, since the results presented in Table II show that, although slightly decreasing the precision (that still is close to 0.75), this threshold improved the recall the most.

C. Training Methodology

Next, we compare three different training methodologies: Solo, Local and FL. Remember that, *Solo training* uses the baseline data of each individual to train a personal model and define an individual threshold, while *Local* is when the baseline data of all individuals is joined and used to create a single model that uses an averaged threshold for detection in several individuals. Finally, *FL training* is when each individual trains a personal model and the different models are integrated into a single model via an aggregation function and using an averaged threshold.

All models in this section use an LR of 0.001 and are trained with data that underwent an RW of 200 without using data

standardization. The threshold used is either $\bar{X} + \sigma$ or the average of this threshold for every individual observation in the train.

Regarding the FL aggregation function, we use the default FedAvg function [6], which averages the weights of the different models received to create a single global model. We have chosen this function as the starting point of this exploration due to its simplicity and lack of parameterization. The individual training and merging are repeated several times - rounds - with the global model from the previous round serving as the starting point for the next models' training [6]. We set this cyclic process to run for 30 rounds.

For the Local and FL models, we performed two separate experiments: (i) having each observation create its threshold, which is used not only in Solo but also with the Local and FL-trained models (identified by using only Local and FL, respectively) in the following figures; and (ii) having all individuals generate a single averaged threshold (next identified as Local Avg and FL Avg). Figures 6 and 7 show the results of these experiments.

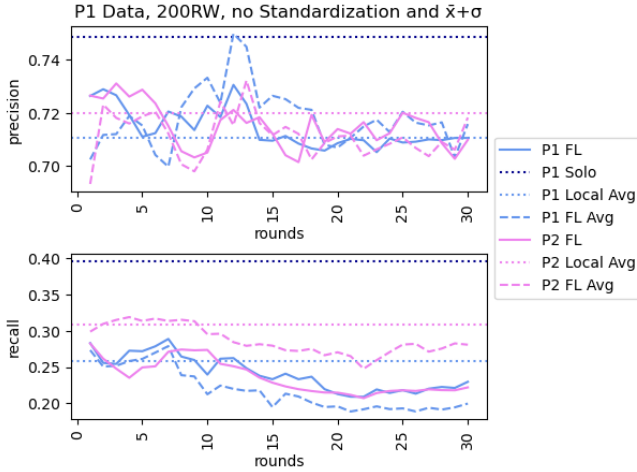


Fig. 6. Scores of the differently trained models when applied to P1 dataset.

As expected, the Solo models show the highest mean scores for both metrics. Notice that the averaged Local (Local Avg) model trained using each one of the P1 and P2 datasets behaves as the best local model for the other test set. In terms of the FL models, the precision of the models evolved in a more erratic way than the recall, but within a smaller range of values. This approach, for both FL and FL Avg experiments, was able to outperform local models at several rounds and for both metrics; and P1 FL Avg models were able to surpass even the Solo models' precision in some rounds. Regarding both test datasets, the model P1 FL Avg shows the worst recall while showing the best precision.

Examining how the mean threshold and loss evolve through the multiple training rounds, it is clear that the model is getting better at reproducing the input, but this does not translate into a better detection of anomalies since there is but a small

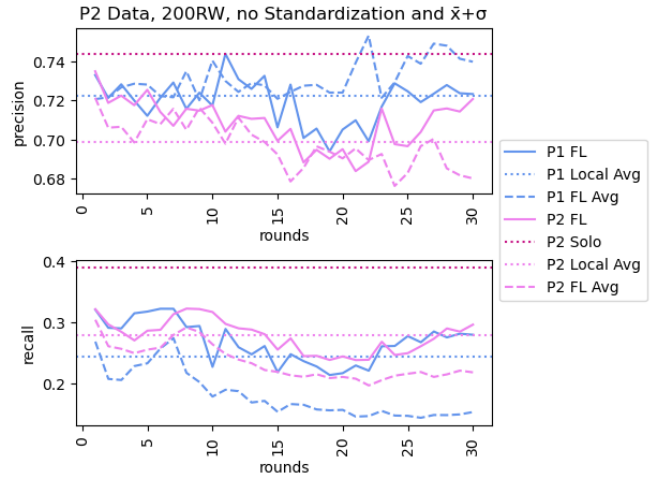


Fig. 7. Scores of the differently trained models when applied to P2 dataset.

difference between the loss obtained by the three types of samples in either dataset (Figure 8).

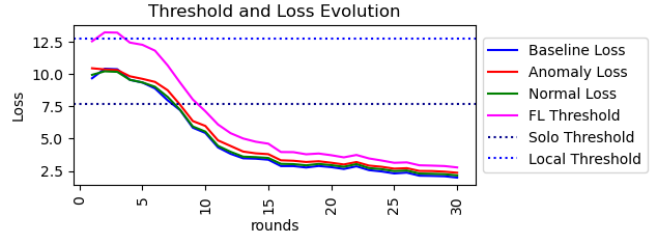


Fig. 8. Evolution of the mean threshold and mean loss obtained by each type of sample found in P1, P2 shows similar values and behaviour.

D. Applying Models to Hospital Data

Now to test the hypothesis that the Local and FL models with averaged thresholds could be applied to detect anomalies in similar data not containing a baseline. this is the case of the hospital dataset since all samples are considered anomalous. Thus, there can be no false positive results, making it unnecessary to explore the variation of the precision.

We proceeded with the application of the models trained in the experiments described in Subsection III-C to the hospital data, and using averaging for the threshold function. The parameters for preprocessing are the same as in the last subsection. It should be noted, however, that results must be taken with due care since both P1 and P2 consist of data collected via wearable devices, while hospital data is acquired via specific cardiac sensing devices.

Figure 9 shows the mean scores of applying both the local and the FL models to the HSM data. As we can see, over the first training rounds, these approaches can obtain recall values competitive with the scores in Table II. However, the best results are still near 0.45 (0.429 and 0.450 for the P1 and P2 models, respectively), making it almost the equivalent of tossing a coin on average.

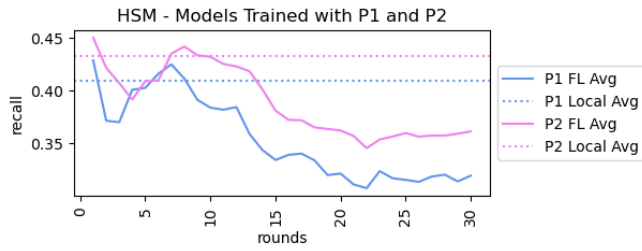


Fig. 9. Recall score of the local and FL models when applied to the hospital data. The AD is done using the mean threshold found with the training data.

The FL approach degrades its performance along the consecutive grounds, raising the conclusion that, in the case of biometric signals like heart rate, it might not be the best choice of action for implementation in a real prototype to be employed in a scenario where no baseline can be established, due to the risk of the model becoming too good at reproducing the input to the point where the threshold can no longer distinguish between normal and anomalous signals.

IV. CONCLUSIONS

To find a methodology that enables the detection of anomalies in data collected using hospital heart rate sensing devices and lacking a baseline, we used models trained using similar data collected via wearable devices. Using the work of Bogu and Snyder [2], we performed several steps of alterations, in terms of preprocessing and models' training, to better suit the hospital data. The experiences here reported also encompass training via federated learning. For a more comprehensive comparison, models were trained under different settings: individually, trained using the data of all the individuals in the dataset - either by joining the baseline data of all the individuals or via FL aggregation - and using different anomaly threshold definitions.

After comparing the mean performance of the several models resulting from the experiments we concluded that, as expected, the individually trained models achieve higher scores. However, when applying models trained with the data of multiple individuals to new users the recall thus obtained is similar in both range and behaviour when compared to the mean scores observed while testing the users that participated in the training process. Furthermore, the models trained via FL can surpass models trained with data grouped from multiple individuals. In terms of threshold variation, using $\bar{X} + \sigma$ resulted in increasing the recall more than worsening the precision, both when using an averaged threshold and using individual ones.

We applied the best models trained with data from multiple individuals and using an averaged threshold to hospital data, that is, data obtained via medical heart rate monitors. Those data contain no samples that can be used to train the models or to calculate a threshold since all are considered anomalous samples in terms of altered heart rate. The mean recall score obtained for this dataset is higher than the ones obtained with the original sets. Nevertheless, the highest value obtained,

0.45, is still below that of a random choice. We conclude that the application of pre-trained anomaly detection models to data from new users needs further exploration. Moreover, for this particular biometric signal treatment (heart rate), FL approaches show a severe decrease in performance over successive rounds, due to decreasing differences between the loss value of normal and anomalous samples.

In terms of limitations in this work, we considered each hour to be a different sample, while related works also considered classifying days, which we intend to explore shortly. Furthermore, we only explored one FL aggregation function, while there are other options to be considered. It would also be interesting to further explore the effect of applying and omitting data augmentation with this type of signal on the FL context, as well as how equivalent the signals in the different datasets are since they vary in the type of machine that collected them, and the context: daily wear versus hospitalization in an ICU.

REFERENCES

- [1] A. Patcha and J.-M. Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends," *Computer Networks*, vol. 51, no. 12, pp. 3448–3470, 2007. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S138912860700062X>
- [2] G. K. Bogu and M. P. Snyder, "Deep learning-based detection of covid-19 using wearables data," *medRxiv*, 2021. [Online]. Available: <https://www.medrxiv.org/content/early/2021/01/09/2021.01.08.21249474>
- [3] F. F. Abir, K. Alyafei, M. E. H. Chowdhury, A. Khandakar, R. Ahmed, M. M. Hossain, S. Mahmud, A. Rahman, T. O. Abbas, S. M. Zughaier, and K. K. Naji, "Pcovnet: A presymptomatic covid-19 detection framework using deep learning model using wearables data," *COMPUTERS IN BIOLOGY AND MEDICINE*, vol. 147, AUG 2022.
- [4] T. Mishra, M. Wang, A. A. Metwally, G. K. Bogu, A. W. Brooks, A. Bahmani, A. Alavi, A. Celli, E. Higgs, O. Dagan-Rosenfeld, B. Fay, S. Kirkpatrick, R. Kellogg, M. Gibson, T. Wang, E. M. Hunting, P. Mamic, A. B. Ganz, B. Rolnik, X. Li, and M. P. Snyder, "Pre-symptomatic detection of covid-19 from smartwatch data," *NATURE BIOMEDICAL ENGINEERING*, vol. 4, no. 12, pp. 1208–1220, DEC 2020.
- [5] A. Alavi, G. K. Bogu, M. Wang, E. S. Rangan, A. W. Brooks, Q. Wang, E. Higgs, A. Celli, T. Mishra, A. A. Metwally, K. Cha, P. Knowles, A. A. Alavi, R. Bhasin, S. Panchamukhi, D. Celis, T. Aditya, A. Honkala, B. Rolnik, E. Hunting, O. Dagan-Rosenfeld, A. Chauhan, J. W. Li, C. Bejikian, V. Krishnan, L. McGuire, X. Li, A. Bahmani, and M. P. Snyder, "Real-time alerting system for covid-19 and other stress events using wearable data," *NATURE MEDICINE*, vol. 28, no. 1, pp. 175+, JAN 2022.
- [6] H. B. McMahan, E. Moore, D. Ramage, and B. A. y Arcas, "Federated learning of deep networks using model averaging," *CoRR*, vol. abs/1602.05629, 2016. [Online]. Available: <http://arxiv.org/abs/1602.05629>