Random Forest for Data Aggregation to Monitor and Predict COVID-19 Using Edge Networks

Mainak Adhikari, M. Ambigavathi, and Varun G Menon, and Mohammad Hammoudeh

Abstract

Medical sensors and distributed edge networks hold promise in advanced control and prediction of infectious diseases, such as COVID-19. Their integration can lower communication latency, bandwidth utilization, and energy consumption while achieving high application scalability and reliability. Existing edge-devices-enabled e-healthcare frameworks face two serious challenges that affect their success. First, real-time medical sensors generate similar or redundant readings used for disease prediction, mostly for non-COVID-19 patients, which increases data transmission latency and energy consumption. Second, predicting the risk level of COVID-19 patients using lightweight machine learning requires minimal training time to be useful. To address these challenges, we develop an edge-centric e-healthcare framework for online health data monitoring and analysis to predict the risk level of COVID-19 patients. In the proposed framework, a statistical data aggregation method (*MEAN* function) is deployed at local gateway devices to remove redundant data to minimize communication latency and energy usage. Geo-distributed edge servers are used to predict with 97 percent accuracy the risk level to each patient with Random Forest (RF) on the aggregated data. Finally, the efficiency of the RF algorithm is demonstrated with standard classification techniques.

Introduction

Since December 2019, the COVID-19 pandemic has heavily affected some of the world's most populated countries such as the United States, India, Brazil, and China. As of 20 May 2021, more than 165 million infections in more than 200 countries were reported by the World Health Organization(WHO),[1] with more than 3.42 million deaths and more than 100.2 million people recovered. As the COVID-19 virus is transferred to the human body through droplets, it is recommended to COVID-19 patients to self-isolate at home to reduce the spread of the virus. Around 80 percent of COVID-19 affected patients recovered while staying at home following advice from medical professionals [1]. Thus, it is vital to have the technology and procedures to remotely monitor COVID-19 patients during their home isolation period.

The availability of wearable body sensor network (WBSN), low-cost medical sensors (MSs), and advanced wireless technologies, such as beyond 5G (B5G) and 6G, lead to the wide adoption of smart healthcare services [2]. Such services provide remote patient monitoring while meeting two key quality of service (QoS) parameters, namely, low data delivery latency and power consumption. MS devices play a critical role in monitoring the current health status of COVID-19 patients for various symptoms such as temperature, pulse rate, oxygen saturation, and heartbeat, and transferring the gathered data to remote computing devices for further analysis [3]. However, existing solutions do not often meet the required level of QoS due to their reliance on centralized cloud services for data collection and processing.

The geo-distribution of edge devices in edge computing has made B5G and 6G networks an attractive solution to localized data collection and processing [4]. Edge-based applications result in low communication latency and power consumption while meeting various user application and service provider constraints [5]. Moreover, artificial intelligence (AI)-enabled edge computing can further enhance data analysis results for real-time applications such as smart healthcare [6]. For instance, standard machine learning techniques can be applied to remotely collected patients' MS data to predict whether a patient is COVID-19 positive or not; this is defined as the "risk level."

To reduce pressure on hospital admissions related to COVID-19, we propose a remote patient monitoring AI-assisted edge computing system. This system provides the real-time status of patients and COVID-19 related risk. The key contributions of this work are:

- Develop a new *edge-centric e-healthcare model* for monitoring COVID-19 symptoms and analyze various risk levels remotely.
- Adopt a statistical data aggregation methodology (*MEAN* function) at the edge of the network to minimize the transmission of redundant MS data to the remote medical edge server (MES) for further analysis.
- Use Random Forest (RF) to predict the risk levels associated wih COVID-19 patients using aggregated/instant monitored symptoms.
- Perform experimental analysis to assess the efficacy of the designed data aggregation and RF algorithms in edge networks in terms of various performance metrics.

Edge-Centric e-Healthcare Model

Edge computing brings the processing power near the deployed network resources. Deploying AI techniques at the edge level helps to analyze the collected MS data with higher accuracy while minimizing latency and power consumption. Alsaeedy *et al.* designed a cellular network to detect the regions at high risk of spreading COVID-19 [7]. Li *et al.* developed efficient training of COVID-19 classification networks using a small number of COVID-19 CT scan datasets with a self-supervised learning technique [8]. Oh *et al.* introduced a patch-based convolutional neural network technique using small-scale trainable parameters for COVID-19 diagnosis [9]. Fan *et al.* developed a semi-supervised segmentation framework to detect lung infection segmentation of COVID-19 patients using CT scan images [10]. Lin *et al.* developed a data fusion strategy for enhancing the security and privacy of data in COVID-19 applications [11].

Recently, Rahman *et al.* developed a distributed COVID-19 management framework that uses deep learning techniques at B5G-enabled edge networks [12]. Similarly, Hossain *et al.* designed another edge-based healthcare model to detect COVID-19 using chest X-ray or CT scan images in the B5G net-

Mainak Adhikari is with the University of Tartu.

M. Ambigavathi is with Anna University.

Varun G. Menon is with the SCMS School of Engineering and Technology.

Mohammad Hammoudeh is with Manchester Metropolitan University.

Digital Object Identifier: 10.1109/IOTM.0001.2100052

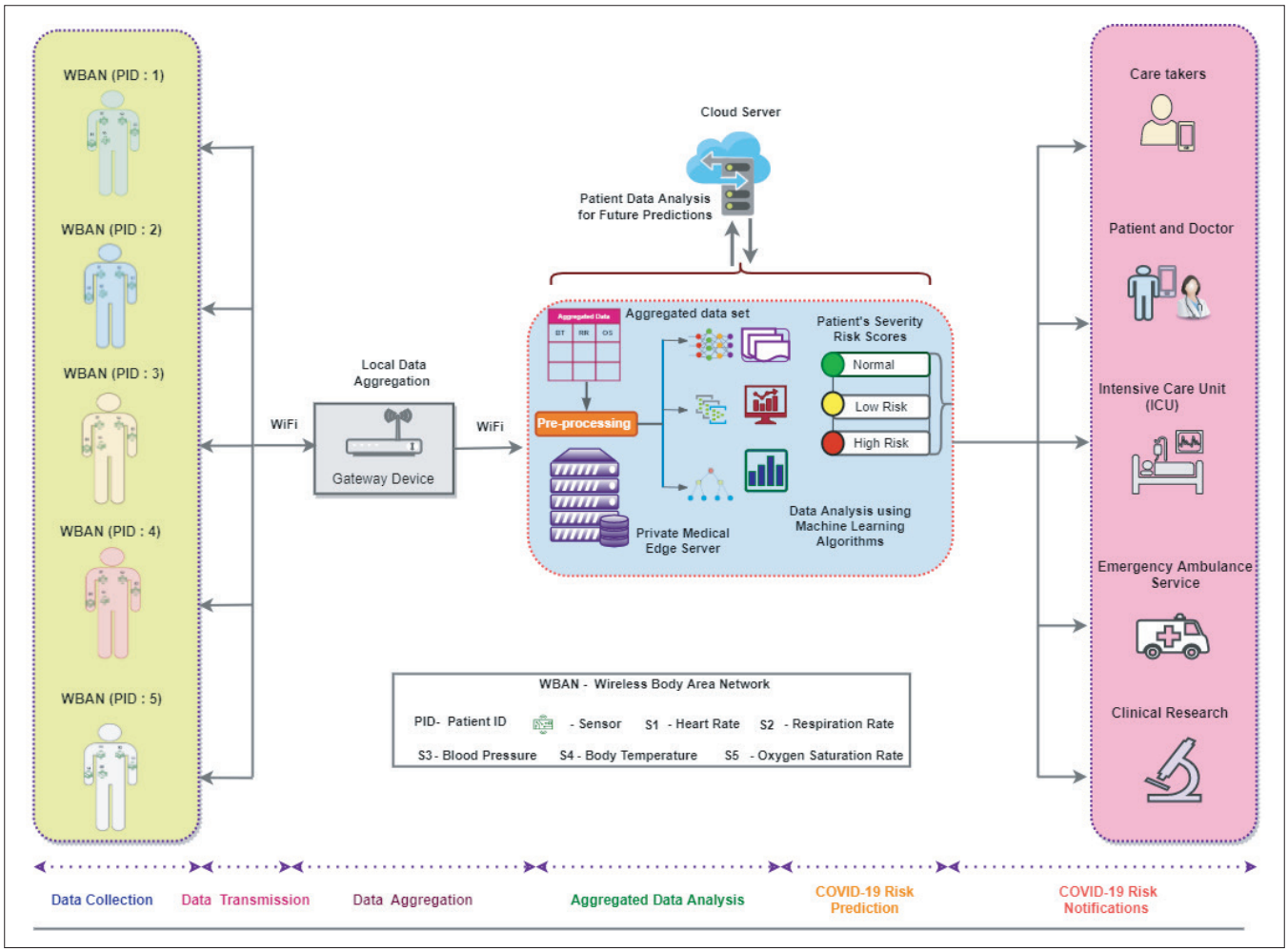

FIGURE 1. A component diagram of the proposed edge-centric e-healthcare model.

work [13]. However, such efforts on the monitoring and detection of the risk factor of COVID-19 patients using standard MSs are not edge-centric, and are often characterized by low accuracy, high latency, and high power consumption.

Cloud computing is often used in the literature to meet the high computational requirements of advanced AI/machine learning techniques. To analyze the health data in the cloud, data must be first moved from the source location to the centralized cloud server location. Data transportation increases latency and bandwidth utilization and endangers data privacy. As hospital administration is at a crisis point due to the increasing number of COVID-19 cases, it is vital to explore and design an intelligent healthcare framework to monitor patients' health symptoms remotely. Using advanced wearable or medical sensors, it is possible to predict and detect abnormality in a patient's health at the edge of the network in near real time using AI techniques.

Motivated by the above-mentioned challenges, we develop a new edge-centric e-healthcare model for remotely monitoring and identifying COVID-19 patients. This model is illustrated in Fig. 1. The proposed model comprises five major components, namely a set of MSs, a local gateway device (LGD), a local MES, a centralized cloud server, and a local hospital. MSs include a number of heterogeneous sensors such as heartbeat (HB), respiration rate (RR), blood pressure (BP), body temperature (BT), and oxygen saturation (OS). MSs are connected to the human body to monitor the body conditions, and they transmit the bio-signal data to the local MES through an LGD for further analysis. Since the monitored health data is time-sensitive, local analysis of the monitored data is a critical task. Transmitting the continuous stream of MS data to the remote MES requires sufficient energy supply and leads to increased network latency and congestion. To address this issue, the proposed model uses statistical data aggregation at the gateway level to remove similar MS readings within a given time interval before transmission to the MES for further analysis. This reduces the overall network transmission latency and power consumption.

When the local MES receives the aggregated MS data from the LGD, it analyzes it for identifying the risk factor of a COVID-19 patient. Initially, the MES uses various types of data preprocessing techniques, namely missing data filtering and data normalization, for removing irrelevant records and noise. Then a standard RF technique is applied to calculate the risk level of a COVID-19 patient [14] as High Risk, Low Risk, and Normal. Finally, a report is generated for each patient with detailed health status, and a recommendation is transmitted to the local hospital authority to take appropriate action as per the risk factor analysis. The MSs' data and the analysis results data are stored on centralized cloud servers for future analysis due to the limited storage and power capacity of the local MES.

Data Aggregation in Edge Networks

Data aggregation in edge networks is the process of compiling the sensed data in an LGD before transmitting it to a remote edge server [15]. Real-time sensors often generate redundant data; thus, transferring all the gathered data can increase network congestion, bandwidth wastage, and communication energy consumption. In the proposed edge-centric e-healthcare model, data generated by MSs in a WBSN are time-sensitive and include several redundant health symptoms, especially in

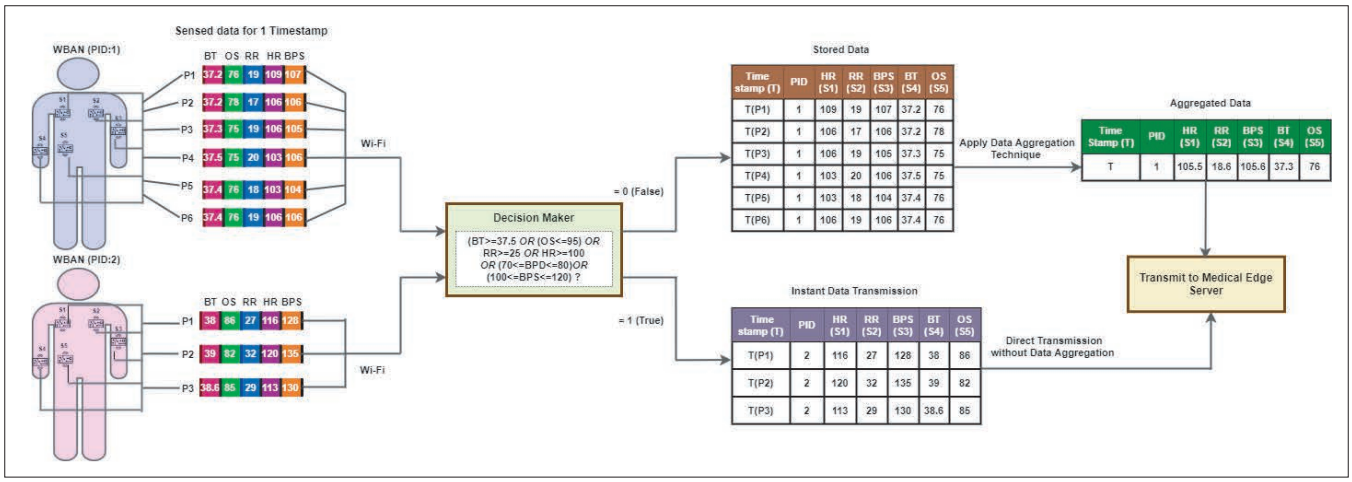

FIGURE 2. Illustration of aggregation for a single timestamp at a gateway.

non-COVID-19 patients, which increases the energy consumption and reduces the disease prediction accuracy. Data redundancy can be reduced by extracting useful information from the gathered health data via a data aggregation methodology that runs at the edge of the networks.

For data aggregation, we use a statistical data aggregation methodology, namely the *MEAN function*. As the symptoms of a non-COVID-19 patient do not fluctuate too much for a single interval, we use the MEAN function to aggregate the data and transmit the average value of each sensed modality for every interval. In the proposed model, initially, the MSs of each monitored patient transmits health readings at periodic time intervals to the decision maker (DM) module of the LGD. The DM makes a decision on whether to transmit the data to the remote MES based on the following rules.

- **IF** $((BT \geq \theta_1)\ OR\ ((OS \leq \theta_2))OR\ (RR \geq \theta_3)OR\ (HR \geq \theta_4)OR\ ((\theta_5 \leq BP_D \leq \theta_6)OR\ (\theta_7 \leq BP_S \leq \theta_8))== 0)$ (i.e., when the symptoms are within their threshold values/limits)
  THEN the monitored data is placed in the local database, and MEAN is applied to aggregate the redundant data. Here, θ_1, θ_2, θ_3, θ_4, θ_5, θ_6, θ_7, and θ_8 represent the threshold values of *BT*, *OS*, *RR*, *HR*, lower and upper limits of *BP* for *diastolic*, and lower and upper limits of *BP* for *systolic health symptoms*, respectively.
- **IF** $((BT \geq \theta_1)OR\ ((OS \leq \theta_2))OR\ (RR \geq \theta_3)OR\ (HR \geq \theta_4)OR\ ((\theta_5 \leq BP_D \leq \theta_6)OR(\theta_7 \leq BP_S \leq \theta_8))== 1)$, that is, when any one of the symptoms is beyond its threshold value (e.g., *BT*, *OS*, *RR*, and *HR*, or not within its range (e.g., *BP*)
  THEN the monitored data is instantly/directly transmitted to the MES without waiting for data aggregation for an immediate decision.

For illustration, the threshold limits for a normal patient of *BT*, *OS*, *RR*, *HR*, lower and upper limits of *BP*, and lower and upper limits BP_S for *systolic symptoms* are 37.5, 95, 25, 100, (70, 80), and (100, 120), respectively. A critical patient has some abnormalities in his/her body, which can vary the threshold limits of those symptoms and increase the likelihood of being at COVID-19-related emergency risk. The sample illustration of the proposed data aggregation scheme in edge networks is shown in Fig. 2, where two COVID-19 patients (*PID:1* and *PID:2*) are monitored through five MSs. The DM of the LGD receives the MSs' readings at a single interval. From Fig. 2, it is observed that the body condition readings of *PID:1* received from different MSs during a single interval are less than the corresponding threshold values. Thus, according to the first condition, the DM stores the data at the local database, which is aggregated with the MEAN function and transmits the aggregated data to the MES instead of transmitting six redundant frames. However, for *PID:2*, some of the monitored health symptoms are greater than their corresponding threshold values. In such a situation, according to the second condition, all the values are transmitted to the remote MES for immediate decision making.

The main motivations of introducing data aggregation in the field of healthcare are listed as follows:

- To reduce the network traffic load and congestion while minimizing communication latency and power consumption
- To reduce the redundancy in MS data to reduce data processing delays
- To increase the lifetime of the MS devices and the network while utilizing the network bandwidth efficiently

Edge-Centric COVID-19 Disease Prediction

The proposed *Edge-Centric COVID-19 Disease Prediction* module predicts the disease and identifies the risk factor of patients using various data preprocessing and standard RF techniques. The RF classifier with a set of multiple decision trees is a powerful multi-way classification technique for disease prediction. The higher number of decision trees helps the RF classifier to provide higher accuracy. The main advantages of the RF classifiers over the standard classification techniques are:

- It handles noisy and missing data efficiently during prediction.
- It performs well over continuous and categorical variables.
- It reduces the variance of the dataset and overfitting problem in decision trees during prediction.

The proposed module requires some careful consideration for disease prediction and identifications at local MES, as shown in Fig. 3.

Initially, the proposed prediction module uses various data preprocessing techniques to remove the noise and inconsistent information from the monitored dataset. While transmitting the aggregated/monitored data to the remote MES from the LGD, some data may be damaged due to signal artifacts. Thus, we used the missing-value filtering approach using the standard Kalman filter to prepare a noise-free dataset. Furthermore, the monitored health symptoms contain several features with different numeral values, which increase the burden of disease prediction. Thus, we use the standard max-min normalization function to re-scale the monitored values for increasing the prediction accuracy.

Next, the proposed module used the standard RF technique because RF technique performs better for categorical and continuous variables while handling the missing values and noise data efficiently with higher accuracy. Therefore, the RF technique is used to predict the symptoms of patients based on the five aggregated/monitored parameters and categorizes the risk factors of patients into three groups: normal, low, and high risk. Based on these severity levels of patients, healthcare professionals take appropriate actions and store the analyzed data to the centralized cloud server for future analysis.

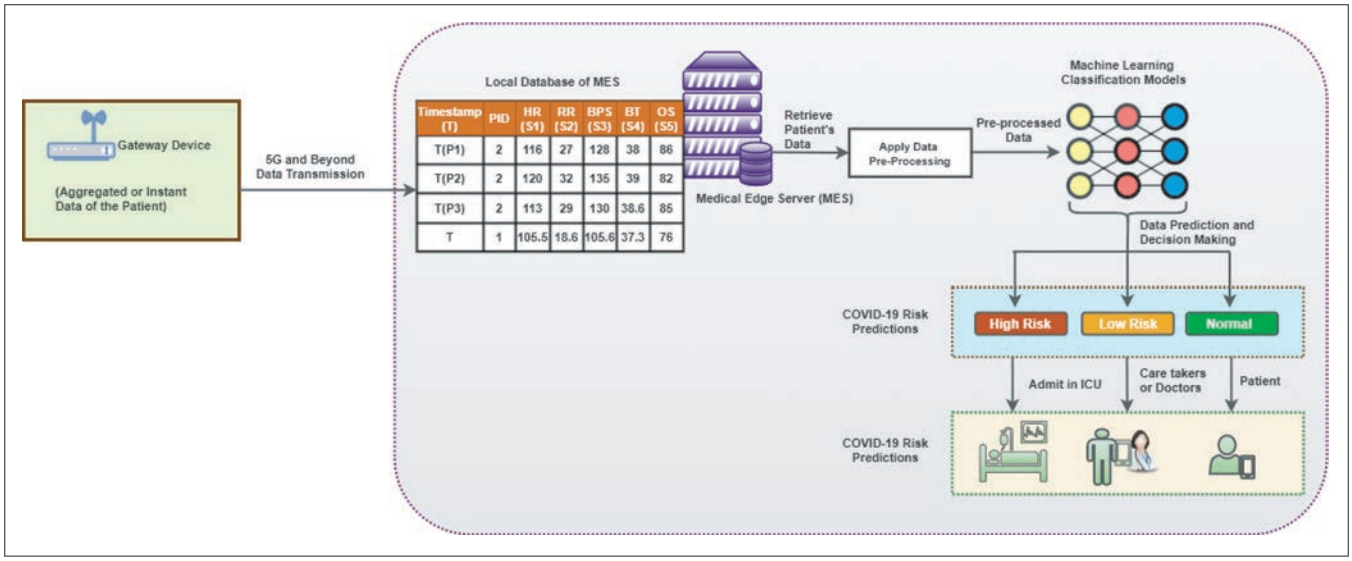

FIGURE 3. Illustration of disease prediction and risk analysis.

Performance Evaluation

Using a working prototype, we conduct a number of empirical tests to validate the proposed edge-centric e-healthcare model using various network performance matrices, namely delay and power consumption. Our experiments aim to illustrate the value of data aggregation at the network edge level for COVID-19 risk level prediction and identification using the RF technique using a synthetic dataset. This dataset is generated based on the observed symptoms of various COVID-19 patients using various statistical parameters. The ranges of the five heath parameters of 100 patients were collected and validated based on the guidance of a medical expert from the government Stanley Medical College (SMC), Chennai. To set up the network parameter, we consider the maximum payload size for this simulation of 750 bytes. The delay for higher threshold vital data is set to 1 ms, and the total number of MSs is 5 for each patient. The idle, initial, transmission, and reception power of each sensor node are set as 0.035 W, 0.615 W, 15 W, and 30 W, respectively.

We also generate a synthetic dataset to validate the standard RF technique over existing classifications models such as decision tree (DT), gradient boosting (GB), stochastic gradient descent (SGD), Gaussian naïve Bayes (GN), support vector classifier (SVC), and *K*-nearest neighbor (KNN) in terms of COVID-19 prediction and identification. This dataset has 50 COVID-19 infected patients and 50 non-COVID-19 affected patients with five features received from the MSs. Next, the dataset is divided into training set (80 percent) and testing set (20 percent), respectively, where the training and testing sets were mutually exclusive for each iteration. Furthermore, the empirical results of all classification models are evaluated using the scikit-learn library package, and the comparative results are visualized using Matplotlib library.

The simulation results of the network parameters for the proposed edge-centric e-healthcare model in terms of delay and power consumption are shown in Figs. 4 and 5, respectively. It is observed that the data aggregation at the network edge level reduced the delay and power consumption of the network while transmitting the health symptoms from the LGD to the MES for further analysis. Also, from Figs. 4 and 5, it is observed that the overall delay and power consumption were reduced for aggregated data due to the removal of redundant data at the edge of the network before transmitting to the remote edge servers for further analysis.

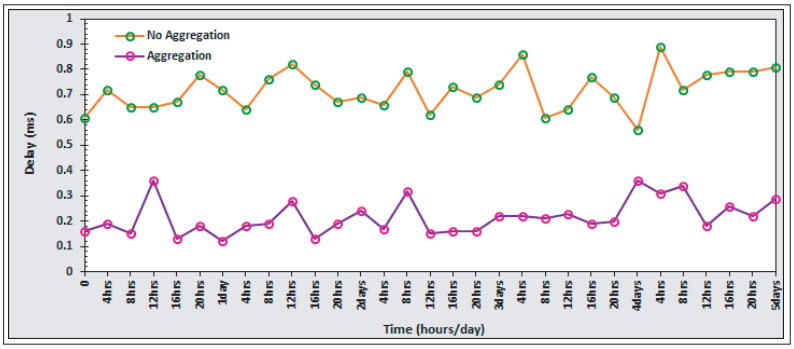

FIGURE 4. Performance analysis in terms of communication delay.

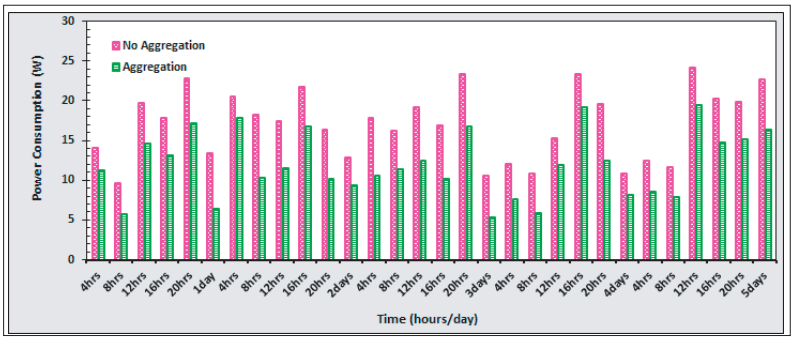

FIGURE 5. Performance analysis in terms of power consumption.

The performance analysis of the RF technique and the standard classification models with the non-aggregated (WoA) and aggregated (WA) data over various statistical parameters are shown in Fig. 6. It is observed that the standard classification models provide higher accuracy of aggregated data instead than the non-aggregated one. Further, the standard RF technique provides higher accuracy (97 percent) and minimum root mean square error (RMSE) (0.29) compared to the standard classification models. Thus, it is evident that the proposed edge-centric e-healthcare model outperforms the existing healthcare model in terms of COVID-19 prediction at the local MES with minimum delay and network power consumption.

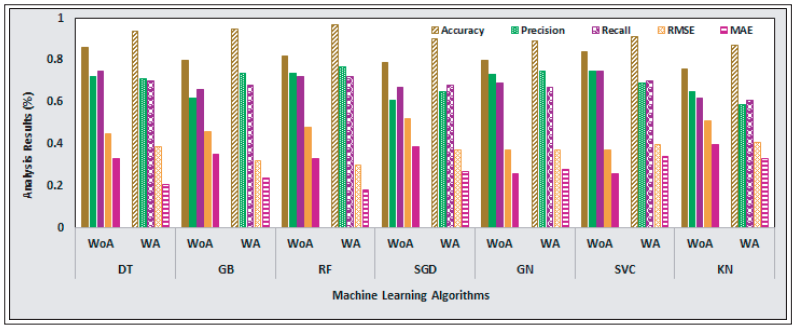

FIGURE 5. Performance analysis in terms of power consumption.

Conclusion and Future Directions

This article advocates an edge-centric e-healthcare model over B5G networks to monitor and analyze the risk level for COVID-19 patients remotely using a standard RF technique. The main objective of this work is to minimize the latency and power consumption during the transmission of the monitored health data through an LGD using a statistical data aggregation technique. The proposed model analyzes the monitored data at a local MES using data preprocessing and a standard RF technique to identify the risk factor of each patient with 97 percent accuracy. This can potentially help to reduce overcrowding in hospitals by measuring the risk to COVID-19 patients remotely.

In the future, we will introduce various deep learning models in the MES for analyzing and predicting medical risk level of COVID-19 patients. Further, we also introduce various data and information fusion techniques to prepare a noise-free dataset at the edge level for increasing the prediction accuracy with minimum MAE. Additionally, we will introduce a pervasive edge computing model for COVID-19 prediction with ultra-low latency and better reliability. The commercial launch of B5G and 6G communication techniques and edge-centric analysis increases the opportunities for further research in the field of healthcare. Some of the important future research directions are:

- Introduce smart-contract-based service subscription with blockchain to provide secure and accurate health services to patients using a smart-contract-based linked service agreement.
- Introduce Tactile Internet and B5G communication to fulfill the requirements of ultra-low latency and ultra-high reliability during monitoring and prediction.
- Introduce a lightweight AI technique at resource-constrained edge devices that can provide higher accuracy during health symptom predictions while utilizing the resources and minimizing the processing cost.
- Develop a distributed intelligent healthcare framework with federated or distributed learning and introduce a software-defined network to increase data analysis accuracy at lower cost.

References

[1] A. Roy *et al.*, "Efficient Monitoring and Contact Tracing for Covid-19: A Smart IoT-Based Framework," *IEEE Internet of Things Mag.*, vol. 3, no. 3, Sept. 2021, pp. 17–23.
[2] M. Hammoudeh *et al.*, "A Service-Oriented Approach for Sensing in the Internet of Things: Intelligent Transportation Systems and Privacy Use Cases," *IEEE Sensors J.*, 2020.
[3] M. Adhikari and A. Munusamy, "icovidcare: Intelligent Health Monitoring Framework for Covid-19 Using Ensemble Random Forest in Edge Networks," *IEEE Internet of Things J.*, vol. 14, 2021, pp. 1–15.
[4] M. J. Piran and D. Y. Suh, "Learning-Driven Wireless Communications, Towards 6G," *2019 Int'l. Conf. Computing, Electronics & Commun. Engineering*, 2019, pp. 219–24.
[5] A. Hazra *et al.*, "Stackelberg Game for Service Deployment of IoT-Enabled Applications in 6G-Aware Fog Networks," *IEEE Internet of Things J.*, 2020, pp. 1–9.
[6] H.-N. Dai, M. Imran, and N. Haider, "Blockchain-Enabled Internet of Medical Things to Combat Covid-19," *IEEE Internet of Things Mag.*, vol. 3, no. 3, Sept. 2020, pp. 52–57.
[7] A. A. Alsaeedy and E. K. Chong, "Detecting Regions at Risk for Spreading Covid-19 Using Existing Cellular Wireless Network Functionalities," *IEEE Open J. Engineering in Medicine and Biology*, vol. 1, 2020, pp. 187–89.
[8] Y. Li *et al.*, "Efficient and Effective Training of Covid-19 Classification Networks with Self-Supervised Dual-Track Learning to Rank," *IEEE J. Biomedical and Health Informatics*, vol. 24, no. 10, 2020, pp. 2787–97.
[9] Y. Oh, S. Park, and J. C. Ye, "Deep Learning Covid-19 Features on CXR Using Limited Training Data Sets," *IEEE Trans. Medical Imaging*, 2020, pp. 1–13.
[10] D.-P. Fan *et al.*, "Inf-Net: Automatic Covid-19 Lung Infection Segmentation from CT Images," *IEEE Trans. Medical Imaging*, vol. 39, no. 8, 2020, pp. 2626–37.
[11] H. Lin *et al.*, "Privacy-Enhanced Data Fusion for Covid-19 Applications in Intelligent Internet of Medical Things," *IEEE Internet of Things J.*, 2020, pp. 1–1.
[12] M. A. Rahman *et al.*, "B5g and Explainable Deep Learning Assisted Healthcare Vertical at the Edge: Covid-i9 Perspective," *IEEE Network*, vol. 34, no. 4, July/Aug. 2020, pp. 98–105.
[13] M. S. Hossain, G. Muhammad, and N. Guizani, "Explainable AI and Mass Surveillance System-Based Healthcare Framework to Combat Covidi9 Like Pandemics," *IEEE Network*, vol. 34, no. 4, July/Aug. 2020, pp. 126–32.
[14] V. K. Gupta *et al.*, "Prediction of Covid-19 Confirmed, Death, and Cured Cases in India Using Random Forest Model," *Big Data Mining and Analytics*, vol. 4, no. 2, 2021, pp. 116–23.
[15] M. Hammoudeh and R. Newman, "Information Extraction from Sensor Networks Using the Watershed Transform Algorithm," *Info. Fusion*, vol. 22, 2015, pp. 39–49.

Biographies

Mainak Adhikari [mainak.ism@gmail.com] is currently working as a postdoctoral research fellow at the University of Tartu, Estonia. He received his Ph.D. from IIT (ISM) Dhanbad, India, in 2019. He obtained his M.Tech. from Kalyani University in 2013. He earned his B.Tech. degree from WBUT in 2011. He is an Associate Editor of *Cluster Computing* and *Physical Communication Magazine*.

M. Ambigavathi [M] (ambigaindhu8@gmail.com) received her B.E. and M.E. degrees in CSE from Anna University in 2010 and 2012, respectively. She completed her Ph.D. at Anna University, Chennai. She is an active professional member of ACM, and also a life member of IETE, ISTE, and CSI. Her research interests include WSN, WBAN, machine and deep learning, and IoT.

Varun G. Menon [varunmenon@ieee.org] is currently an associate professor and head of the Department of Computer Science Engineering, and International Collaborations and Corporate Relations in charge at SCMS School of Engineering and Technology, India. He completed his Ph.D. in computer science and engineering and holds an M.Tech. degree in computer and communication with University First Rank. He also holds an M.Sc. in applied psychology, an M.B.A. in human resource management, and a Diploma in training and development.

Mohammad Hammoudeh [SM] (hammoudeh@ieee.org) is a professor and Chair of Cyber Security in the Department of Computing and Mathematics at Manchester Metropolitan University. He heads the CfACS Internet of Things Lab he founded in 2016, where he leads a multi-disciplinary group of research associates and Ph.D. students. He is a Fellow of the Higher Education Academy United Kingdom. He currently investigates ways of improving industry practice to allow for guaranteed security and distributed computing applications that work effectively every time.

Footnotes

[1] https://www.who.int/