

Criticality and Utility-aware Fog Computing System for Remote Health Monitoring

Moirangthem Biken Singh, Navneet Taunk, Naveen Kumar Mall, and Ajay Pratap, *Member, IEEE*

Abstract—Growing remote health monitoring system allows constant monitoring of the patient's condition and performance of preventive and control check-ups outside medical facilities. However, the real-time smart-healthcare application poses a delay constraint that has to be solved efficiently. Fog computing is emerging as an efficient solution for such real-time applications. Moreover, different medical centers are getting attracted to the growing IoT-based remote healthcare system in order to make a profit by hiring Fog computing resources. However, there is a need for an efficient algorithmic model for allocation of limited fog computing resources in the criticality-aware smart-healthcare system considering the profit of medical centers. Thus, the objective of this work is to maximize the system utility calculated as a linear combination of the profit of the medical center and the loss of patients. To measure profit, we propose a flat-pricing-based model. Further, we propose a swapping-based heuristic to maximize the system utility. The proposed heuristic is tested on various parameters and shown to perform close to the optimal with criticality-awareness in its core. Through extensive simulations, we show that the proposed heuristic achieves an average utility of 96% of the optimal, in polynomial time complexity.

Index Terms—IoT, WBAN, Fog Server, Smart Healthcare, Algorithm.



1 INTRODUCTION

IoT assisted remote healthcare has recently gained popularity as it seemed to be an efficient solution for the challenges faced in remote healthcare sector [1]. The lack of healthcare facilities in rural *India* can be greatly assisted by IoT-based remote health monitoring in a cost effective way [2]. However, the development of IoT tackles various problems such as limitation of available resources, limited accessibility of IoT systems for people living in rural areas.

In remote health monitoring, the patient is equipped with sensors, and the data generated by the sensor is sent to a gateway through Wireless Body Area Network (WBAN). Further, the gateway device sends the data to the base station through beyond-WBAN. In beyond-WBANs, 5G communication is emerging an efficient solution for fast and real-time transmissions [3], [4]. The assistance of fog computing within beyond-WBAN has emerged as an efficient way to compute the data sent through the network [5]. Although their computational capacities are not as much as centralized cloud servers, they are capable enough of computing medical data packets near to patients [6]. Therefore, fog computing can reduce the latency to a greater extent, thus improving the quality of the monitoring system [7].

The people living in rural areas have a higher rate of poverty [8] and thus, they cannot afford Local Devices (LDs) or sensors on their own. They need to be assisted by the government with these pieces of equipment. Thus, a better solution is to provide these things for free or at

low price. Moreover, the charge for monitoring should be kept low. Therefore, the profit of the medical center becomes an important factor for the system model as the revenue generated would be less due to cheap monitoring cost. The medical centers providing remote healthcare to these patients would see the technological advancement for their benefit and thus, they would invest in it to make a profit out of it.

In this work, the proposed system is divided into two parts, one is intra-WBAN and the other is beyond-WBAN. Intra-WBAN consists of sensors deployed on the patients and the LD provided to the patients whereas beyond-WBAN consists of different LDs that send the data to the Fog Servers (FSs) located near the base station/access point. The encouragement of such model is described as below:

1.1 Motivation

The emerging FS assisted remote health monitoring system can be fruitful for both the patients and the medical centers as discussed below:

- As the medical data is highly critical, it has to be processed in real-time without much delay.
- The data of patients with a serious disease or other critical conditions should be given higher priority over others.
- In a practical scenario, a medical center would charge the patients for the medical service it provides.
- Therefore, there is a need of a well established cost effective mechanism between patient and healthcare service provider to monitor criticality, latency and revenue via resource allocation in remote healthcare system.

Motivated by the above objectives, we aim to formulate an FS assisted beyond-WBAN based remote health monitor-

• M. B. Singh, N. Taunk, N. K. Mall and A. Pratap are with the Department of Computer Science and Engineering, Indian Institute of Technology (Banaras Hindu University) Varanasi 221005 India. E-mail: {moirangthembsingh.rs.cse21, navneettaunk.cse18, naveenkumar-mall.cse18, ajay.cse}@iitbhu.ac.in.

ing system that minimizes the cost of patients with the profit of the medical center in consideration. Inspired by [9], we aim to use dedicated LDs to not only collect the data sent by the sensors but also those LDs have some computation power, which can be utilized to compute the patients' data locally if the condition of the patient is not much critical.

1.2 Contribution

In this paper, we design an FS assisted remote health monitoring system. The main objective of the system is to maximize the utility that depends on the profit of the medical center and the cost of patients (which depends on latency delay and their criticality described in Section 3). The main contributions of this paper are summarized below:

- A criticality and utility-aware remote health monitoring system is proposed.
- A cost-function is formulated for the patients that measure the loss of patients in terms of latency and their criticality in such a way that more critical patients are given priority over less critical patients. Moreover, the problem is formulated depending on the patients' loss and the medical center's profit, trying to balance between the two.
- A swapping-based heuristic is proposed to maximize the system utility under the constraints of permissible latency for the computation of patients' data in polynomial time complexity.
- Through extensive simulations, the proposed heuristic is found to achieve a utility of 96% of the optimal on an average.

The rest of the paper is organized as follows: Section 2 reviews the relevant work. The system model and the problem definition are introduced in Sections 3 and 4, respectively. The proposed solution and analysis are given in Sections 5 and 6, respectively. The performance study is presented in Section 7. Finally, Section 8 offers conclusions and future research directions.

2 RELATED WORKS

The authors in [10] propose a haptic communication framework for e-health systems. The primary focus of the paper is to improve haptic communications under three factors (system stability, energy consumption, and network delay). They propose a time-varying swarm algorithm to solve the formulated problem. The authors in [11] propose a cost-aware medical cyber-physical system assisted by fog computing. Their work jointly focuses on task allocation, base station association, and virtual machine placement. They propose linear-programming based heuristics to solve the formulated problem. The authors in [12] propose an energy-aware medical cyber-physical system assisted by fog computing. Their primary focus is on resource allocation to minimize energy consumption and response time. They propose a dynamic-cluster algorithm to solve the formulated problem.

In [13], the authors investigate the factors such as energy consumption, transmission delay, QoS requirement, the power limit and wireless fronthaul constraint in fog computing-based Internet of Medical Things (IoMT) for

remote health monitoring. They propose a low time-complexity sub-optimal scheme to solve the problem. The authors in [14] propose a queue-based transmission of time-sensitive medical data packets in beyond-WBAN. They propose a non-cooperative game for the above mentioned scenarios and then, they propose an analytical framework to solve the formulated problem. The authors in [9] propose a health monitoring system for IoMT considering criticality, energy and delay constraints. They propose a decentralized non-cooperative game based scheme to solve the formulated problem.

The authors in [15] propose a priority-aware time-slot allocation in WBANs. They extend the evolutionary game theory to solve the formulated problem. The authors in [16] propose a Nash bargaining solution for a cooperative game based priority-aware data-rate tuning in WBANs model. Moreover, Table 1 summarizes the closely related works available in the literature.

Shortcomings of Existing Approaches: In some of the existing approaches, only intra-WBAN transmission is considered on the basis of latency and criticality. Some works consider both intra-WBAN and beyond-WBAN transmission under latency and criticality constraints. However, none of the existing approach have considered the profit of the medical center which is one of the main objective of our system. Therefore, different from the above work, we propose a novel criticality-aware health monitoring system with the profit of the medical center in consideration. Moreover, we propose a novel swapping-based heuristic to solve the formulated problem in polynomial time complexity.

TABLE 1: A relative comparison

Author	Primary Problem	Brief Description
Feng et. al. [10]	Resource Allocation, Packet Drop, Energy Harvesting	Swarm Intelligence
Gu et. al. [11]	Task Allocation, Base Station Association, Machine Placement	LP-based heuristics
Apat et. al. [12]	Energy Consumption, Response Time, Resource Allocation	Dynamic clustering
Qui et. al. [13]	QoS requirement, power limit and wireless fronthaul constraint.	Lagrange Multipliers based
Yi et. al. [14]	Data Priority, Latency	Non-cooperative Game
Ning et. al. [9]	Medical Criticality, Age Of Information, Energy Consumption	Non-cooperative Game
Misra et. al. [15]	Data Priority, Time-slot Allocation	Evolutionary Game Theory
Misra et. al. [16]	Data Priority, Data-rate Tuning	Cooperative Bargaining Game
Proposed model	Criticality, Latency, Profit and Resource Allocation	Swapping-based heuristic

3 SYSTEM MODEL

As shown in Fig. 1, we consider a remote health monitoring system, provided by medical center to a set of patients

TABLE 2: Symbol description

Symbol	Description
\mathbb{S}	Set of sensors for a patient
\mathbb{X}	Set of medical criticalities for all sensors.
x_s	Medical Criticality of sensor s .
$\theta_{s,t}$	Physiological data value sensed by the sensor s at time t
$\theta_{l,s}$	Lower limit of the normal value for sensor s .
$\theta_{u,s}$	Upper limit of the normal value for sensor s .
$d_{s,t}$	Packet severity index for sensor s at time t .
$c_{s,t}$	Overall criticality index for sensor s at time t .
\mathbb{F}	Set of FSs.
\mathbb{P}	Set of patients.
$\rho_{p,t}^c$	Patient Criticality for patient p at time t .
\mathbb{H}_t	Set of strategy for patients.
$u_{p,t}$	Whether LD is chosen for computation
$q_{p,t}$	Whether FS is chosen for computation
$\eta_{p,t}$	Overall data size for patient p at time t .
$\beta_{p,t}$	CPU cycles required to compute patient p 's data at time t .
$T_{p,t}^{c,l}$	Computation time for patient p at time t by local device.
Υ	Computation capacity of patient's local device.
Γ	Computation capacity of an edge server.
m	Price per unit time for computation at FS
l	Price per unit time for computation at LD
χ_t	Revenue earned by medical center.
ϕ_t	Expenditure of medical center per edge server.
g	Expenditure of medical center per CPU cycle of computation at edge server.
k	Fixed charge per FS.
δ	Latency constraint.

$\mathbb{P} = \{1, 2, \dots, P\}^1$. The proposed problem setting is equivalent to project assignment problem in colleges where a student approaches to a professor for project assignment and professor assigns an available project to the student. Similarly in our case, patient (LD) approaches to medical center for remote monitoring and medical center assigns an available FS to the patient (LD) as shown in Fig. 1.

The proposed problem setting is divided into two parts: intra-WBAN and beyond-WBAN described as follows:

3.1 Intra-WBAN

Consider a set of sensors, $\mathbb{S} = \{1, 2, \dots, S\}$, deployed on each patient. Each sensor generates data packets and transmits them to the LD. The data packets generated by different sensors belong to different classes depending on their criticality. For instance, a sensor measuring heart rate should have higher priority over a sensor that measures skin temperature. Moreover, heart-related diseases are more serious and should be prioritized over general diseases. To facilitate this, we have considered medical criticality as a measure to prioritise the data packet.

Let medical criticalities of data generated by sensors be the set $\mathbb{X} = \{x_1, x_2, \dots, x_S\}$. If data generated by sensor s is more critical than that of sensor s' , then $x_s > x_{s'}$. It can be possible for two sensors of the same criticality class to have different medical criticalities. Here, $x_s \in [0, \infty]$. Let

1. The symbol description is given in the Table 2.

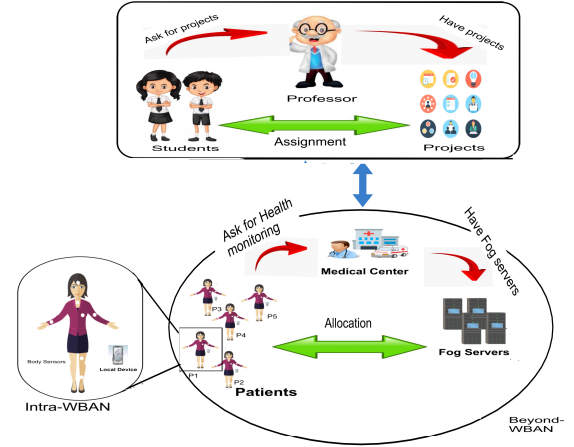


Fig. 1: System Model

$\theta_{s,t}$ be the the parameter value sensed by sensor s at time t . Let $\theta_{l,s}$ and $\theta_{u,s}$ be the reference range of that parameter under normal conditions for a healthy person. Then, packet severity index [15] at time t can be defined as,

$$d_{s,t} = \left| \frac{(\theta_{u,s} - \theta_{s,t})^2 - (\theta_{s,t} - \theta_{l,s})^2}{(|\theta_{u,s}| + |\theta_{l,s}|)^2} \right| \quad (1)$$

Let overall criticality index, $c_{s,t}$ of a sensor s at time t be the product of packet severity index and the medical criticality as follows:

$$c_{s,t} = x_s d_{s,t}. \quad (2)$$

Let p^{th} patient's criticality at time t be defined as the summation of overall criticality indices of sensors, as follows:

$$\rho_{p,t}^c = \sum_{s=1}^S c_{p,s,t}, \quad (3)$$

where $c_{p,s,t}$ is the overall criticality index for patient p and sensor s at time t . The patient criticality indicates the extent to which the health of a patient is critical. Higher the criticality value indicates more severe is the condition of the patient. After collecting the data at LD, there is a need to make a decision for its computation either at LD or FS in-order to achieve the system's constraints. Moreover, computation capacity of all LDs are considered to be uniform and equal to Υ . Let $u_{p,t}$ be a binary variable defined as below:

$$u_{p,t} = \begin{cases} 1, & \text{System selects LD for computation of } p\text{'s data;} \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

Moreover, computation time at LD for a patient p can be calculated as:

$$T_{p,t}^{c,l} = \frac{\beta_{p,t}}{\Upsilon} \quad (5)$$

where $\beta_{p,t}$ is the number of CPU cycles required for the computation of patient p 's data at time t . In the next section, we describe the transmission² and computation of data at FSs.

2. Transmission of data in intra-WBAN is beyond the scope of this work. However, the existing approach [9], can be utilized for intra-WBAN communication.

3.2 Beyond-WBAN

In this model, transmission and computation latencies of patient's data are evaluated. Consider a set of FSs $\mathbb{F} = \{1, 2, 3, \dots, F\}$ and let $q_{p,t}$ be a binary variable defined as follows:

$$q_{p,t} = 1 - u_{p,t}. \quad (6)$$

Let \mathbb{H}_t be a $P \times F$ binary matrix which denotes the choice of the system for computation of patients' data at time t where, \mathbb{H}_t is a globally accessible variable maintained at cloud server responsible for execution of the proposed heuristic discussed in the Section 5.

$$\mathbb{H}_t = \begin{bmatrix} h_{1,t}^1 & h_{1,t}^2 & \cdot & \cdot & h_{1,t}^F \\ h_{2,t}^1 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ h_{P,t}^1 & \cdot & \cdot & \cdot & h_{P,t}^F \end{bmatrix} \quad (7)$$

where,

$$h_{p,t}^f = \begin{cases} 1, & \text{FS } f \text{ computes patient } p\text{'s data;} \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

The transmission rate between patient p and FS f can be computed as follows:

$$BR_{p,f,t} = \Omega \log_2 \left(1 + \frac{W_{p,f,t}^{bey}}{N_t^{bey}} \right), \quad (9)$$

where, Ω is the channel bandwidth, $W_{p,f,t}^{bey}$ be the transmission power, and N_t^{bey} be the noise power. Mathematical expression for $W_{p,f,t}^{bey}$ can be written as, $W_{p,f,t}^{bey} = w_p G_{p,f,t}$, where w_p is the power of transmission and $G_{p,f,t}$ is the channel gain of patient p , if f server is chosen. Here, we have assumed that all patients communicate through different channels, so interference is not considered³. Moreover, transmission time between patient p and FS f can be written as:

$$T_{p,f,t}^{tr} = \frac{\eta_{p,t}}{BR_{p,f,t}}, \quad (10)$$

where, $\eta_{p,t}$ is the size of the patient p 's data. Let computation capacity of an FS be Γ . Similar to the studies [18], [19] we are assuming that all the patients use the resource of an FS equally and all FS have the same computation capacity. The computation time for a patient p due to fog computing can be calculated as:

$$T_{p,t}^{c,f}(\mathbb{H}_t) = \frac{\beta_{p,t}}{\gamma_p(\mathbb{H}_t)}, \quad (11)$$

where $\gamma_p(\mathbb{H}_t)$ is the fraction of resource used by the patient p which can be computed by dividing total resources of a server by the number of patients utilizing it. Mathematically,

$$\gamma_p(\mathbb{H}_t) = \frac{\Gamma}{\sum_{f=1}^F h_{p,t}^f \sum_{p'=1}^P h_{p',t}^f}. \quad (12)$$

The criticality-awareness with low-latency is an important factor for the system. Thus, one of our objectives is to minimize the loss incurred for the patients which is a

parameter of the criticality-awareness. The cost function for the patients can be computed as the weighted sum of computation time and the transmission time, where the weights are the patients' criticality. Mathematically,

$$J(t) = \sum_{p \in \mathbb{P}} \rho_{p,t-1}^c \left(\sum_{f=1}^F h_{p,t}^f T_{p,f,t}^{tr} + T_{p,t}^{c,f}(\mathbb{H}_t) + u_{p,t} T_{p,t}^{c,l} \right). \quad (13)$$

We can see that the cost function depends on the criticality of patients, i.e., if a patient is more critical, it will add up more to the cost, thus we have to lower the latency for that patient in the beyond-WBAN in order to reduce the cost.

As mentioned in Fig. 1, a medical center is considered for providing remote healthcare to the patients in return of service charges. As the flat-type pricing scheme is emerging as a good business model for the health providers [20], we are considering a flat-type pricing scheme to calculate the revenue of the medical center in the following.

3.2.1 Flat-type Pricing Scheme

For the computation on the LD, the medical center will charge l unit price per time slot. If the computation is done on the FS, the medical center will charge m unit price per time slot. As FS charge should be greater than that of LD, thus $m > l$. So, the revenue earned by the medical center can be calculated as,

$$\chi_t = \sum_{p \in \mathbb{P}} (u_{p,t} l + q_{p,t} m). \quad (14)$$

Now, as FSs are involved, the medical center has to bear its cost as well. Let k be the fixed expenditure per FS per time slot and g be the expenses of medical center per CPU cycle due to computation on the FS. Thus, the expenditure of medical center can be calculated as,

$$\phi_t = kF + g \sum_{p \in \mathbb{P}} q_{p,t} \beta_{p,t}. \quad (15)$$

Therefore, the profit gained by the medical center can be calculated as,

$$\Delta(t) = \chi_t - \phi_t = \sum_{p \in \mathbb{P}} (u_{p,t} l + q_{p,t} m) - kF - g \sum_{p \in \mathbb{P}} q_{p,t} \beta_{p,t}. \quad (16)$$

On observing equation (16), we can see that, for a patient p , the profit depends on whether that patient is allocated an FS or LD. Let the maximum value of $\beta_{p,t}$ be $\beta_{p,t}^{max}$. Here, $\beta_{p,t}^{max}$ can be approximated by the medical center before deciding the values of m and l . The profit earned by the medical center depends on the following constraint:

$$m - l \geq g \beta_{p,t}^{max} + \frac{kF}{P} \quad (17)$$

Here, it is ensured that if a patient is using an FS, then the profit of the medical center will be more compared to the case when he uses an LD independent of the CPU cycles of the patient data as discussed in Lemma 1.

Lemma 1. *The profit of the medical center either increases or remains constant as more patients utilize FS rather than LD for their computation.*

3. However, interference can be solved by applying methods given in [3], [17].

Proof. Let P' be the number of patients utilizing FS. So, the profit in this scenario can be given by:

$$\Delta_{t,1} = P'm + (P - P')l - kF - g \sum_{p \in P} q_{p,t} \beta_{p,t}. \quad (18)$$

Now, take any patient p' that is utilizing LD and assign him any FS for his data computation. So, the new profit in this scenario is (assuming allocation of all other patients remains the same):

$$\Delta_{t,2} = (P' + 1)m + (P - P' - 1)l - kF - g \sum_{p \in P} q_{p,t} \beta_{p,t} - g \beta_{p',t}. \quad (19)$$

Now, $\Delta_{t,2} - \Delta_{t,1}$ is given by:

$$\Delta_{t,2} - \Delta_{t,1} = m - l - g \beta_{p',t}. \quad (20)$$

From Eqs. (17) and (20), we can conclude that:

$$\Delta_{t,2} - \Delta_{t,1} \geq 0. \quad (21)$$

As we increase the number of patients that utilize FSs for their computation, the profit also increases or remains constant. Hence, proved. \square

4 PROBLEM FORMULATION

In the proposed system, we are considering two factors: the profit of the medical center and the cost of patients. We know that a medical center will try to maximize its profit under the condition that no patient will face any delay in monitoring. Moreover, the patients would want to minimize their cost to ensure that they are being properly monitored. However, both objectives cannot be achieved at the same time. Thus, we are considering utility defined as the linear combination of profit of medical center and the cost of patients as follows:

$$U(t) = \lambda_1 \Delta(t) - \lambda_2 J(t), \quad (22)$$

where, λ_1 and λ_2 are the weights assigned to the profit of the medical center and the cost of the patients respectively. The weights are taken as inverse units of the profit and the latency cost respectively, so that utility becomes unitless. The weights are dependent on the system requirements and should be considered accordingly. That means, if the system is more profit aware then, $\lambda_1 > \lambda_2$, or if the system is more criticality aware then, $\lambda_1 < \lambda_2$, or if it is equally balancing between the two, then $\lambda_1 = \lambda_2$. Moreover, we are considering a constraint on the permissible latency defined as δ so as to ensure that no patient would face a delay of more than $\frac{\delta}{\rho_{p,t-1}^c}$. Furthermore, we have considered if the criticality of the patient is more, then the permissible latency is less. Thus, the optimization problem is formulated as follows:

$$\arg \max_{\mathbb{H}_t} U(t) \quad (23)$$

Subject to the constraints:

$$\lambda_1 \geq 0, \lambda_2 \geq 0, \quad (24)$$

$$\lambda_1 + \lambda_2 = 2, \quad (25)$$

$$\forall p \in \mathbb{P}, \sum_{f'=1}^F h_{p,t}^{f'} T_{p,f,t}^{tr} + T_{p,t}^{c,f}(\mathbb{H}_t) + u_{p,t} T_{p,t}^{c,l} \leq \frac{\delta}{\rho_{p,t-1}^c}, \quad (26)$$

$$l \leq l_{max}, \quad (27)$$

$$m \leq m_{max}, \quad (28)$$

$$m - l \geq g \beta_{p,t}^{max} + \frac{kF}{P}, \quad (29)$$

$$\forall p \in \mathbb{P}, \sum_{f \in \mathbb{F}} h_{p,t}^f = q_{p,t}. \quad (30)$$

The constraints given in Eqs. (24) and (25) are the bounds on the weights. Eq. (26) is the latency constraint. Eqs. (27) and (28) put a constraint on the service charge. Eq. (29) refers to the additional constraint as defined in Eq. (17). Eq. (30) ensures that every patient is allocated at most one FS.

The formulated problem in Eqs. (23-30) is a Binary Integer Programming problem in \mathbb{H}_t decision variables, that is generally NP-hard to solve as its feasibility problem is strongly NP-complete [21]. Due to the high conditionality and hardness of the formulated problem, this paper proposes a complete framework to provide a sub-optimal solution for the maximization problem based on a swapping-based heuristic in the following section.

5 PROPOSED SOLUTION

To save the high computation charge at FS, each patient would like to compute the task at LD itself. However, while doing so, they may not meet the latency constraint, Eq. (26). So, we have to prioritise these patients for utilizing the FSs. Thus, a sub-problem here is to allocate the FSs to such patients. Let \mathbb{P}^v be the set of patients that violate the latency constraint if their data is computed at the LD. Formally,

$$\mathbb{P}^v = \{p \in \mathbb{P} : \rho_{p,t-1}^c T_{p,t}^{c,l} > \delta\} \quad (31)$$

As per Lemma (1), we can say that the profit only depends on patients who utilize the FSs and, increasing the number of patients who utilize FSs increases the earned profit of medical center. So, if all the patients in \mathbb{P}^v utilize FSs, the profit does not depend on how they are allocated the FSs. Thus, the utility depends only on the cost function as defined in Eq. (13).

In this sub-problem, we offload the patient's data to one FS. After allocation of patients in the set \mathbb{P}^v (patients violating latency constraint), our next objective is to allocate a subset of the remaining patients such that it maximizes the utility under the system constraints. We can notice that it is not possible to allocate all the patients to FSs due to the limited resources. Doing so may result in violation of the latency constraint Eq. (26) and could increase the patients' cost drastically, which would result in low utility. Consider the constraint given in Eq. (26), and let $n_{p,f,t}^{max}$ be the maximum number of patients (see Theorem 1) that can utilize the FS f for their computation if patient p utilizes it.

Theorem 1. *The maximum number of patients that can utilize the FS f for their computation, if patient p utilizes that FS, can be given as:*

$$n_{p,f,t}^{max} = \left\lfloor \left(\frac{\Gamma}{\beta_{p,t}} \right) \left(\frac{\delta}{\rho_{p,t-1}^c} - \frac{\eta_{p,t}}{\Omega \log_2 \left(1 + \frac{W_{p,f,t}^{bey}}{N_t^{bey}} \right)} \right) \right\rfloor \quad (32)$$

Proof. According to Eq. (26), if a patient p utilizes FS f , then,

$$T_{p,f,t}^{tr} + T_{p,t}^{c,f}(\mathbb{H}_t) \leq \frac{\delta}{\rho_{p,t-1}^c}, \quad (33)$$

Putting values of $T_{p,f,t}^{tr}$ and $T_{p,t}^{c,f}(\mathbb{H}_t)$ from Eqs. (10) and (11) respectively, we get

$$\frac{\eta_{p,t}}{BR_{p,f,t}} + \frac{\beta_{p,t}}{\gamma_p(\mathbb{H}_t)} \leq \frac{\delta}{\rho_{p,t-1}^c}, \quad (34)$$

where, $\gamma_p(\mathbb{H}_t) = \frac{\Gamma}{n'_{p,f,t}}$, where $n'_{p,f,t}$ are the number of patients utilizing FS f including p . After solving the inequality and putting the value of $BR_{p,f,t}$, we get,

$$n'_{p,f,t} \leq \left(\frac{\Gamma}{\beta_{p,t}} \right) \left(\frac{\delta}{\rho_{p,t-1}^c} - \frac{\eta_{p,t}}{\Omega \log_2 \left(1 + \frac{W_{p,f,t}^{bey}}{N_t^{bey}} \right)} \right), \quad (35)$$

where, $n'_{p,f,t}$ is an integer. So, the maximum value of $n'_{p,f,t}$ is the greatest integer value of the right-hand side expression. Hence, proved. \square

To solve the formulated problem, we have proposed Utility Maximization Patient Monitoring (UMPM) algorithm. The main idea behind the proposed heuristic is, to begin with, an initial allocation and then, re-positioning the patients by swapping their positions in order to achieve higher utility, iteratively, as shown in Fig. 2. The elaboration of each sub-algorithm is described as following:

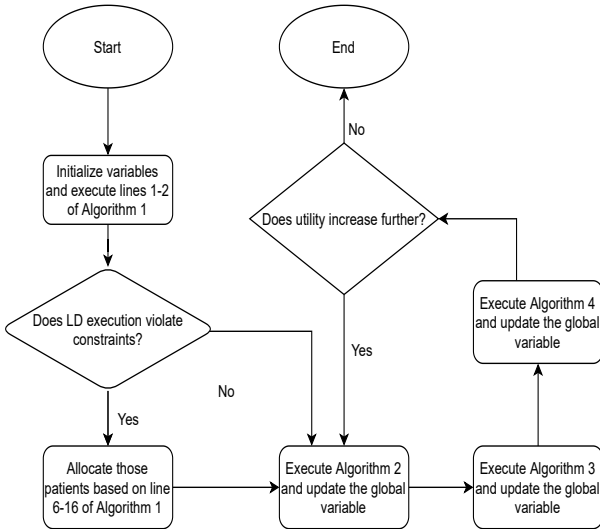


Fig. 2: Flow chart of the proposed heuristic.

5.1 UMPM Algorithm

Keeping the allocation of other patients as it is, we define U_{diff}^4 as the difference between utility before and after a

4. The profit is changed only for the patient p and the patient p affects the cost function with a value of the difference of p 's local computation time and its transmission and fog computation time. The computation time of the remaining patients using f also changes.

patient p is allocated an FS f as follows:

$$U_{diff} = \lambda_1(m - l - g\beta_{p,t}) + \lambda_2 \left(\rho_{p,t-1}^c \left(\frac{\beta_{p,t}}{\Upsilon} - \left(\frac{\beta_{p,t}(n_f + 1)}{\Gamma} + \frac{\eta_{p,t}}{BR_{p,f,t}} \right) \right) \right) - \lambda_2 \sum_{p' \in P^f} \frac{\rho_{p',t-1}^c \beta_{p',t}}{\Gamma}. \quad (36)$$

The set of patients that utilize FS f can be calculated as:

$$\mathbb{P}^f = \{p \in \mathbb{P} : q_{p,t} = 1\}. \quad (37)$$

Moreover, the number of patients utilizing FS f can be estimated as:

$$n^f = \sum_{p=1}^P h_{p,t}^f. \quad (38)$$

Algorithm 1: UMPM Algorithm

Input: $\Gamma, \Upsilon, g, m, l, \delta, \mathbb{H}_t = \{0\}, \Omega, N_t^{bey}; \forall p \in \mathbb{P}$:

$\rho_{p,t-1}^c, \beta_{p,t}, \eta_{p,t}; \forall p \in \mathbb{P}, \forall f \in \mathbb{F}: W_{p,f,t}^{bey};$

$\forall f \in \mathbb{F} : \mathbb{P}^f, n^f = 0;$

$temp_p, temp_f, U_{diff}^{max} = -\infty$

Output: Allocation Strategy (\mathbb{H}_t) .

1 Calculate $n_{p,f,t}^{max}$ for all the patients and FS using Theorem 1;

2 Calculate local computation time for the patients using Eq. (5);

3 **for** $p \leftarrow 1$ **to** P **do**

4 **if** $T_{p,t}^{c,l} > \frac{\delta}{\rho_{p,t-1}^c}$ **then**

5 insert p into \mathbb{P}^v

6 Sort patients in set \mathbb{P}^v in the order of decreasing criticalities;

7 **for** $p \in \mathbb{P}^v$ **do**

8 **for** $f \in F$ **do**

9 **if** $n_f \geq \min_{p' \in \mathbb{P}^f \cup \{p\}} n_{p',f,t}^{max}$ **then**

10 **continue**

11 Calculate U_{diff} as per Eq. (36);

12 **if** $U_{diff} > U_{diff}^{max}$ **then**

13 $U_{diff} \leftarrow U_{diff}^{max};$

14 $temp_p \leftarrow p;$

15 $temp_f \leftarrow f;$

16 Allocate $temp_p$ to $temp_f$ and update variables accordingly;

17 Run Algorithm 2;

18 Run Algorithm 3;

19 $\mathbb{P}^{rem} \leftarrow \mathbb{P} - \mathbb{P}^v$; // Update \mathbb{P}^{rem} as the set of patients who have not yet assigned an FS

20 Run Algorithm 4;

21 **repeat**

22 Run Algorithm 2;

23 Run Algorithm 3;

24 Run Algorithm 4;

25 Update \mathbb{P}^{rem} ;

26 **until** Utility does not increase;

The UMPM algorithm begins by calculating the constraint parameter $n_{p,f,t}^{max}$ as defined in Eq. (32). Then, it selects a subset of patients that violates the latency constraint if their data is computed on LD. Further, the algorithm sorts the patients in \mathbb{P}^v in the order of their decreasing criticalities. By doing so, the patients with higher criticalities are given priority by the algorithm, and thus, the algorithm will be able to accommodate all the patients in \mathbb{P}^v on FS. After that, re-allocation is done using Algorithms 2 and 3. After the initial allocation, the algorithm constructs the set \mathbb{P}^{rem} , which is the set of patients who have not been allocated any FS. It calls Algorithm 4 to allocate more patients on FS. It then calls Algorithms 2 and 3 which re-position the patients on the FSs. After that, Algorithm 4 is further called to allocate more patients on FS so as to increase utility. This process is repeated until there is no such possibility of increment in the utility (Fig. 2). Moreover, we have considered \mathbb{H}_t as a global matrix available at cloud server and accessed by all the proposed Algorithms 1-4.

The reason for having Algorithms 2 and 3 is to obtain better utility by swapping the association between FSs and patients as described below:

5.2 Two Way Swap based Algorithm

The utility difference due to two way swap, J_{diff}^{tr} ⁵ can be calculated as:

$$J_{diff}^{tr} = \frac{\rho_{p,t-1}^c \eta_{p,t}}{BR_{p,f,t}} - \frac{\rho_{p,t-1}^c \eta_{p,t}}{BR_{p,f',t}} + \frac{\rho_{p',t-1}^c \eta_{p',t}}{BR_{p',f',t}} - \frac{\rho_{p',t-1}^c \eta_{p',t}}{BR_{p',f,t}} + \frac{\rho_{p,t-1}^c \beta_{p,t} n_f}{\Gamma} - \frac{\rho_{p,t-1}^c \beta_{p,t} n_{f'}}{\Gamma} + \frac{\rho_{p',t-1}^c \beta_{p',t} n_{f'}}{\Gamma} - \frac{\rho_{p',t-1}^c \beta_{p',t} n_f}{\Gamma}. \quad (39)$$

The algorithm at each iteration picks a pair of patients already allocated to different FSs (lines 2-7). It then checks whether swapping the position of these two patients can increase utility or not (line 10). If the utility can be increased by satisfying the constraints (line 8), the patients are swapped (lines 11-12). It repeats the process until there is no such pair of patients (line 13). The convergence proof can be found in Section 6.

5. When two patients p and p' utilizing servers f and f' respectively are swapped, then the change in utility is caused by the difference of their transmission latencies and their computation latencies, as considered in Eq. (39).

Algorithm 2: Two way swap

Input: Globally accessible \mathbb{H}_t , Information of all patients (as per Algorithm 1) and FSs.

Output: \mathbb{H}_t

```

1 repeat
2   for  $f \leftarrow 1$  to  $F$  do
3     for every  $p \in \mathbb{P}^f$  do
4       for  $f' \leftarrow 1$  to  $F$  do
5         if  $f' == f$  then
6           continue
7         for every  $p' \in \mathbb{P}^{f'}$  do
8           if  $n_{f'} > n_{p,f',t}^{max}$  or  $n_f > n_{p',f,t}^{max}$  then
9             continue
10          if  $J_{diff}^{tr} > 0$  then
11            Swap  $p$  and  $p'$ ;
12            Update all the values accordingly;
13            Go to the Repeat loop;
14 until No swap increases utility;
```

5.3 One Way Swap based Algorithm

The utility difference due to one way swap, J_{diff} ⁶ is calculated as follows:

$$J_{diff} = \frac{\rho_{p,t-1}^c \eta_{p,t}}{BR_{p,f,t}} - \frac{\rho_{p,t-1}^c \eta_{p,t}}{BR_{p,f',t}} + \frac{\rho_{p,t-1}^c \beta_{p,t} (n_f - n_{f'} - 1)}{\Gamma} + \sum_{p' \in \mathbb{P}^f \setminus \{p\}} \frac{\rho_{p',t-1}^c \beta_{p',t}}{\Gamma} - \sum_{p' \in \mathbb{P}^{f'}} \frac{\rho_{p',t-1}^c \beta_{p',t}}{\Gamma}. \quad (40)$$

Algorithm 3: One Way Swap

Input: Globally accessible \mathbb{H}_t , Information of all patients (as in Algorithm 1) and FS.

Output: \mathbb{H}_t

```

1 repeat
2   for  $f \leftarrow 1$  to  $F$  do
3     for every  $p \in \mathbb{P}^f$  do
4       for  $f' \leftarrow 1$  to  $F$  do
5         if  $f' == f$  then
6           continue
7         Compute  $J_{diff}$  according to Eq. (40);
8         if  $J_{diff} > 0$  and
9            $n_{f'} + 1 \leq \min_{p' \in \mathbb{P}^{f'} \cup \{p\}} (n_{p',f',t}^{max})$  then
10          Add  $p$  to  $\mathbb{P}^{f'}$  and remove  $p$  from  $\mathbb{P}^f$ ;
11          Update the values correspondingly;
12          Go to the Repeat loop;
13 until No swap increases utility;
```

The algorithm at each iteration picks a patient allocated to an FS and checks if assigning that patient a different

6. When a patient p utilizing FS f is allocated FS f' , we can observe that profit does not change. The change in the cost function can be calculated as the difference between the transmission times when p uses f and f' . Also, the computation time of the patient p changes. Other than that, the computation time of the patients utilizing f and f' , other than p changes. All these changes are considered in J_{diff} .

FS can increase utility or not, satisfying the constraints (lines 2-8). If the utility can be increased by satisfying the constraints, the patients are assigned to a different FS (lines 9-10). Similar to the two-way swapping, this algorithm also repeats the process until there is no such pair of patients (line 11). The convergence proof can be found in Section 6.

5.4 Patient-FS Allocation Algorithm

The algorithm considers all patient-FS pairs at each iteration and selects the one that increases the utility by maximum value. In this way, the algorithm selects a subset of the patients given as input and allocates them to the FSs satisfying the constraints and maximising the utility value. Algorithm terminates when there is no improvement in the utility value compared to utility obtained in previous iteration. The following Lemma 2 establishes an iterative utility correlation across different iteration of Algorithm 4.

Lemma 2. Let $U_{diff,i}^{max}$ be the U_{diff}^{max} calculated by the algorithm at i^{th} iteration, then $U_{diff,i}^{max} \geq U_{diff,i+1}^{max}$. In other words, the maximum utility difference decreases with each iteration of Algorithm 4.

Proof. Let p be the patient assigned to FS f at i^{th} iteration and p' be the patient assigned to FS f' at $(i+1)^{th}$ iteration. Then, consider two following cases:

Case 1: $f = f'$

$$U_{diff,i}^{max} = \lambda_1(m - l - g\beta_{p,t}) + \lambda_2 \left(\rho_{p,t-1}^c \left(\frac{\beta_{p,t}}{\Upsilon} - \left(\frac{\beta_{p,t}(n_f + 1)}{\Gamma} + \frac{\eta_{p,t}}{BR_{p,f,t}} \right) \right) \right) - \lambda_2 \sum_{p'' \in P^f} \frac{\rho_{p'',t-1}^c \beta_{p'',t}}{\Gamma}. \quad (41)$$

Also,

$$U_{diff,i+1}^{max} = \lambda_1(m - l - g\beta_{p',t}) + \lambda_2 \left(\rho_{p',t-1}^c \left(\frac{\beta_{p',t}}{\Upsilon} - \left(\frac{\beta_{p',t}(n_f + 2)}{\Gamma} + \frac{\eta_{p',t}}{BR_{p',f,t}} \right) \right) \right) - \lambda_2 \sum_{p'' \in P^f \cup p} \frac{\rho_{p'',t-1}^c \beta_{p'',t}}{\Gamma}, \quad (42)$$

where n^f are the number of patients utilizing f before i^{th} iteration and similarly P^f is the set of such patients. On subtracting Eq. (41) from Eq. (42), it is clear that,

$$U_{diff,i+1}^{max} \leq U_{diff,i}^{max}. \quad (43)$$

Case 2: $f \neq f'$

In this case, as both the FSs are different, if $U_{diff,i+1}^{max}$ would have been greater than $U_{diff,i}^{max}$, then the algorithm would have picked p' at the $(i+1)^{th}$ iteration only, but that is not the case. Hence $U_{diff,i+1}^{max} \leq U_{diff,i}^{max}$.

Combining both the cases, we conclude, $U_{diff,i+1}^{max} \leq U_{diff,i}^{max}$. Hence, proved. \square

Algorithm 4: Patient-FS Allocation

Input: Globally accessible \mathbb{H}_t , Set of Patients, P^{rem} , $flag = 0$, $temp_p$, $temp_f$, $U_{diff}^{max} = 0$

Output: Allocation Strategy (\mathbb{H}_t).

```

1 repeat  $|P^{rem}|$  times
2   for every  $p \in P^{rem}$  do
3     if  $q_{p,t} == 1$  then
4       continue;
5     for  $f \leftarrow 1$  to  $F$  do
6       if  $n_f \geq \min_{p' \in P^f \cup \{p\}} n_{p',f,t}^{max}$  then
7         continue
8       Calculate  $U_{diff}$  as per Eq. (36);
9       if  $U_{diff} > U_{diff}^{max}$  then
10         $flag \leftarrow 1$ ;
11         $U_{diff} \leftarrow U_{diff}^{max}$ ;
12         $temp_p \leftarrow p$ ;
13         $temp_f \leftarrow f$ ;
14   if  $flag == 0$  then
15     break
16   Assign patient  $temp_p$  to FS  $temp_f$ ;
17   Update  $n^{temp_f}$  and  $P^{temp_f}$ ;
18   Update  $\mathbb{H}_t$ ;

```

5.5 Illustration of UMPMA

Let there be 3 FSs and 10 patients. We illustrate the proposed heuristic using a randomly generated example under the simulation parameters as mentioned in Table 3. The yellow and blue colours represent patients and FSs respectively, in subsequent figures.

Let Fig. 3 (a) shows an allocation of patients to FSs after execution of lines 1-18 (including Algorithms 2-3). We can see that P8 and P9 are allocated to F1 and F2 respectively. Moreover, Algorithm 4 executes to see any possible update

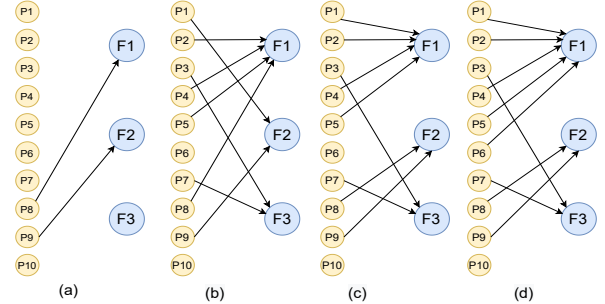


Fig. 3: (a) Initial allocation. (b) Updated allocation after execution of Algorithm 4. (c) Updated allocation in next iteration. (d) Final allocation.

in the allocation for improving utility value. In this algorithm, an allocation is found out while maximizing overall utility value. Thus, we get an outcome as shown in Fig. 3 (b). Next, after re-iterating the given swapping algorithms, Fig. 3 (c) is obtained as an outcome improving the utility over the allocation given in Fig. 3 (b). Furthermore, after convergence of the proposed heuristic, we achieve Fig. 3 (d) as a final outcome maximizing the overall utility over all previous allocations.

6 ANALYSIS OF PROPOSED HEURISTIC

This section discusses convergence and time complexity of the proposed heuristic as follows.

Lemma 3. *The proposed UMPMA converges.*

Proof. The convergence of the UMPMA relies on the convergence of one-way and two-way swap algorithms. Both the swapping algorithms swaps only if utility increases and if there is no such swap possible, their execution is terminated. As the total possible combinations (patient-FS) are finite, hence the utility will also be a finite value. Thus, both the swapping mechanisms converge. Algorithm 4 converges because the number of iterations are bounded by a finite value i.e., the number of remaining patients. Further, the proposed heuristic repeatedly executes two-way swap, one-way swap and Patient-FS allocation schemes. Every iteration converges and the algorithm goes to the next iteration only when utility increases. A similar argument regarding the finiteness of the utility concludes the proof. \square

Theorem 2. *The time complexity of UMPMA is $O(P^2F)$.*

Proof. The time complexity of the UMPMA depends on the complexity of the three sub-algorithms it calls. The time complexity of Algorithm 1 from lines (1-16) is $O(PF)$. Through amortized analysis, we can see that Algorithm 2 considers $O(P^2)$ pairs of patients. The algorithm repeats until it converges. Thus, the required number of iterations is bounded by a finite value. Thus, the time complexity of Algorithm 2 is $O(P^2)$. Similarly, through amortized analysis, the time complexity of Algorithm 3 is $O(PF)$ as it considers $O(PF)$ number of possibilities of swapping. In Algorithm 4, the number of iterations are bounded by $O(P)$. In every iteration, $O(PF)$ pairs are considered. Thus, the time complexity of Algorithm 4 is $O(P^2F)$. Hence, the time complexity of proposed UMPMA is $O(\max\{P^2F, P^2, PF\})$, i.e., $O(P^2F)$. \square

Performance study of the proposed heuristic is given in the following section over different simulation parameters.

7 PERFORMANCE STUDY

The simulation setup and the parameters are given in the following:

7.1 Simulation Setup

To simulate the proposed health monitoring system, we have considered different number of patients and FSs in simulation environment. The size of the patient's data is randomly considered to be within 1 to 3 MB and the CPU cycles required are randomly considered to be between 100 to 1000 Megacycles [9]. The value of δ is considered to be 250 ms. The bandwidth of a channel is taken as 5 MHz. Transmission power and noise are considered to be 0.1 Watts and -100 dBm, respectively as shown in Table 3. Distance between any patient and FS is randomly taken as between 50 to 100 m. Channel gain is considered to be as $(\text{dist}_{p,f})^{-3}$ [22], where $\text{dist}_{p,f}$ is the distance between the LD and the FS. The computation capacity of the FS is taken as 22.4 GHz and that of LD is taken as 2.4 GHz. The patient criticalities are considered between 0 to 1 and are randomly

TABLE 3: Simulation Parameters

Parameter	Value
Number of Patients (P)	20-1000
Number of FS (F)	2-200
Data Size ($\eta_{p,t}$)	[1,3] MB
CPU Cycles ($\beta_{p,t}$)	[100,1000] Megacycles
Time Constraint (δ)	250 ms
Patient Criticality ($\rho_{p,t-1}^c$)	[0,1]
Local Computation Price (l)	100 units
Fog Computation Price (m)	200 units
FS Charge per CPU Cycles (g)	0.1 units
Fixed Charge per FS (k)	0 units
Distance between LD and FS	[50,100] m
Path loss factor	3
Channel Bandwidth (Ω)	5 MHz
Noise (N_t^{bey})	-100 dBm
Transmission power ($P_{p,t}^{\text{bey}}$)	0.1 W
Fog Computation Capacity (Γ)	22.4 GHz
Local Computation Capacity (Υ)	2.4 GHz
Patients' Weight (λ_1)	1
Medical Center's Weight (λ_2)	1

generated. Although patient criticalities can be anything between 0 to ∞ , it can be normalized between 0 to 1 for each patient. We are considering various parameters to evaluate the performance of our proposed algorithm. Python 3.9.0 platform is used to model the above simulation setup and execution of proposed heuristic.

IBM ILOG CPLEX Optimization Studio has been utilized to obtain optimal solutions for the comparison [23]. The simulations were performed on a personal computer with processor Intel(R) Core(TM) i5-8250U CPU @ 1.60ghz.

Moreover, we propose another scheme, named *Base*, to compare it with our proposed heuristic. Base scheme first considers patients that violates the constraint, sort them in decreasing order of their criticalities. Then, it allocates the patient one-by-one to an FS that results in increasing the utility by the maximum amount. It then considers, the remaining patients and order them by criticalities in decreasing order, and then repeats the same process of allocation.

7.2 Simulation Results

In this section, the results are presented on various aspects such as mentioned below:

7.2.1 System Utility

In Fig. 4, we have considered three cases, when the number of patients are 20, 40 and 60 and the number of FSs varies from 2 to 12. A general trend can be seen among all three cases. The proposed heuristic performs better than the Base scheme. The utility obtained by the proposed heuristic is 96% of the optimal value compared to 56% of that of Base scheme on an average. The reason is that the Base scheme allocates the patients in a particular order. Although the Base scheme is based on patient criticality which is an important factor for the system, however, it does not consider the data size and the CPU cycles of data packets. The proposed heuristic considers all the above factors to reach a sub-optimal utility within polynomial time complexity.

7.2.2 Patient Cost

In this section, we present simulation results for the patients' cost. From Fig. 5, it can be observed that the proposed

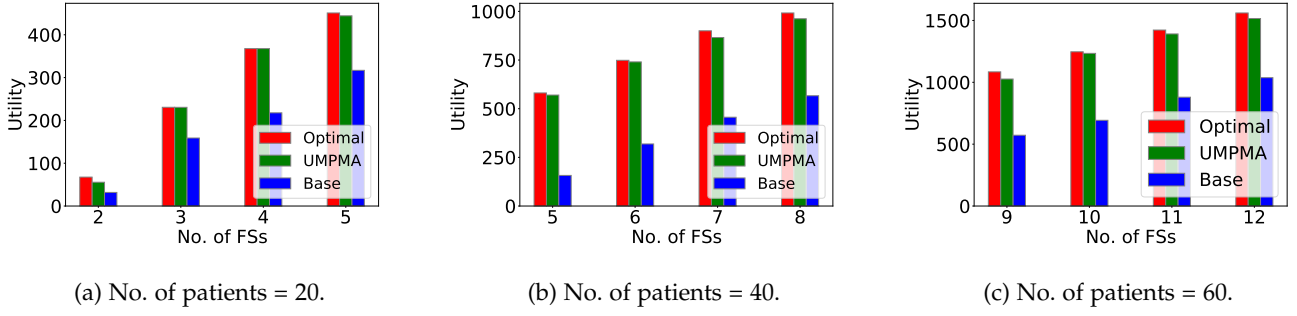


Fig. 4: Utility comparison among different schemes.

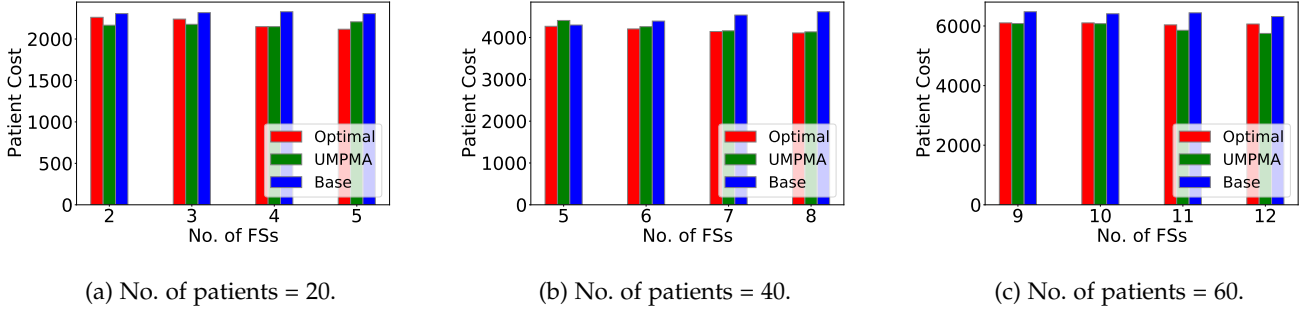


Fig. 5: Patients' cost comparison among different schemes.

heuristic generally results in lower patients' cost than the Base scheme. The reason is that the proposed heuristic considers different parameters and it allocates and re-allocates patients to maximise utility, resulting in lower patients' cost. However, the Base scheme does not re-allocate patients, resulting in higher patients' cost and lower utility. It can be seen that sometimes the patients' cost obtained by the proposed heuristic is lower than that obtained by the optimal solution. It is because optimal solution considers all possibilities and it could be a possibility that profit becomes a dominating factor in the optimal solution. However, our algorithm is more criticality-aware, thus it tries to increase the utility by lowering the patients' cost.

7.2.3 Convergence Analysis

Fig. 6a depicts the convergence of the algorithm for three different cases. From the result, we can observe that when $P = 60$, the algorithm converges in 6 iterations, and in 2 and 4 iterations for $P = 20$ and $P = 40$ respectively, on an average. As considered in the Section 6, the number of iterations are quite small and can be taken as a constant.

7.2.4 Trade-off between utility and criticality

We analyze the trade-off between utility and criticality through simulation. We have considered nine different median criticalities from 0.1 to 0.9. We have considered a fixed order of patients and then, assigned a median criticality to the middle patient. We randomly assign a criticality lower and higher than the median to the patients before and after the middle patient, respectively. In Fig. 6b, it can be observed that as the median criticality increases, the system utility decreases. However, the relative trends remain the same and our proposed algorithm achieves a system utility

of 97% of the optimal whereas the Base scheme could only achieve 83% of the optimal on an average.

7.2.5 Utility comparison for dense networks

We consider a large number of patients and FSs to analyse the proposed algorithm. However, due to the large number of patients and FSs, it is difficult to obtain an optimal solution using IBM ILOG CPLEX tool [23]. Thus, we compare the results with the Base scheme only. In Fig. 6c, we can see that the proposed algorithm gives better result than the base scheme. For a small number of patients and FSs, the utility gap between the proposed heuristic and the Base scheme is less, but as the number of patients and FSs increases, the gap also increases. Thus, the proposed heuristic is quite beneficial for dense networks. The Base scheme could only achieve a utility of 78% of the utility achieved by the proposed heuristic.

7.2.6 Execution time comparison

In this section, we compare the execution time of three schemes. For Optimal, we consider the execution time as the time required by the ILOG CPLEX tool to reach the utility achieved by the proposed heuristic. All three schemes were run on the same machine and the execution time is taken as the average of execution times obtained by extensively executing these schemes. The proposed heuristic and the Base scheme complete their execution in a few milliseconds, thus, their plot is overlapping each other. However, the optimal takes some seconds to reach the utility achieved by the proposed heuristic. Therefore, we can say that the proposed heuristic achieves a sub-optimal utility in a small time compared to the optimal and its execution time is comparable to the Base scheme for smaller inputs. For more

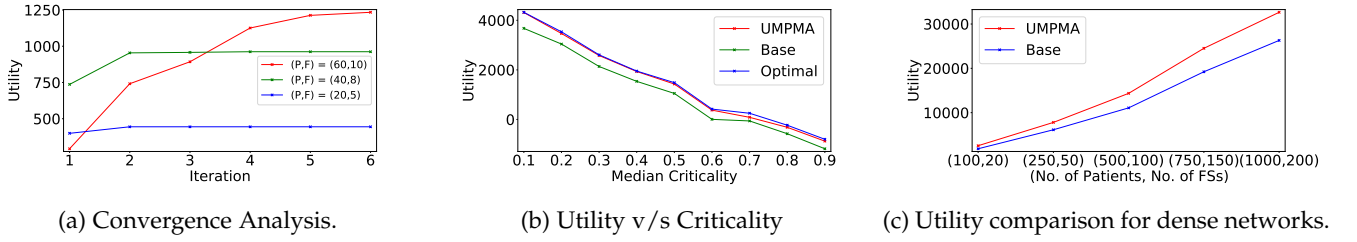


Fig. 6: Utility comparison against different parameters.

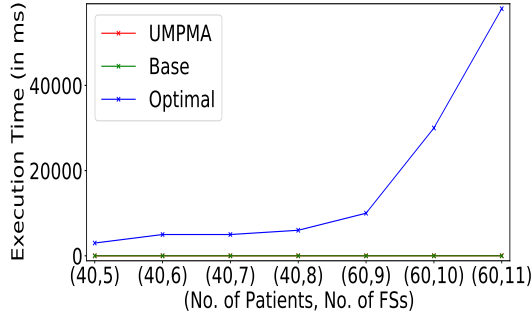


Fig. 7: Execution time comparison

dense networks, the Base scheme is quite fast, but, the utility achieved is quite low. Thus, UMPMA emerges as a better scheme, balancing both the execution time and system utility.

8 CONCLUSION

In this paper, we have designed a beyond-WBAN based fog computing system for remote health monitoring. The main contribution was in the beyond-WBAN, where we formulated a problem based on the profit of medical center and the loss of patients, measured in terms of latency and criticalities. We then proposed a criticality-aware utility maximization heuristic (i.e., UMPMA) to maximize the utility in beyond-WBAN. The proposed heuristic is based on the swapping mechanism. Simulation results and evaluation of the UMPMA were presented to show the effectiveness of the proposed heuristic on various parameters. The proposed algorithm was demonstrated as criticality-aware, thus serving the purpose of the system. Through extensive simulations, we show that the proposed heuristic achieves an average utility of 96% of the optimal, in polynomial time complexity.

This study leads to some future directions. The interference can be included in the model and thus, a sub-channel allocation problem can be considered with the current utility maximization problem. The role of doctors can be directly included in the system based on the criticality of the patients and it gives rise to a different pricing model as doctors are directly involved. Moreover, energy consumption is an important factor along-with the latency. Thus, a future study can be in the direction of energy-awareness together with the criticality-awareness.

Acknowledgments: This work is partially supported by the Science and Engineering Research Board (SERB), Government of India under Grant SRG/2020/000318.

REFERENCES

- [1] F. Wu, C. Qiu, T. Wu, and M. R. Yuce, "Edge-Based Hybrid System Implementation for Long-Range Safety and Healthcare IoT Applications," *IEEE Internet of Things Journal*, pp. 1–1, 2021.
- [2] S. P. Dash, "The impact of iot in healthcare: Global technological change & the roadmap to a networked architecture in india," *Journal of the Indian Institute of Science*, pp. 1–13, 2020.
- [3] A. Pratap and S. K. Das, "Stable matching based resource allocation for service provider's revenue maximization in 5g networks," *IEEE Transactions on Mobile Computing*, pp. 1–1, mar 5555.
- [4] P. K. Bishoyi and S. Misra, "Enabling Green Mobile Edge Computing for 5G-Based Healthcare Applications," *IEEE Transactions on Green Communications and Networking*, 2021.
- [5] A. A. Mutlag, M. K. Abd Ghani, N. a. Arunkumar, M. A. Mohammed, and O. Mohd, "Enabling technologies for fog computing in healthcare iot systems," *Future Generation Computer Systems*, vol. 90, pp. 62–78, 2019.
- [6] W. Yu, F. Liang, X. He, W. G. Hatcher, C. Lu, J. Lin, and X. Yang, "A survey on the edge computing for the internet of things," *IEEE access*, vol. 6, pp. 6900–6919, 2017.
- [7] S. Yi, C. Li, and Q. Li, "A survey of fog computing: concepts, applications and issues," in *Proceedings of the 2015 workshop on mobile big data*, pp. 37–42, 2015.
- [8] P. Deshingkar, *Migration, remote rural areas and chronic poverty in India*. ODI, 2010.
- [9] Z. Ning, P. Dong, X. Wang, X. Hu, L. Guo, B. Hu, Y. Guo, T. Qiu, and R. Y. Kwok, "Mobile edge computing enabled 5g health monitoring for internet of medical things: A decentralized game theoretic approach," *IEEE Journal on Selected Areas in Communications*, 2020.
- [10] L. Feng, A. Ali, M. Iqbal, A. K. Bashir, S. A. Hussain, and S. Pack, "Optimal haptic communications over nanonetworks for e-health systems," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 5, pp. 3016–3027, 2019.
- [11] L. Gu, D. Zeng, S. Guo, A. Barnawi, and Y. Xiang, "Cost efficient resource management in fog computing supported medical cyber-physical system," *IEEE Transactions on Emerging Topics in Computing*, vol. 5, no. 1, pp. 108–119, 2015.
- [12] H. K. Apat, K. Bhaisare, B. Sahoo, and P. Maiti, "Energy efficient resource management in fog computing supported medical cyber-physical system," in *2020 International Conference on Computer Science, Engineering and Applications (ICCSEA)*, pp. 1–6, IEEE, 2020.
- [13] Y. Qiu, H. Zhang, and K. Long, "Computation offloading and wireless resource management for healthcare monitoring in fog-computing based internet of medical things," *IEEE Internet of Things Journal*, 2021.
- [14] C. Yi and J. Cai, "Transmission management of delay-sensitive medical packets in beyond wireless body area networks: A queueing game approach," *IEEE Transactions on Mobile Computing*, vol. 17, no. 9, pp. 2209–2222, 2018.
- [15] S. Misra and S. Sarkar, "Priority-based time-slot allocation in wireless body area networks during medical emergency situations: An evolutionary game-theoretic perspective," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 2, pp. 541–548, 2014.
- [16] S. Misra, S. Moulik, and H.-C. Chao, "A cooperative bargaining solution for priority-based data-rate tuning in a wireless body area network," *IEEE transactions on wireless communications*, vol. 14, no. 5, pp. 2769–2777, 2015.
- [17] B. Di, L. Song, and Y. Li, "Sub-channel assignment, power allocation, and user scheduling for non-orthogonal multiple access networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 11, pp. 7686–7698, 2016.

- [18] A. Pratap, R. Gupta, V. S. S. Nadendla, and S. K. Das, "On maximizing task throughput in iot-enabled 5g networks under latency and bandwidth constraints," in *2019 IEEE International Conference on Smart Computing (SMARTCOMP)*, pp. 217–224, IEEE, 2019.
- [19] A. Pratap, R. Gupta, V. S. S. Nadendla, and S. K. Das, "Bandwidth-constrained task throughput maximization in iot-enabled 5g networks," *Pervasive and Mobile Computing*, vol. 69, p. 101281, 2020.
- [20] T. Tanwar, U. D. Kumar, and N. Mustafee, "Optimal package pricing in healthcare services," *Journal of the Operational Research Society*, vol. 71, no. 11, pp. 1860–1872, 2020.
- [21] H. R. Lewis, "Computers and intractability. a guide to the theory of np-completeness," 1983.
- [22] T. S. Rappaport *et al.*, *Wireless communications: principles and practice*, vol. 2. prentice hall PTR New Jersey, 1996.
- [23] "CPLEX Optimizer." <https://www.ibm.com/analytics/cplex-optimizer>. Accessed: 2021-05-01.



Ajay Pratap is an Assistant Professor with the Department of Computer Science and Engineering, Indian Institute of Technology (BHU) Varanasi, India. Before joining IIT (BHU), he was associated with the Department of Computer Science and Engineering, National Institute of Technology Karnataka (NITK) Surathkal, India, as an Assistant Professor from December 2019 to May 2020. He worked as a Postdoctoral Researcher in the Department of Computer Science at Missouri University of Science and Technology, USA, from August 2018 to December 2019. He completed his Ph.D. degree in Computer Science and Engineering from the Indian Institute of Technology Patna, India, in July 2018. His research interests include Cyber-Physical Systems, IoT-enabled Smart Environments, Mobile Computing and Networking, Statistical Learning, Algorithm Design for Next-generation Advanced Wireless Networks, Applied Graph Theory, and Game Theory. His current work is related to HetNet, Small Cells, Fog Computing, IoT, and D2D communication underlaying cellular 5G and beyond. His papers appeared in several international journals and conferences including IEEE Transactions on Mobile Computing, IEEE Transactions on Parallel and Distributed Systems, and IEEE LCN, etc. He has received several awards including the Best Paper Candidate Award and NSF travel grant for IEEE Smartcom'19 in the USA.



Moirangthem Biken Singh completed the B.Tech degree in Computer Science and Engineering from the National Institute of Technology Manipur, India, in 2018 and the M.Tech degree from National Institute of Technology Kurukshetra, India, in 2021. He is currently pursuing Ph.D. degree in Computer Science and Engineering, Indian Institute of Technology (BHU) Varanasi, India. His current research interest include AI, machine learning and FL in Smart Healthcare.



Navneet Taunk is an undergraduate student with the Department of Computer Science and Engineering, Indian Institute of Technology (BHU) Varanasi, India. His research interests include Design of Algorithms, Internet of Things (IoT) and Mathematical Modelling.



Naveen Kumar Mall is an undergraduate student with the Department of Computer Science and Engineering at the Indian Institute of Technology (BHU) Varanasi, India. His research interests include Deep learning, Machine learning and Internet of Things (IoT).