A Dynamic Cost-Efficient Task Offloading Framework for Resource-constrained Edge-based Smart Healthcare Systems

1st Subhranshu Sekhar Tripathy
School of Computer Engineering
KIIT Deemed to be University
Bhubaneswar, 751024, India
subhranshu.008@gmail.com

2nd Sujit Bebortta
Department of Computer Science
Ravenshaw University
Cuttack, 753003, India
sujitbebortta1@gmail.com

3rd Aishwarya Nayak
Department of Computer Science & Engineering
DRIEMS University
Cuttack, 754022, India
aishwaryanayak22@gmail.com

4th Jnana Ranjan Behera
School of Electronics
KIIT Deemed to be University
Bhubaneswar, 751024, India
jnanaranjan07@gmail.com

***Abstract*—The advancements in Internet of Healthcare Things (IoHT) has significantly transformed the healthcare industry. To meet the computing needs of the healthcare industry, the Multi-access Edge Computing (MEC) serves a favorable solution towards processing of offloaded computational workloads from resource constrained IoHT devices. As the offloaded data to MEC servers increases in volume, it becomes difficult for the resource constrained servers to process these tasks for advanced health analytics. In this view, the present work proposes a dynamic cost-aware solution for handling the computationally intensive healthcare data packets offloaded by resource constrained IoHT applications. Here, an adaptive offloading technique is put forth that makes use of the $M/M/c$ queuing model and models the system as a non-birth death stochastic process. The effectiveness of the suggested CARE framework is investigated in conjunction with benchmark computing platforms like MEC and the conventional remote Cloud computing platform under various sensitivity criteria, like the ideal workload, the response time, and the system processing cost. Our simulation results showed that the suggested framework outperformed the other two approaches, making it appropriate for time-sensitive IoHT applications.**

***Index Terms*—Task Offloading, Internet of Healthcare Things, Response Time Minimization, Cost Optimization.**

I. Introduction

The Quality of Service (QoS) for Internet of Things (IoT)-enabled applications has significantly improved with the introduction of fifth generation (5G) connectivity infrastructures [1], [2]. The healthcare sector has greatly benefited from the advancements in IoT and Information Communication Technologies (ICT) for facilitating processing of time-critical tasks [3]. Wireless Body Area Networks (WBANs) have been instrumental in the healthcare sector for capturing vital health parameters using wearable sensory devices [4]. However, as a result of heterogeneous multimedia data, more complicated and computationally intensive tasks have emerged evidencing the rise of Internet of Healthcare Things (IoHT) paradigms [1]. Multi-access Edge Computing (MEC) has been proven to be a promising resource rich computing platform for supporting higher system throughput and low latency computation of healthcare data on the edge computing servers, as demonstrated by the studies made in references [1], [4].

The management of energy consumption and resource allocation in the MEC environment continues to be a major difficulty in scenarios for rapidly scaling IoHT applications [1]. In order to do this, the deployment of offloading techniques is crucial in managing the MEC platforms' constrained performance. By addressing the needs of a large user base, computation offloading in IoHT applications helps reduce the energy consumption by the MEC servers as a consequence of processing demanding workloads. Additionally, this controls the compute and storage needs on the MEC servers for processing healthcare data and improves the users' Quality of Experience (QoE). According to [1], it is clear that the number of applications connected to cloud computing platforms has increased, and the energy used by cloud servers now accounts for more than 2% of the overall electricity used in the United States. These developing technologies, including the IoHT, call for an effective offloading framework to provide low latency, energy-efficient computation, system throughput, and optimal resource allocation. These offloading choices, however, are frequently context-based, and depend on the objectives of the healthcare application that will be developed.

The effectiveness of task offloading strategies for IoT-based MEC systems has been reported in a number of recent research [1]. It was noted that these systems' performance was optimized in terms of latency, computing costs, task waiting times, and the typical quantity of data packets associated with the system. Additionally, task offloading for large-scale IoTdriven systems has been seen to reduce IoT device energy usage, and reduced the effects of carbon foot-printing. Since the healthcare sector deals mostly with the processing of time-sensitive tasks, the role of task offloading and latency minimization of IoHT systems deserves substantial attention [5]. However, very few works have discussed the task offloading mechanism associated with MEC-based IoHT systems in light of optimising server utilization, waiting times for healthcare tasks in the queue, and processing costs associated with the execution of healthcare data.

This work provides a cost-aware solution for computation offloading in resource-constrained IoHT applications by proposing the CARE framework. We model the task queuing mechanism of the packets produced by IoHT service layer as stochastic processes to address the computation offloading issues. The task arrival behavior at the MEC and Cloud servers was modelled as a non-birth death process using the $M/M/c$ queuing system. The aforementioned model was used to derive performance measurements for three operating modes, namely the MEC platform, the Cloud only platform, and the proposed CARE platform, in order to facilitate the processing of IoHT data. The heterogeneous $M/M/c$ queue's steady state solutions were first determined, and performance metrics for the system's workload, average number of data packets, average response time, and processing costs were obtained. It was suggested that the proposed operating platform can be used to enable adaptive offloading for processing time-sensitive IoHT tasks in light of the limited computational resources of MEC servers. Through the numerical simulations including various workload scenarios and task arrival rates, it was found that the proposed CARE platform performs significantly better than the MEC and Cloud-based platforms. The suggested CARE platform's ideal workload significantly outperformed the MEC platform and the Cloud platform, increasing by 15.873% and 63.492%, respectively. The CARE platform showed an improvement over MEC of 0.8107% and over Cloud platform of 64.4471% in terms of average response time for these platforms. Therefore, based on our observations, it can be concluded that the proposed CARE platform effectively manages the QoS requirements of the system by timely offloading computationally intensive tasks, which enables it to handle IoHT generated data.

II. Related Work

In this section, some cutting-edge works that aim to achieve computation offloading in IoHT systems have been studied. The application of machine learning models to several challenging problems in the healthcare sector has greatly benefited IoHT. Recently, it has been shown that machine learning models can be used to monitor serious clinical illnesses like Parkinson's disease, where they can be used to track the disease's progression and problems in patients [6]. These frameworks have aided in raising patients' quality of life (QoL). Machine learning classifiers like the Deep Neural Network (DNN) have shown to be effective in terms of handling signal complexity and extracting features from the recorded EEG signals in the context of recognising epileptic structures in patients with seizures. The human activity recognition (HAR) function of machine learning models for IoHT is vital for gathering user physiological data. According to work by [7], Convolutional Neural Networks (CNNs) have been proven effective for providing real-time prediction in HAR systems. In order to predict the data from wearable inertial sensors for categorizing diverse human activities, the Stochastic Gradient Descent (SGD) approach was used in convergence with CNN in order to reduce the loss function [7].

IoHT applications have profited significantly from the use of computer vision and machine learning algorithms for quickly identifying various patient postures, identifying the primary symptoms of falls, and identifying other signs of pain in indoor environments [8]. The use of an optimised multivariate linear regression model was observed to track abnormal blood serum parameters such as blood glucose, high density lipoprotein (HDL), low density lipoprotein (LDL), lamotrigine concentration, total cholesterol, and thyroid stimulating hormone in order to develop a recommendation system for patients with progressive diabetes [9].

The challenges of resource allocation for processing clinical information have emerged as a result of the adoption of IoHT and the development of 5G communication technologies in the healthcare industry. This has led to big data challenges in IoHT and has had negative effects on system performance and timely processing of vital healthcare data, as investigated by [10]. In particular for the healthcare industry, the implementation of efficient resource allocation algorithms for cloud edge computing platforms has helped minimize latency overheads and limit effects on system performance. Additionally, the combination of big data analytics and deep learning models like the deep belief network (DBN) has improved the convergence rate of the prediction model and helped handle extremely huge healthcare datasets [11].

Healthcare big data analytics have been quite helpful in providing individualized care for IoHT systems, particularly in addressing patients' clinical and assistive needs [12]. This aids in the development of effective medical decision support systems (m-DSS) for facilitating services like the real-time prediction of patients' underlying clinical conditions via data streams produced by IoHT apps, according to a study by [12].

III. System Model

The IoHT is a new field that enables real-time smart healthcare services for users. In order to capture the patients' essential health metrics and provide timely recommendations, a vast number of sensory devices and applications must work cooperatively. The processing of tasks generated by these systems has become difficult to execute locally at the users' end due to the recent growth seen in wearable technologies; as a result, the role of Cloud computing platforms has evolved in IoHT. The Cloud platform has some restrictions due to the remotely distributed servers in terms of latency, response time, transmission and processing costs that are peculiar to the healthcare sector.

According to this perspective, the MEC platform's function of enabling processing closer to the IoHT devices on the edge can be accomplished in order to give services to users on time. Although MEC satisfies the majority of the requirements for IoHT systems, due to limited resource availability, this platform experiences restrictions in resource allocation for processing computationally intensive jobs, which results in latency problems. Therefore, a framework that makes use of both the unrestricted Cloud platform resources and the MEC platform's computing efficiency is needed. In order to address the aforementioned difficulties, this work suggests CARE framework that enables adaptive offloading of computationally heavy operations from MEC servers to the closest Cloud server to enable the continued execution of healthcare tasks.

The proposed hybrid architecture's operating mechanism is shown in Fig. 1. After being routed through the proper gateways, it is shown that the healthcare tasks generated by various IoHT devices follows a Poisson arrival rate at the servers' task queue. The system also enables the use of MEC platform and the Cloud platform, two heterogeneous platforms for carrying out the execution of these tasks.

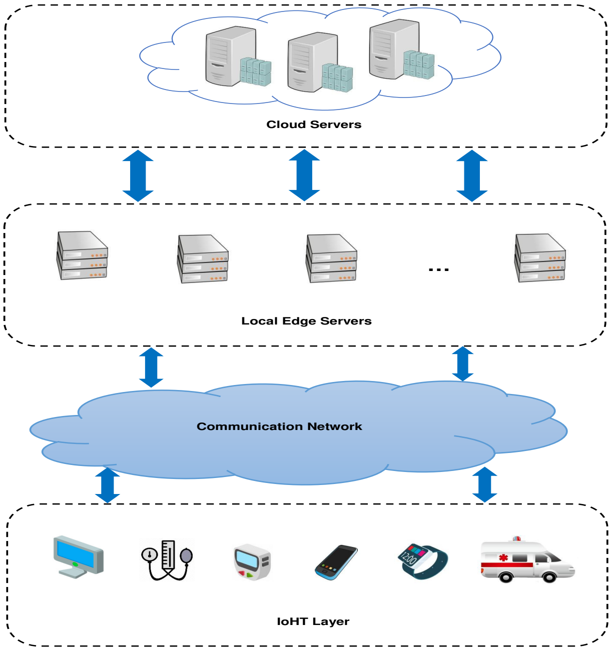

Fig. 1. An overview the basic IoHT mechanism in Cloud-Edge environment.

According to the offloading factor $beta$, the proposed CARE framework alternates control between the MEC servers and Cloud servers to enable the offloading policy and operate as a hybrid computing platform. This lowers the likelihood of service interruptions brought on by limited resource availability. The arrival pattern of the tasks generated by IoHT is defined by a Poisson process, which follows the $M/M/c$ queuing scheme. For the completion of each task, the MEC and Cloud platforms' service rates follow an exponential distribution. The MEC servers offload computationally intensive tasks to the closest Cloud server because they have a tendency to consume MEC's computing resources, which compromises the QoS requirements of IoHT systems.

IV. Problem Formulation

We leverage the $M/M/c$ stochastic queuing framework to realize the inherent behavior of IoHT systems. For MEC, Cloud, and the suggested CARE framework, we formulate the problem for predicting performance metrics in terms of the ideal workload, typical response time, and processing costs. The framework is depicted as a system of heterogeneous servers where an adaptive method is taken over MEC servers to offload computationally heavy activities to nearby cloud servers in order to maintain the IoHT system's QoS. The MEC system is regarded as a server with a limited amount of storage space that can process healthcare data packets produced by wearables and IoHT devices. Data packets arrive at the MEC server using a Poisson process, where the arrival rate of each packet is represented by λ. Data packets handled by a MEC server have a service rate of μ_1, whereas cloud servers have a service rate of μ_2.

The process has a probability of $p_{n,m}(t)$ of having n data packets at time t, despite the fact that it was first detected to have m. The suggested system architecture's differential-difference equations [13], are denoted by the following notation:

$$\frac{dP_{0,0}(t)}{dt} = -\lambda P_{0,0}(t) + \mu_1 P_{1,0}(t) + \mu_2 P_{0,1}(t), \tag{1}$$

$$\frac{dP_{1,0}(t)}{dt} = -(\lambda + \mu_1) P_{1,0}(t) + \mu_2 P_{1,1}(t) + \lambda P_{0,0}(t), \tag{2}$$

$$\frac{dP_{0,1}(t)}{dt} = -(\lambda + \mu_2) P_{0,1}(t) + \mu_1 P_{1,1}(t), \tag{3}$$

$$\frac{dp_{n,m}(t)}{dt} = -(\lambda + \mu_1 + \mu_2) p_{n,m}(t) + (\mu_1 + \mu_2)$$

$$\times p_{n+1,m}(t) + \lambda p_{n-1,m}(t), \;\; n > 0, \;\; m = 1. \tag{4}$$

Transient solutions are produced by solving the differential-difference equation system mentioned above. For the sake of simplicity, the study considers the steady-state solution of the proposed system by setting all time derivative terms $dP_{n,m}(t)/dt$ to zero [13]. The steady-state probability $p_{n,m} = \lim_{t\to\infty} P_{n,m}$. Now the prior system of differential-difference equations reduces to

$$\lambda p_{0,0} = \mu_1 p_{1,0} + \mu_2 p_{0,1}, \tag{5}$$

$$(\lambda + \mu_1) p_{1,0} = \mu_2 p_{1,1} + \lambda p_{0,0}, \tag{6}$$

$$(\lambda + \mu_2) p_{0,1}(t) = \mu_1 p_{1,1}(t), \tag{7}$$

$$(\lambda+\mu_1+\mu_2)p_{n,m}(t) = (\mu_1+\mu_2)p_{n+1,m}(t)+\lambda p_{n-1,m}(t), \;\; n > 0, \; m = 1. \tag{8}$$

Here, $\rho_1 = \lambda/\mu_1$, $\rho_2 = \lambda/\mu_2$ and ρ represents different work-load of MEC server, cloud server, and CARE platform respectively. This paper the service rate μ_1 of MEC server is faster than the cloud service rate μ_2 i.e., $\mu_1 = \beta\mu_2$, hence ρ is defined as

$$\rho = \frac{\lambda}{\mu_1+\mu_2} = \frac{\rho_2}{1+\beta} \tag{9}$$

Using the value of ρ, we obtain the following relationship as

$$p_{n,1} = \rho p_{n-1,1} = \rho^{n-1} p_{n-1,1}, \; n > 1. \tag{10}$$

Using the steady-state system of difference Eqs.(5) to (8), we result the following state probabilities:

$$p_{0,1} = \frac{\rho}{1+2\rho}\frac{\lambda}{\mu_2}p_{0,0} \tag{11}$$

$$p_{1,0} = \frac{1+\rho}{1+2\rho}\frac{\lambda}{\mu_1}p_{0,0} \tag{12}$$

$$p_{1,1} = \frac{\rho}{1+2\rho}\frac{\lambda(\lambda+\mu_2)}{\mu_1\mu_2}p_{0,0} \tag{13}$$

Assuming the sum of all possible state probabilities is equal to one, we have

$$\sum_{n=0}^{\infty}\sum_{m=0}^{1} p_{n,m} = 1 \tag{14}$$

On plugging all prior state probabilities on Eq.(14), we get:

$$\left[\frac{\lambda(\lambda+\mu_2)}{(1-\rho)(1+2\rho)\mu_1\mu_2} + 1\right] p_{0,0} = 1 \tag{15}$$

The probability of system is at idle,

$$p_{0,0} = \frac{(1-\rho)(1+2\rho)\mu_1\mu_2}{(1-\rho)(1+2\rho)\mu_1\mu_2 + \lambda(\lambda+\mu_2)} \tag{16}$$

A. Average Response Time Model

Average number of health care data packets in the proposed system is calculated as:

$$E(N) = \sum_{n=0}^{\infty} np_{n,0} + \sum_{n=0}^{\infty}(k+1)p_{n,1} \tag{17}$$

On simplification,we obtain

$$E(N) = \frac{\lambda(\lambda+\mu_2)}{(1-\rho)^2(1+2\rho)\mu_1\mu_2 + \lambda(\lambda+\mu_2)(1-\rho)} \tag{18}$$

Using Little's formula [13] the average response time of the MEC server and cloud server are defined as $E(R_1) = 1/\mu_1(1-\rho_1)$ and $E(R_2) = 1/\mu_2(1-\rho_2)$. Similarly, the average response time of CARE framework is stated as

$$E(R) = \frac{E(N)}{\lambda} = \frac{1+\rho+\rho\beta}{\mu_2(1-\rho)(\beta+\rho+\rho^2+2\beta\rho+\rho^2\beta^2)} \tag{19}$$

B. Processing Cost Evaluation

The processing cost the health care data packets at MEC server is given by,

$$C_{MEC} = c_1\frac{\mu_1}{\beta} + c_2E(R_1) \tag{20}$$

where c_1 and c_2 denote the processing cost per data packet and response time cost for each data packet. Using average response time MEC server the prior equation becomes

$$C_{MEC} = c_1\frac{\mu_1}{\beta} + \frac{c_2}{\mu_1-\lambda} \tag{21}$$

The processing cost the health care data packets at could server is given by,

$$C_{Cloud} = c_1\mu_2 + c_2E(R_2) \tag{22}$$

Plugging the value of average response time of cloud server in the above equation, we have

$$C_{Cloud} = c_1\mu_2 + \frac{c_2}{\mu_2-\lambda} \tag{23}$$

Similarly, processing cost corresponding to CARE platform can be evaluated using the average response time of CARE platform(i.e., Eq.(17)):

$$C_{CARE} = c_1\mu_2 + c_2E(R) \tag{24}$$

or

$$C_{CARE} = c_1\mu_2 + \frac{c_2(1+\rho+\rho\beta)}{\mu_2(1-\rho)(\beta+\rho+\rho^2+2\beta\rho+\rho^2\alpha^2)}. \tag{25}$$

V. Results and Discussions

The analytical findings regarding the proposed CARE platform's convergence with the benchmark computing platforms like MEC and Cloud platforms are discussed in this section for different system parameters. The experimental setup was made using MATLAB 2020a, running over an Intel(R) Core(TM) i5 computer with 8GB of RAM and a 2.6GHz processor. The healthcare data packet arrival rate λ is regarded as varying within the range of [10:30] packets/second. The processing rate for the cloud computing servers is fixed at a rate of μ_2 =31 packets/second, and μ_1 gives the processing rate for MEC servers as $\mu_1 = \beta \times 31$ packets/second, with workload factor $\beta = 2$. From Fig. 2, it is observed that the workload for the Cloud platform is the highest, whereas the workload for MEC and CARE platforms is relatively smaller, as evidenced from the graphical illustration.

The average response time for processing various tasks offloaded by the IoHT layer is depicted in Fig. 3. The observed lowest response time for the suggested CARE platform demonstrates its effectiveness, whereas the greatest average response time for the Cloud platform demonstrates a compromised QoS for processing time-sensitive IoHT tasks.

The average response time for processing varying range of workload factor $\beta = [1, 3]$ is shown in Fig. 4. From the graphical representation, it was observed that the average response time incurred by the MEC and Cloud platforms were initially high due to the workload factor being low. However, the average response time of the MEC platform depicted an exponential decay in its response time with increase in the workload factor β. The proposed CARE platform outperformed the benchmark techniques by providing the lowest average response time of 0.0155 ms.

The effect of different task arrival rate λ over the processing costs for the suggested CARE framework along with other benchmark platforms like the MEC and cloud is shown in Fig. 5. With more data packets arriving at the server, it is observed that the Cloud platform's processing cost witnesses an exponential growth, whereas processing cost for the MEC and CARE platform observes a very steady growth, with CARE platform incurring the lowest processing cost among the other two benchmark techniques.

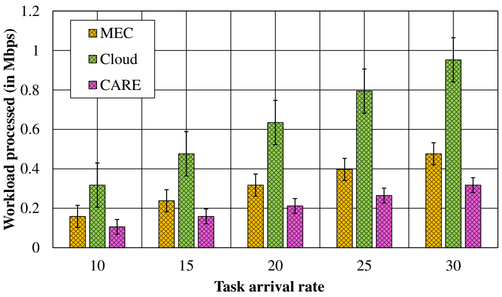

Fig. 2. Workload analysis for different task arrival rates for MEC, Cloud, and CARE platforms.

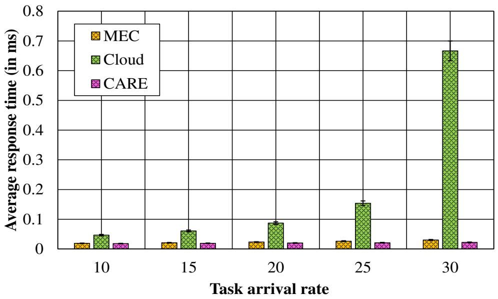

Fig. 3. Average response time (in milliseconds) for MEC, Cloud, and CARE platforms corresponding to different task arrival rates.

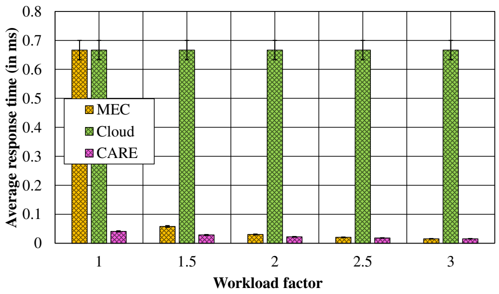

Fig. 4. Sensitivity analysis for average response time of benchmark platforms with proposed CARE approach with different workload factor β.

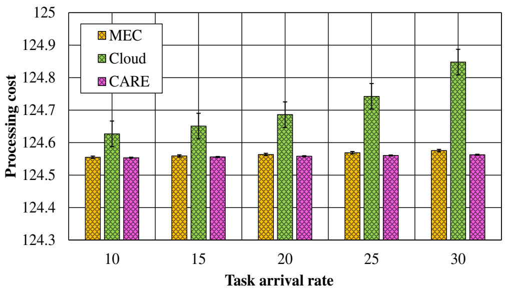

Fig. 5. Processing cost analysis of MEC, Cloud, and CARE approach.

VI. Conclusions

The healthcare sector has witnessed a significant rise as a result of the introduction of wearables and IoHT connected devices. The workforce for traditional healthcare and remote patient monitoring has been further supported by the IoHT's superior computational capabilities. Existing computer frameworks have faced difficulties because to the ongoing creation of health metrics that are collected from a diverse range of sensory devices. In order to satisfy the real-time computational demands of the healthcare industry, the MEC has achieved significant popularity. Recently, the resource-constrained MEC platforms have seen significant difficulties in meeting the QoS requirements of IoHT systems due to the development in healthcare data analytics techniques and the generation of healthcare big data.

In order to facilitate the execution of computationally intensive operations, an adaptive offloading technique acronymed as CARE was proposed that shifts control between the MEC servers and Cloud server to facilitate cost-aware computation of healthcare data. The framework was characterized as a heterogeneous collection of servers in a non-birth death stochastic process. Additional critical performance indicators, such as the ideal workload, typical response time, and processing costs for the various computing platforms, including MEC and Cloud platform, were studied in conjunction with the suggested CARE strategy. The analytical findings revealed that the suggested CARE framework performed better than the MEC and traditional Cloud-based platforms under various sensitivity criteria. The proposed cost-aware computation offloading solution CARE, thus demonstrates effectiveness in managing workloads produced by IoHT systems with reduced response times and processing costs.

References

[1] Li Ping Qian, Yuan Wu, Bo Ji, Liang Huang, and Danny HK Tsang. Hybridiot: Integration of hierarchical multiple access and computation offloading for iot-based smart cities. *IEEE network*, 33(2):6–13, 2019.

[2] Tiago Koketsu Rodrigues, Jiajia Liu, and Nei Kato. Application of cybertwin for offloading in mobile multi-access edge computing for 6g networks. *IEEE Internet of Things Journal*, 2021.

[3] En Wang, Dawei Li, Boxiang Dong, Huan Zhou, and Michelle Zhu. Flat and hierarchical system deployment for edge computing systems. *Future Generation Computer Systems*, 105:308–317, 2020.

[4] Sujit Bebortta, Manoranjan Panda, and Shradhanjali Panda. Classification of pathological disorders in children using random forest algorithm. In *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, pages 1–6. IEEE, 2020.

[5] Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government printing office, 1948.

[6] Mohsin Raza, Muhammad Awais, Nishant Singh, Muhammad Imran, and Sajjad Hussain. Intelligent iot framework for indoor healthcare monitoring of parkinson's disease patient. *IEEE Journal on Selected Areas in Communications*, 39(2):593–602, 2020.

[7] Daniele Ravi, Charence Wong, Benny Lo, and Guang-Zhong Yang. A deep learning approach to on-node sensor data analytics for mobile or wearable devices. *IEEE journal of biomedical and health informatics*, 21(1):56–64, 2016.

[8] Imran Ahmed, Gwanggil Jeon, and Francesco Piccialli. A deep-learning-based smart healthcare system for patient's discomfort detection at the edge of internet of things. *IEEE Internet of Things Journal*, 8(13):10318–10326, 2021.

[9] VK Daliya, TK Ramesh, and Seok-Bum Ko. An optimised multivariable regression model for predictive analysis of diabetic disease progression. *IEEE Access*, 9:99768–99780, 2021.

[10] Jianxi Wang and Liutao Wang. A computing resource allocation optimization strategy for massive internet of health things devices considering privacy protection in cloud edge computing environment. *Journal of Grid Computing*, 19(2):1–14, 2021.

[11] Denis A Pustokhin, Irina V Pustokhina, Poonam Rani, Vineet Kansal, Mohamed Elhoseny, Gyanendra Prasad Joshi, and K Shankar. Optimal deep learning approaches and healthcare big data analytics for mobile networks toward 5g. *Computers & Electrical Engineering*, 95:107376, 2021.

[12] V Jagadeeswari, V Subramaniyaswamy, R Logesh, and Varadarajan Vijayakumar. A study on medical internet of things and big data in personalized healthcare system. *Health information science and systems*, 6(1):1–20, 2018.

[13] Kishor S Trivedi. *Probability & statistics with reliability, queuing and computer science applications*. John Wiley & Sons, 2008.