UNIVERSITY OF RWANDA
COLLEGE OF SCIENCE AND TECHNOLOGY

AFRICAN CENTRE OF EXCELLENCE IN INTERNET OF THINGS

OPTIMIZED PATIENT FLOW PROCESS – A CASE OF OUTPATIENT AND SURGICAL DEPARTMENTS IN SUB-SAHARAN AFRICA HEALTHCARE SYSTEMS

PhD. Thesis submitted in the fulfilment of requirements of award of PhD Degree in Internet of Things – Wireless Sensor Networking

Kambombo Mtonga

SEPTEMBER/2022

UNIVERSITY OF RWANDA
COLLEGE OF SCIENCE AND TECHNOLOGY

AFRICAN CENTRE OF EXCELLENCE IN INTERNET OF THINGS

OPTIMIZED PATIENT FLOW PROCESS – A CASE OF OUTPATIENT AND SURGICAL DEPARTMENTS IN SUB-SAHARAN AFRICA HEALTHCARE SYSTEMS

PhD. Thesis submitted in the fulfilment of requirements of award of PhD Degree in Internet of Things – Wireless Sensor Networking

Kambombo Mtonga
218014372

Thesis Supervisor: Prof. Santhi Kumaran
Thesis Co-Supervisor(s): Assoc. Prof. Kayalvizhi Jayavel
Dr. Omar gatera

JUNE/2022

DECLARATION

I hereby declare that the dissertation entitled “Optimized Patient Flow Process – A Case of Outpatient and Surgical Departments in sub-Saharan Africa Healthcare Systems” to be submitted for the Degree of Doctor of Philosophy is my original work and the dissertation has not formed the basis for the award of any degree, diploma, associateship or fellowship of similar other titles. It has not been submitted to any other University or Institution for the award of any degree or diploma. Where other people’s work has been used (either from a printed source, internet or any other source), this has been properly acknowledged and referenced in accordance with the requirements as stated in the University of Rwanda's antiplagiarism policy.

Kambombo Mtonga, a Ph.D. student of UR-ACEIoT student ID 218014372, successfully defended the thesis/dissertation entitled "OPTIMIZED PATIENT FLOW PROCESS – A CASE OF OUTPATIENT AND SURGICAL DEPARTMENTS IN SUB-SAHARAN AFRICA HEALTHCARE SYSTEMS", which he prepared after fulfilling the requirements specified in the associated legislations, before the thesis examination members whose signatures are below.

Thesis Supervisor : **(Prof.) Santhi Kumarani** ……
The Copperbelt University

Co-Supervisor (s) : **(Assoc. Prof.) Chomora Mikeka** ……..
SRM Institute of Science and Technology

(Dr.) Omar Gatera …….
University of Rwanda

Viva Voce Members : **(Prof.) Vincent Havyarimana** …….
Burundi Higher Institute of Education (ENS) (External)

(Prof.) Hung Tran …….
Phenikaa University (External)

(Prof.) Denis Ndanguza …….
University of Rwanda (Chair)

Date of Submission : 15/09/2022
Date of Defense : 28/06/2022

This thesis is dedicated to my beloved children,

ACKNOWLEDGEMENT

I would like to give my at most thanks to my supervisor Prof. Santhi Kumaran, for her patient and timeless support over the period it has taken me to complete this thesis. Special appreciation should also go to Assoc. Prof. Chomora Mikeka for the mentorship role you played throughout the period of this research. Special mention should also go to Assoc. Prof. Kayalvizhi Jayavel for her warm, kind and sincere friendship. No words can capture how grateful I am for your immeasurable support towards the completion of my research. Special thanks should go to Dr. Omar Gatera for his continued guidance on professionalism and ethical conduct in research. Special thanks should also go to my colleagues for helping to create a friendly atmosphere. Special appreciation and recognition belong to Lucia Kabanga mother to my two beloved daughters for your support and care, and Jessie Victoria Chavinda and Gordon Jeremiah Chavinda for your motherly and fatherly care. Last, but not least, special thanks belong to all my entire family.

09/2022 Kambombo Mtonga

TABLE OF CONTENTS

ABBREVIATIONS

API : Application Program Interface
ARIMA : Auto-Regressive Integrated Moving Average
CNN : Convolutional Neural Network
DBA : Deep Belief Architectures
DBM : Deep Belief Machines
ED : Emergency Department
FIFO : First In First Out
GARCH : Generalized Autoregressive Conditional Heteroscedasticity
GLM : General Linear Model
GPS : Global Positioning System
GSM : Global System for Mobile communication
HIES : Health Information Exchange System
ILP : Integer Linear Programming
IoT : Internet of Things
JSON : JavaScript Object Notation
KCMC : Kilimanjaro Christian Medical Centre
LDR : Light Dependent Resistor
LPT : Largest Processing Time
MICU : Mobile Intensive Care Unit
MQTT : Message Queuing Telemetry Transport
MRF : Markov Random Field
NICU : Neonatal Intensive Care Unit
OPD : Outpatient Department
OT : Operating Theatre
PST : Passenger Services Time
REST : Representational state transfer
RF : Readio Frequency
SD : Standard Deviation
SPT : Shortest Processing Time
TORSCHE : Time Optimisation, Resources, Scheduling

UCM : Unobserved Components Model

VAR : Vector Autoregression

SYMBOLS

G	: A graph – a network of health facilities
E	: Edge set of G
P	: A set of patients
L	: A set of hospitals in a particular region
M	: Total number of hospitals in a particular region
R	: Average number of patients in the catchment are of a hospital
$\boldsymbol{w_{ij}}$	: Weight of link between i and j
$\boldsymbol{b_i}$	: Bias of unit i
K	: Total number of time intervals
$\boldsymbol{\vartheta}$	: Convolution operation overhead
$\boldsymbol{\alpha}$	: Percentage computations due to convolution
$\boldsymbol{\beta_j}$	: Computation percentage of neural network layer
$\boldsymbol{\lambda}$	: Patient arrival rate
$\boldsymbol{\mu}$	: Patient service rate
$\boldsymbol{c}$	: Number of servers
$\boldsymbol{\rho}$	: System utilization
$\boldsymbol{P_\theta}$	: Probability of having θ patients in the system
$\boldsymbol{\sigma_i}$	: Average number of patients arriving over $[t_{i-1}, t_i]$
$\boldsymbol{M/M/c}$	: A multi-channel queueing system with Poisson arrival and exponential distribution
$\boldsymbol{r_{Task_j}}$	: Release time (Ready time) for task j
$\boldsymbol{d_{Task_j}}$	: Due date of task j
$\boldsymbol{\bar{d}_{Task_j}}$	: Deadline time of task j
$\boldsymbol{C_{Task_j}}$	: Completion time of task j
$\boldsymbol{L_{Task_j}}$	: $L_{Task_j} = C_{Task_j} - d_{Task_j}$
$\boldsymbol{\omega_{Task_j}}$	: Priority of task j
$\boldsymbol{F_{Task_j}}$	: Task flowtime, i.e., $F_{Task_j} = C_{Task_j} - r_{Task_j}$
$\mathbb{R}^m$	m dimensional vectors of real numbers

LIST OF TABLES

LIST OF FIGURES

Optimized Patient Flow Process – A Case of Outpatient and Surgical Departments In sub-Saharan Africa Healthcare Systems

Kambombo Mtonga

University of Rwanda, Kigali, 2022

Thesis Supervisors: Prof. Santhi Kumaran, Assoc. Prof. Chomora Mikeka, Dr. Omar Gatera

This thesis adds to the plethora of knowledge on interventions/mechanisms that can be employed to overcome the challenges rocking the healthcare systems in sub-Saharan Africa. Extended waiting time due to overcrowding of patients in hospital waiting rooms, a consequence of few and sparsely located health facilities is negatively impacting provision of quality healthcare services and optimal usage of resources in healthcare facilities. The outpatient and surgical departments are two of the sections where patients constantly experience extended waiting times. This research demonstrates how existing mathematical tools such as integer linear programming, queuing theory and other technologies, such as machine learning and the internet of things can be utilized to uncover and address bottlenecks in the delivery and access to healthcare, with special focus on patient flow process. Through a modelling approach we investigate how sharing of patient load information among health facilities can help to reduce patient waiting time. We propose re-assigning excess patient loads to nearby facilities that have minimal load as a way to control overcrowding and reduce queue abandonment. An Internet of things integrated smart bus system is proposed to aid the movement of patients to less crowded health facilities.

The relationship between patient flow and staff scheduling is investigated using queuing theory. We investigate how the various parameters that govern patient movement in the outpatient can be optimized to improve quality of healthcare delivery and access. Furthermore, using an integer linear programming approach, we study the problem of optimal assignment of operating rooms, under conditions of limited available operating facilities and specialized equipment. This research shows that quality healthcare service defined by low access time and reduced queue abandonment is possible in sub-Saharan Africa region.

Chapter 1: Introduction

Clinical operations are characterized with tensions among competing needs of clinicians, patients and healthcare managers. At the center of these competing needs is time. Time is an important factor in determining the efficiency of almost all sections of the hospital. For example, for patients, lengthy waiting times lowers the quality of the clinical experience, whilst increased face time implies improved experience [1]. For clinicians, delays in the clinical system builds pressure to stay on schedule, a situation that leads to reduced job satisfaction. For hospital managers, their goal is keeping operational cost to a minimum and this is possible if they control inefficient use of costly labour by avoiding unnecessary overtime and minimize usage of other capital resources [2] [3]. Waiting time is among the key quality indicators of efficiency of healthcare systems; since lengthy waiting times among others; increases dissatisfaction, stress, health care costs and may lead to fatalities [4] [5].

In general, for general consultation visits to a healthcare facility, a patient's waiting time can be described as the time between a patient's arrival and the time the patient enters a consultation room. The time a patient spends in the consultation room is referred to as productive time. However, a fine-grained characterization of a patient's waiting time has been proposed by other researchers. This is so because upon entry into the consultation room, a patient may still continue to wait for service, due to unavailability of the medical personnel. Hence, true patient waiting time is defined by the sum of the initial wait time (i.e., the time a patient enters the consultation room minus a patient's arrival time) and room delay (i.e., the time a patient spends in the consultation room without presence of the attending medical personnel) [6] [7] [8]. With regard to other health facility environments, the productive time is basically the sum of the times that the patient makes use of clinical resources in activities for which the process completion is dependent on. For example, hospitalized patients may interact with other clinical resources such as rooms, beds, technicians and nurses.

This research identifies two hospital areas in which time is key and takes a modeling approach to propose solutions that can lead to robustness in healthcare service provision. We explore the causes and the impacts of extended waiting times in the provision of outpatient services and also the provision of elective surgical services in the sub-Saharan Africa healthcare environments.

Figure 1.1 shows a functional-based general outlay of a healthcare facility.

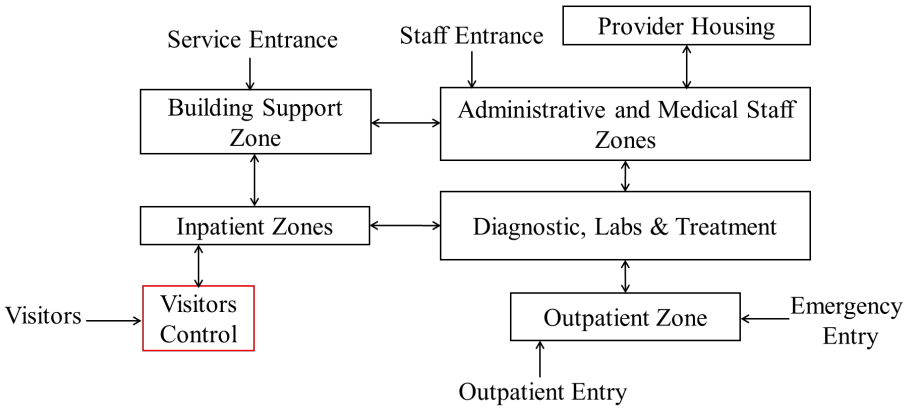

Figure 1.1. General hospital functionality-based outlay

1.1 The outpatient department and patient waiting time

The outpatient department (OPD) plays an integral part in the provision of health care services, as it performs many functions beyond diagnosis and treatment. For example, patients admitted for inpatient services are also screened in the OPD; and patients who are discharged receive follow-up treatment in the OPD. These OPDs also serve as teaching facilities for medical students, nursing students, and residents, and for health care research [9].

In the OPD, the extended waiting times are a result of overcrowding of patients in the waiting rooms. Overcrowding implies long patient queues; hence, patients have to endure a painful and lengthy wait for treatment [10] [11]. In the low to middle income regions (e.g., the sub-Saharan Africa), overcrowding in the OPD is further worsened by the fact that the healthcare facilities are few and scattered, frequently characterized by low-skilled staffs and unreliable and error-prone diagnostics resulting from manual data capture processes. In addition, the OPD is flooded with non-appointment patients, a situation which leads to demand for health care services exceeding the capacity of the healthcare facilities that are seldom appropriately staffed. Furthermore, traditionally, in the OPD, patients are normally served on a first-come-first-serve policy, a situation which disadvantages patients who arrive late due to long distances. This underscores the need for triage to ensure that patients are treated in order of urgency, then in order of arrival. Unfortunately, in the developing world, where shortage of healthcare personnel is a big problem, triage can take long, leading to unnecessary delays in treatment [12] [13]. Hence, in such environments, there is a need to adopt other sophisticated approaches to study

patient flow behaviors which can inform the staff scheduling process, thereby enabling matching of staffing ratios to patient demand, leading to low access times.

1.2 The surgical department and patient waiting time

As regards the surgical department, surgical waiting time is a big problem, especially for the elective surgeries. In the sub-Saharan Africa, where most health facilities do not have exclusively dedicated theaters for elective and emergency surgeries, it is the elective surgical requests which are likely to be sacrificed in case of a sudden surge in demand for emergency surgical services. The cancellation or shifting of elective cases as healthcare facilities prioritize emergency surgical cases leads to prolonged waiting times for the service resulting in deterioration of the patients as the diseases progresses, which is commonly the case for cancer patients [14]. For example, in a study carried out in [15], where the authors assessed reasons for cancellation of elective surgeries in Malawi, it was discovered that, out of the 10, 000 elective surgeries, 4,740 were cancelled with infrastructural limitations accounting for 84.8% as reason for cancellation. Similarly, Rajaguru et al. in [16] carried out a cross-sectional study to asses surgical care delivery in Tanzania and they discovered that, out of the 3,817 scheduled surgeries, 238 were cancelled, representing a 20.8% cancellation rate. In [17], Kajja et al. assessed the factors causing delayed elective surgery at Mulago Hospital, a teaching hospital in Uganda. It was discovered that 33% of the 133 elective surgeries were delayed due to shortage of operating space.

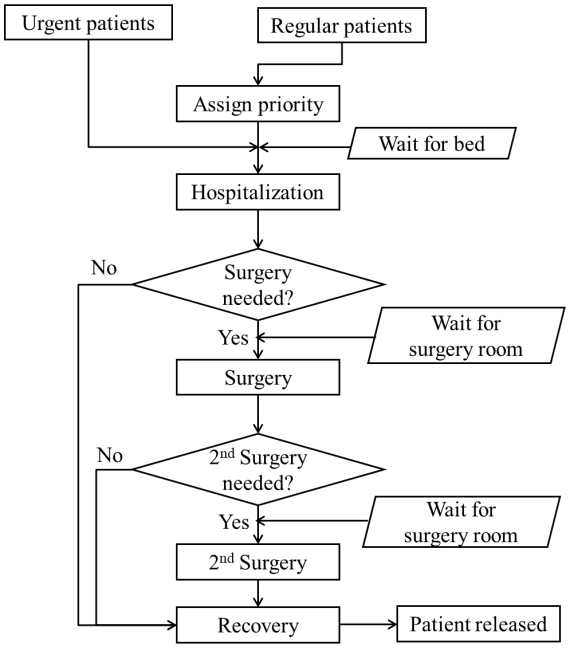

Figure 1.2. Flow diagram for patient flow in the elective surgery department

Considering that operation theaters are an expensive investment, the optimal usage of these facilities can ensure that the dividend from the investment on the facilities is maximized. Proper scheduling approaches are key to the maximized utilization of the operating theatres [18]. The integration of proper technology in the scheduling process can aid the creation of optimal schedules of patients who want to undergo surgeries leading to minimization of patient waiting times under conditions of limited available operating facilities, limited human resources and specialized equipment [19] [12]. Figure 1.2 shows the process flow in the elective surgery department.

1.3 Overview of the major contributions and thesis outline

To address the above issues, this research takes a modeling approach to demonstrate how low access times can be achieved in the access of outpatient services and surgical services. Below we briefly discuss the main focus of this research, which also highlights the main contribution of this thesis.

The first specific focus of this research is to address the issue of unpredictable flow of patients to the OPD, which makes scheduling of staff to match demand difficult. Matching demand for health care services and staff levels requires patient flow prediction systems that are based on accurate patient flow models. To this end, we predict patient traffic flow using a deep learning approach. Knowledge of the anticipated patient load would help healthcare facility managers to create more accurate staff schedules, timely asses financial needs and strategically plan for the facility. In case of excess patient load, we propose directing patients to nearest facilities where predicted patient load is less. This creates a need for efficient transportation method. To avoid putting pressure on the ambulatory service, an IoT integrated smart public bus service architecture is proposed.

The second focus of this research is the problem of optimized patient flow within a health center and how it relates to staff scheduling. We apply queuing theory techniques to assess and evaluate the relationship between staffing ratios and waiting times at Ntaja health center of Malawi. Specifically, we model queue parameters such as arrival rate, service rate, etc. and validate them using secondary data for the health center. Since adults and children patients join different queues, we treat the service times of each patient group independently and derive queue

parameters governing each queue. By considering the patient flow behavior as a closed network, we approximate recommended staffing ratios to achieve steady state so as to reduce patient waiting times.

The third and final focus of this research is the problem of suboptimal scheduling of elective surgical cases. We model the problem of scheduling of elective surgical cases in the limited operating rooms at the Kilimanjaro Christian Medical Center of Tanzania. The operating room scheduling problem for elective surgeries is modeled as an integer linear programming problem. The model is solved using Torsche toolbox with the help of MatLab routines and functions. The solution is tested using secondary data from the Kilimanjaro Christian Medical Centre, a referral hospital located on the northern corridor of the Republic of Tanzania.

The above highlighted points represent the main deliverables (which correspond to the objectives) of this research work. The remainder of this thesis is organized as follows: Chapter 1: proposes a Deep Learning-based patient flow prediction model and an IoT integrated smart transportation system. The Chapter 3: presents the modeling of patient flow within a health center by applying queueing theory techniques. The Chapter 1: presents the modeling of the problem of scheduling of elective surgery cases in the limited operating rooms. The Chapter 1: concludes the thesis and give insight on future works.

REFERENCES

[1] E. Batbaatar , J. Dorjdagwa, A. Luvsannyam, M. Savino and P. Amenta, "Determinants of patient satisfaction: a systematic review," *Perspectives in public health,* vol. 137, no. 2, pp. 89-101, 2017.

[2] C. Lin, G. ALbertson , L. Schilling, E. Cyran, S. ANderson, L. Ware and R. Anderson, "Is patients' perception of time spent with the physician a determinant of ambulatory patient satisfaction?," *Archives of internal medicine,* vol. 161, no. 11, pp. 1437-1442, 2001.

[3] M. Pedrazza, S. Berlanda, E. Trifiletti and F. Bressan, "Exploring physicians' dissatisfaction and work-related stress: development of the PhyDis scale," *Frontiers in psychology,* vol. 18, no. 7, p. 1238, 2016.

[4] A. Baker, Crossing the quality chasm: a new health system for the 21st century, British Medical Journal Publishing Group, 2001.

[5] A. Boonma, K. Sethanan, S. Talangkun and T. Laonapakul, "Patient waiting time and satisfaction in GP clini at a tertiary hospital in Thailand," in *In MATEC Web of Conference*, 2018.

[6] A. Aeenparast, F. Farzadi, F. Maftoon and H. Yahyazadeh, "Patient flow analysis in general hospitals: how clinic disciplines affect outpatient wait times," *Hospital Practice and Research,* vol. 4, no. 4, pp. 128-133, 2019.

[7] S. Asefzadeh, "Patient flow analysis in a children's clinic," *International Journal for Quality in Health Care,* vol. 9, no. 2, pp. 143-147, 1997.

[8] K. Makwana and D. Dave, "A study of improving hospital out patient department (OPD) using queuing network analysis methodology," *Journal of Management (JOM),* vol. 6, no. 3, 2019.

[9] T. Hong, P. Shang, M. Arumugam and R. Yusuf, "Use of simulation to solve outpatient clinic problems: a review of literature," *South Africa Journal of Industrial Engineering,* vol. 24, no. 3, pp. 27-42, 2013.

[10] M. Bahadori, E. Teymourzadeh, R. Ravangard and M. Raadabadi, "Factors affecting the overcrowding in outpatient healthcare," *Journal of Educ Health Promot.,* vol. 6, no. 21, 2017.

[11] M. Yarmohammadian, F. Rezaei, A. Haghshenas and N. Tavakoli, "Overcrowding in emergency departments: a review of strategies to decrease future challenges," *Journal of Res Med Sci.,* vol. 22, no. 23, 2017.

[12] T. Tran, U. Nguyen, V. Nong and B. Tran, "Patient waiting time in the outpatient clinic at a central surgical hospital of Vietnam: implications for reource allocation," *F1000 Research,* vol. 6, no. 454, 2017.

[13] Y. Shukla, R. Tiwari, B. Rohit and P. Kasar, "An assessment of OPD registration counter services and. channelization of patients in NSCB medical college hospital," *International Journal of Medical Research Science and Public Health,* vol. 4, no. 10, pp. 1468-1472, 2015.

[14] S. Acharya, D. Dharel, S. Upadhyaya, N. Khanal, S. Dahal, S. Dahal and K. Aryal, "Study of factors associated with waiting time for patients undegoing emergency surgey in a

tertiary care center in Nepal," *Journal of Society of Anesthesiologists of Nepal,* vol. 1, no. 1, pp. 7-12, 2014.

[15] M. Prin, J. Eaton, O. Mtalimanja and A. Charles, "High elective surgery cancellation rate in Malawi primarily due to infrustructural limitations," *World Journal of Surgery,* vol. 42, no. 6, pp. 1597-602, 2018.

[16] P. Rajaguru, M. Jusabani, H. Massawe, R. Temu and N. Sheth, "Understanding surgical care delivery in Sub-Saharan Africa: a cross sectional analysis of surgical volumes, operations and financing at a tertiary referral hospital in rural Tanzania," *Global health Research and Policy,* vol. 4, no. 1, pp. 1-9, 2019.

[17] I. Kajja and C. Sibinga, "Delayed elective surgery in a major teaching hospital in Uganda," *International Journal of Clinical Transfusion Medicine,* vol. 4, no. 1, pp. 1-6, 2014.

[18] D. Antonelli and T. Taurino, "Application of a patient flow model to a surgery department," in *In 2010 IEEE workshop on health care management (WHCM)*, 2010.

[19] C. Granja, B. Almada-Lobo, F. Janela, J. Seabra and A. Mendes, "An optimization based on simulation approach to the patient admission scheduling problem using a linear programming algorithm," *Journal of Biomedical Informatics,* vol. 52, pp. 427-437, 2014.

[20] S. Sardana, S. GS, A. Vij and S. Kale, "Analysis of waiting time for elective surgical procedures in neurosurgery department at a tertiary care teaching hospital in NCT, India," *Int J Res Med Sci.,* vol. 5, no. 10, pp. 4538-4, 2017.

Chapter 2: Machine Learning-based Patient Load Prediction and IoT Integrated Intelligent Patient Transfer Systems

A mismatch between staffing ratios and service demand leads to overcrowding of patients in waiting rooms of health centers. Overcrowding consequently leads to excessive patient waiting times, incomplete preventive service delivery and disgruntled medical staff. Worse, due to the limited patient load that a health center can handle, patients may leave the clinic before the medical examination is complete. It is true that as one health center may be struggling with an excessive patient load, another facility in the vicinity may have a low patient turn out. A centralized hospital management system, where hospitals are able to timely exchange patient load information would allow excess patient load from an overcrowded health center to be re-assigned in a timely way to the nearest health centers. In this chapter, a machine learning-based patient load prediction model for forecasting future patient loads is proposed. Given current and historical patient load data as inputs, the model outputs future predicted patient loads. Furthermore, we propose re-assigning excess patient loads to nearby facilities that have minimal load as a way to control overcrowding and reduce the number of patients that leave health facilities without receiving medical care as a result of overcrowding. The re-assigning of patients will imply a need for transportation for the patient to move from one facility to another. To avoid putting a further strain on the already fragmented ambulatory services, we assume the existence of a scheduled bus system and propose an Internet of Things (IoT) integrated smart bus system. The developed IoT system can be tagged on buses and can be queried by patients through representation state transfer application program interfaces (APIs) to provide them with the position of the buses through web app or SMS relative to their origin and destination stop. The back end of the proposed system is based on message queue telemetry transport, which is lightweight, data efficient and scalable, unlike the traditionally used hypertext transfer protocol.

2.1 Introduction

Overcrowding of hospital waiting rooms by patients characterizes the outpatient departments (OPDs) of health care centers of developing countries [1] [2]. It is true that most of the patients that visit health centers have varying health challenges and some may require prompt medical

attention. However, due to overcrowding, these patients are made to endure a lengthy and painful wait for treatment [3] [4]. In the developing world, the problem is aggravated by the fact that health facilities are few, scattered and under-staffed. In such countries, it is a common experience for patients visiting an OPD to return home without receiving medical attention. For example, data collected from the patients record book at one general OPD showed that, within a one month period, of the 2211 patients who walked-in into the OPD department, only 1870 (84.58%) patients were seen and 341 (15.42%) patients were not seen [5].

To deal with such excess demand, physicians may opt to work overtime or delegate some work to nurses. Consequently, this may cause the healthcare center to incur extra costs and/or provide reduced quality of service to the patients [6] [7]. Reduced quality of service might impact patient's trust towards the health facility and/or practitioners. Overcrowding in hospitals also relates to patient's safety. Due to limited space, patients may be placed in inappropriate spaces, which can be recipe for complications and fatalities [8] [9]. It is the case in many situations that as one heath center is being pressed with high demand beyond its capacity, the facilities nearby may be having few patients. However, there is no mechanism for timely exchange of information about patient load between health facilities. This leads to patients in overcrowded health facilities having to endure a lengthy and painful wait for treatment which may not even be given.

Patients are usually transferred between health facilities and the purpose for such transfers is to maintain the continuity of medical care. Normally, these are patients that require specialized treatment or procedure and are transferred to facilities with such specialized equipment and personnel, i.e., the benefits of care available at another facility against the potential risks involved. The patient transfer process involves such key elements as; decision to transfer and notification, patient pre-transfer stabilization and preparation, choice of appropriate mode of transfer (e.g., land or air transport), personnel to accompany the patient, equipment and monitoring required during the transfer, and finally, the documentation and handover of the patient at the receiving facility [10] [11]. In each transfer these key elements are followed so as not to affect patient prognosis.

Since this works' focus is the outpatient department, we make the following assumptions: (1) Upon arrival, patients are triaged such that the critical patients are served first and then the least

critical the last. (2) Excess patient load can only be transferred once, i.e., a patient can only be transferred once. As pointed out above, patient transfer process starts with a decision to transfer the patient because of the exposure of the patient and the staff to additional risks and additional expenses for the guardians. A senior consultant level doctor is responsible for making the decision to transfer the patient. A patient's guardians are made aware of the benefits and risks involved during such a transfer. Written and informed consent of patients' relatives along with the reason(s) to transfer is mandatory before the transfer. In this work, the role of the senior doctor will be played by the Central Controller of the system. Since the system provides the patient with alternative health care service source, it is up to the patient to go to the recommended facility or wait in the queue and risk returning home without getting medical care. Hence, a patients' consent will be by virtual of accepting to go to the recommended facility. Since patients are triaged upon arrival (with the most critical served first), we assume that the patients being recommended for transfer are stable enough such that pre-transfer stabilization and preparation are not required. The choice of appropriate mode of transport is key in this work and is covered in Section 2.7. The consideration of guardian, patient monitoring during transfer and documentation are beyond the scope of this work.

By intelligently re-assigning patients from congested facilities to less congested ones, overcrowding of OPDs can be managed in a timely way. Re-assigning excess patient loads reduces the number of patients that return to their homes without being attended to by physicians. Consequently, this may lead to improvement in the quality of service since the service demanded of medical staff would be kept to a minimum. Patient flows in hospitals are usually not continuous and stable; rather, patient flows are complex and may change suddenly [12]. Hospitals experience two kinds of service demands: (1) Event driven demand, i.e., ambulatory arrivals during accidents and natural disasters, and (2) regular demand from the catchment area. However, a lack of real time patient flow information may result in some health facilities straining to satisfy the demand while at the same time facilities in the same vicinity may have minimal demand. While patients may be wary about being transferred to a hospital they know little or nothing about, however, such transfers may significantly reduce the amount of time a patient must wait to get medical care [13]. Attempts in the literature exist to predict patient traffic, but most of the work focuses on patient flow in the emergency department (ED) [14] [15]. Furthermore, independent attempts have been made to predict and schedule traffic flow [16] [17]. However, to our knowledge, no literature has explored the integration of the two

processes using machine learning techniques.

In this chapter, a deep learning-based patient load prediction model is proposed. The model aid in overcoming the effects of overcrowding in hospital waiting rooms, which results from a mismatch between hospital staffing ratios and the demand for health care services. Overcrowding of hospital waiting rooms leads to excessive patient waiting times, incomplete service delivery and unhappy medical staff. Worse, due to the limited patient loads that a health facility can handle, patients may leave the facility before the medical examination is complete. However, it is true that, as one health facility may be struggling with excessive patient load, another facility in the vicinity may have low patient turn out. In this work, we assume the existence of a hospital information management system, which enables timely sharing of excess patient load information among hospitals [18]. The proposed machine learning-based patient load prediction model takes current and historical patient load data as inputs and outputs future predicted patient load. Furthermore, in the case of excess patient loads, we propose to re-assign excess loads to nearby facilities that have a minimal load as a way to control overcrowding and reduce the number of patients that leave health facilities without receiving medical care.

The re-assigning of patients will imply a need for transportation for the patient to move from one facility to another. To avoid putting a further strain on the already fragmented ambulatory services, we assume the existence of a scheduled bus transport system which can support the timely movement of patients from the bus stop nearest to the source facility to the destination facility. Building on this assumption we propose an Internet of Things (IoT) integrated smart bus framework. We develop an Arduino-based smart bus system kit that can be tagged on public buses and can be queried by patients through representation state transfer application program interfaces (APIs) to provide them with the position of the buses through web app or SMS relative to their origin and destination stop. The back end of the proposed system is based on message queue telemetry transport (MQTT), which is lightweight, data efficient and scalable, unlike the traditionally used hypertext transfer protocol (HTTP). To ensure the reliability of the smart transport system, our solution makes use of the real time location of the buses to compute the approximated time for the bus to reach a particular destination. By saving a bus's location data on the server together with corresponding timestamps, the system is able to estimate the arrival time of the bus to a particular bus stop. Alternatively, the arrival time can also be approximated using services like Google maps [19].

IoT enables advanced services by interconnecting physical and virtual things based on existing and evolving inter-operable information and communication technologies. The IoT gives immediate access to information about physical objects and leads to innovative services with high efficiency and productivity [20]. Building on the power of IoT, we demonstrate its application in the transportation system by developing an IoT-based smart bus system. Generally, a smart bus system consists of four basic components, namely smart bus depots, smart bus stops, smart buses and interactive citizen interfaces (web portal based and smart phone app based). Each of these components are connected through the Internet. The smart bus depots, smart bus stops and smart buses consist of a number of heterogeneous wireless and embedded sensors. These sensor networks are connected to the city internet backbone through Wi-Fi hotspots in the bus stops, depots and inside buses. These intelligent autonomous devices attached to or embedded into the system senses the user requirements and interacts with them, shares information with other devices and takes decisions without any human intervention. Buses are widely used public transportation in many cities today. Normal buses can be converted into smart buses with the incorporation of intelligent sensors and IoT devices. Our contributions in this work are three-fold:

- Using a deep learning approach, we propose a patient traffic flow prediction model.
- We propose combining the deep learning patient load prediction and hospital assignment using the predicted patient load as criterion to perform the intelligent hospital assignment.
- We develop an IoT-based smart bus system. We explore scalable and efficient techniques that allow the system to handle increased requests for data.

The remainder of this chapter is organized as follows. In Section 2.2, a review of the relevant literature and machine learning structures is presented. In Section 2.3, we model our system architecture and training model. In Section 2.4, we propose a deep learning-based patient load prediction method. In Section 2.7, we present our proposed IoT-based smart bus system that allows the estimation of the time for the smart bus to arrive at a bus stop, or the time for it to reach a destination such that transferred patients do not experience unnecessary delays due to transport congestion. Finally, Section 2.8 concludes this chapter.

2.2 Related work

A lot of research efforts have been made to deal with the problem of overcrowding with regard to the ED of the hospital [21] [22]. To deal with this problem, inter-hospital transfers have been proposed as a possible solution [23]. However, overcrowding is also a problem in the OPD and equally affects patient satisfaction, which is affected by the efficiency of the services rendered. Efficient provision of health services encompasses such issues as waiting time to consultation, duration of consultation, timely response to emergencies, quick drug dispensation, and timely and accurate diagnosis [24] [25]. The challenge in dealing with excess demand results from the challenges in predicting patient flow in the OPD. Patient load forecasting involves predicting patient loads for future time periods. Depending on time horizons, patient load prediction can be categorized into: Long-range forecasting, medium-range forecasting, short-range forecasting and real-time or very short-term forecasting. Accurate patient flow predictions may aid in improving hospital management efficiency.

Traditional methods have been applied in forecasting hospital visits. In [26], Dan and Qualls explored the problem of predicting ED patient volume, length of stay and acuity. They studied five models, namely raw observations, moving averages, mean values with moving averages, seasonal indicators with moving averages and auto-regressive integrated moving averages (ARIMAs). It was discovered in this study that simpler models performed best. In [27], Rotstein et al. explore the problem of short-range forecasting of patient volume. A general linear model (GLM) is formulated that can be applied for short-range forecasting of patient volume. In [28], Batal et al. developed equations for predicting daily patient volumes via stepwise linear regression analysis. In [29], Reis and Mandl developed a trimmed mean seasonal model for the expected number of daily patient visits to an ED. In [30], Brillman et al. constructed a first order cyclical regression model with fixed-width sine and cosine harmonics as the seasonal component and a hierarchical model with a scalable Gaussian function as the seasonal component for ED daily respiratory chief complaints. In [31], Flottemesch et al. formulated a mathematical model for forecasting ED censuses. In [32], Boyle et al. investigated the performance of a general linear regression model formed with 11 dummy variables in forecasting monthly patient admissions. In [33], Au-Yeung et al. forecasted patient arrivals to an accident and ED via a structural time series model. In [34], Kam et al. studied the problem of predicting daily patient numbers for a regional medical center via the application of time series analysis. Capan et al. applied time series analyses; specifically, they applied best-fitting

models of ARIMA and linear regression to various prediction models and compared the results using error statistics. Their work aimed at forecasting censuses in neonatal intensive care units (NICU) [35]. Other interesting studies relating to predicting patient flow in ED have been reported in [36] [37] [38] [39]. While these traditional techniques are popular in forecasting hospital visits, they are not good at dealing with complexity in hospital visits data.

Artificial Intelligence techniques, e.g., artificial neural networks (ANNs), form another approach for hospital visit forecasting [40]. However, ANN is not as prevalent as the traditional methods in this field. In [41], Jones et al. explored the performance comparison of exponential smoothing, seasonal ARIMA, ANN and time series regression in daily ED patient volume prediction with linear regression. Their results show that the former four methods did not perform consistently in forecasting with a sample, although they all performed better in sample fitting. In [42], Aladag and Aladag modeled the number of outpatient visits by ANN using different activation functions. In [43], Xu et al. modeled daily patient arrivals at ED via ANN. Kottalanka et al. developed an artificial intelligence model based on back-propagation Neural Network for predicting patient inflow [44]. Clearly, more research is needed to demonstrate and harness the power of machine learning to improve the provision of healthy services, which can lead to improved quality of care and also efficient utilization of resources.

The patient flow data is basically time series data. In [45], a time series is defined as a vector $X = \{x^{(1)}, x^{(2)}, \cdots, x^{(n)}\}$, where each element $x(t) \in \mathbb{R}^m$ pertaining to X is an array of m values $\{x^{(1)}, x^{(2)}, \cdots, x^{(n)}\}$, with each of the m values corresponding to the input variables measured in the time series. Several traditional techniques of manipulating time series data including traditional artificial neural networks (ANNs) have been exploited in the literature towards application in Economics, Engineering and Medicine [46] [47]. Deep learning however has shown exceptional performance in both classification and forecasting applications [48] [49]. For classification, the existing methods generally relied on the usage of domain specific features normally crafted manually by human experts. However, getting best features was a daunting task and performance of the classifier was heavily dependent on their quality. The advantage of deep learning is in its ability to learn such features by themselves, reducing the need for human experts [50] [51].

Convolutional neural network (CNN) models have continued to gain popularity among the deep

learning community in various domains [52]. Due to their efficiency with data that has topological structure in the features space, CNN models have achieved the best results in computer vision [53]. They have also been applied successfully to sequential data such as sentences, time series, and speech [54] [55] [56] [57]. CNNs employ sets of shared weights across the whole input, which is efficient both statistically and computationally. CNNs are particularly interesting in domains associated with processing of large amounts of data [58]. Since patient load data can be categorized as big data, the application of deep learning-based solution is suited to solve the combined problem of patient load prediction and intelligent inter hospital patient transfer.

We point out that in recent times, deep reinforcement learning (Deep-RL) has gained popularity. Deep learning basically models a scenario in which no direct interaction exists between the algorithm and the environment it is operating in. Basically, RL enables a feedback loop between the algorithm and the environment. It allows the algorithm to experience a dataset that varies with time as a result of the interaction with the surrounding environment. RL is applicable to scenarios that can be modeled by a Markov decision process (MDP). Recent works based on RL techniques have shown comparable performance against NNs. In [59], Negnevitsky et al. applied an adaptive neural fuzzy inference system (ANFIS) to study the problem of load forecasting in power systems. They further pointed out the difficulty of load forecasting in power systems resulting from the complexity of the power load series as it exhibits seasonality levels and the fact that power load series have various weather related exogenous variables. In [60], the authors applied principles of RL and game theory to develop an autonomous evacuation process to support distributed and efficient evacuation planning. Deep-RL methods results when deep NNs are used to approximate any of the components of RL, e.g., action-value and/or the policy functions [61]. A combination of NN with RL will enable solving even more complex problems.

Despite the wide range of successes, current state-of-the-art Deep-RL methods still face a number of significant drawbacks [62]. As the training of NNs requires huge amounts of data, Deep-RL demonstrates unsatisfying results in settings where data generation is expensive. Even in cases where interaction is nearly free (e.g., in simulated environments), Deep-RL algorithms tend to require excessive amounts of iterations, which raises their computational and wall-clock time cost. Furthermore, Deep-RL suffers from random initialization and hyperparameter

sensitivity, and its optimization process is known to be uncomfortably unstable [63]. An especially embarrassing consequence of these Deep-RL features turned out to be the low reproducibility of empirical observations from different research groups [64]. The main reason for NN becoming so popular lies in its ability to learn complex and nonlinear relationships that are difficult to model with conventional techniques [65] [66]. Hence our choice of deep learning is based on its many documented exciting achievements [67], which are a function of big data, powerful computation, new algorithmic techniques, mature software packages and architectures and strong financial support.

2.2.1 Deep learning architectures

Deep Neural Networks are a part of the broad field of artificial intelligence (AI). AI is the science and engineering of creating intelligent machines that have the ability to achieve goals like humans do. Figure 2.1 shows the relationship of deep learning to the entire AI [68]. There exist different deep learning structures including; the Deep Boltzmann Machines (DBMs) and Deep Convolutional Neural Networks (DCNN). These architectures can be modeled to help control patient flow systems. The chosen Deep Belief Architecture (DBA) has L layers, including one input layer, one visible output layer and (L-2) hidden layers. The units comprising each layer except for the input layer has its own weight value called bias. The units for two adjacent layers are connected with each other via weighted links while no inner layer connection exists.

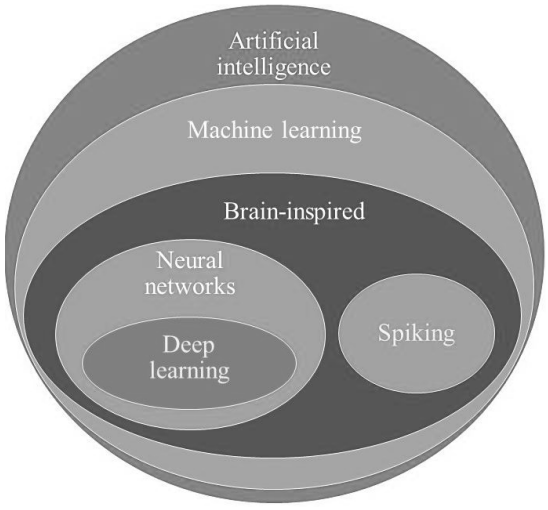

Figure 2.1: Relationship between deep learning and the rest of AI

2.3 System modeling

In this section we describe our system architecture and the training process. We start by presenting the notations used throughout the remainder of this chapter in Table 2.1.

Table 2.1: Notations

Symbol	Meaning
G	A graph- a network of health facilities
E	Edge of a graph G
P	A set of patients
L	A set of hospitals in a particular region
G	Total number of hospitals in a particular region
R	Average number of patients in the catchment area of a hospital
w_{ij}	Weight of link between i and j
b_i	Bias of unit i
ζ	Learning rate
$f(\cdot)$	Activation function

We model the system as a graph $\mathrm{G} = (\mathrm{P} \cup \mathrm{L} \cup \mathrm{C}, \mathrm{E})$, where

- C is the central controller. The central controller can be played by the big health facility within a defined catchment area.
- $\mathrm{L} = \{l_1, l_2, \cdots, l_M\}$ is the set of local hospitals under the control of C.

Given that M is the number of hospitals under consideration, we define the following:
Let $p_m(t)$ be the number of patients at m^{th} hospital at time t. Then;

$$P(t) = \{p_1(t), p_2(t), \cdots, p_M(t)\}, \quad (2.1)$$

is the set of patients over the set of hospitals at time t.

Note that here, M is the total number of hospitals located in a particular region with each hospital serving a population in its catchment area. If we let R to be the average number of patients in

the catchment area of L, i.e., the total catchment area covering the M hospitals, then $L = M \times R$. R is also time variant, since the population of a particular area varies with time. Each local hospital maintains a local database of patient information.

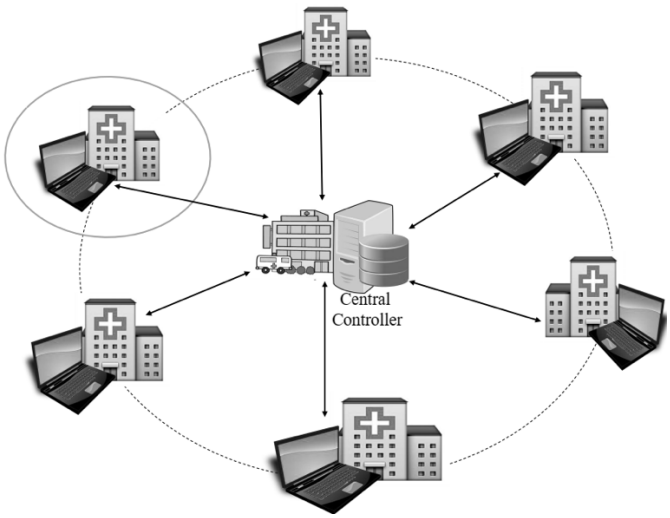

Figure 2.2: A centralized hospital management system – dotted circle defines a network of hospitals, whilst the solid circle defines a catchment area of a particular hospital

Define E to represent the edges in a graph G. Let an edge $e \in E$ in the graph represent the route connecting two nearest hospitals. Then the weight, $w(e)$, represents the obstacles that will impact a patient's ability to get from one facility to another. These obstacles include; the distance, traffic congestion, the transport cost and the physical terrain. Here, we consider the shortest possible road connecting two facilities. In the Figure 2.2, the arrows indicate information flow between the central controller and each of the facilities in it's domain and the solid circular line defines a catchment area, while the dotted circular line defines a network of health facilities under control of Central Controller. From now on we abbreviate Central Controller as Cc.

2.3.1 Training

Let the pair $(x_{\text{input}}, y_{\text{output}})$ represent the values of units in the input and output layers, respectively. Let w_{ij} represent the weight of the link between units i and j, and b_i represents the bias of unit i. Let w be the matrix of weights of all links and b be the matrix of all the bias values, then the training of the Deep Belief Architecture (DBA) comprises two steps, namely; forward propagation and back propagation processes. Here, the forward propagation is used for constructing the structure and activating the output, whilst back propagation is used for adapting

the structure and fine-tuning the values of the weight and bias matrices. Forward propagation process can be modeled as a log-likelihood function and is given as;

$$l(w, b, x_{\text{input}}, y_{\text{output}}) = \sum_{t=1}^{m} log\, p(v^t)\,, \tag{2.2}$$

where v^t denotes the t^{th} training data. Here, the DBA training can be seen as a log-linear Markov Random Field (MRF). As such, $p(v^t)$ represents the probability of (v^t). m is the total number of training data.

The purpose of the training process is to minimize $l(w, b, x_{\text{input}}, y_{\text{output}})$, in the back-propagation process. Back propagation is an efficient way to compute the partial derivatives of the gradient. It is a computation derived from the *Chain Rule* of calculus, and it operates by passing values backwards through the network to compute how the loss is affected by each weight. The link weight w and the bias b are adjusted using the gradient descent method. We represent w and b as;

$$w = w + \zeta \frac{\partial(w, b, x_{\text{input}}, y_{\text{output}})}{\partial w}, \tag{2.3}$$

$$b = b + \zeta \frac{\partial(w, b, x_{\text{input}}, y_{\text{output}})}{\partial b}, \tag{2.4}$$

where ξ is the learning rate of the training process. The Gradient descent is a first-order optimization method, this means it uses only information of the first derivative of the error, hence it can be used in combination with error back propagation. The challenge with this method is the difficult to choose the learning rate so as to get fast learning but at the same time avoid oscillation.

As the input layer increases (i.e., gets larger) and less connected in high dimensions, the DBA structure becomes inefficient as it fails to capture the spatial features efficiently. The CNN tend to be powerful in this regard. The convolution operation extracts features of the input, and the parameters of the convolution operation comprises a set of learnable filters. Letting w_{ij} to denote the filters and the k^{th} filter is represented by w_k^{ll}, then the convolution operation outputs a feature map given as;

$$
\begin{aligned}
u_{i,j,k}^{(l_l)} &= \left(U^{l_l} * W_k^{l_l}\right)_{(i,j)} + w_{b_k}^{l_l} \\
&= \sum_{p=1}^{P} \sum_{m=1}^{M'} \sum_{n=1}^{N'} w_{m,n,p} a_{i+m,j+n,p}^{l_l} + w_{b_k}^{l_l} \\
a_{i,j,k}^{l_l} &= f\left(u_{i,j,k}^{l_l}\right),
\end{aligned}
\tag{2.5}
$$

where $f(\cdot)$ is the activation function and $a_{i,j,k}^{l_l}$ is the activated value of the unit in the i^{th} row and j^{th} column of the feature map. Therefore, $u_{i,j,k}^{l_l}$is the value before activation. $w_{b_k}^{l_l}$ denotes the bias of the k^{th} filter and is usually a single numeric value. $a_{i+m,j+n,p}^{l_l}$ is the activated value of unit in the $(i+m)^{th}$ row and $(j+n)^{th}$ column. The Rectified Linear Units (ReLU) has gained popularity among activation functions. Introduced by [69], ReLU works by thresholding values at 0, i.e., $f(x) = \max(0, x)$ [70].

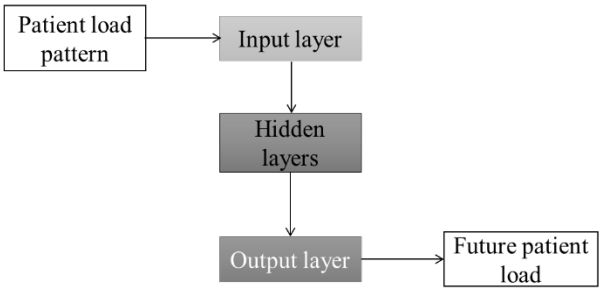

Figure 2.3: Proposed deep learning system

2.4 Proposed deep learning-based patient load prediction model

The proposed deep learning-based system (see Figure 2.3) assumes the existence of a Health Information Exchange (HIE) system that allows health facilities to share healthy relevant information. HIE systems are typically categorized by how patient health information is stored and how the participants can access patient health information [71]. The common HIE models (from here on, we use the words model and system interchangeably) are:

- **Decentralized model-** in this model, each participating health facility controls its data separately in special "edge servers" at a unique location. Patient-specific data is only shared with other participants upon request. In a strictly decentralized model, every request for patient data must be made to every participating data source.
- **Centralized model-** here participants agree to share data and the data is normalized in a common format and terminology and are housed together in a central data repository where they can be accessed and used by participants in line with agreed policies and

procedures. More than one repository may exist for different kinds of data. This model may offer best technical performance in terms of data availability and response time.

- **Hybrid-federated model-** this model is similar to the decentralized model, but it adds a "record locator service" to track patient movement.

In this work, the required information to be shared by the health facilities is patient load information. Hence, the proposed model can suit any of the existing HIE models. However, in this work, two different systems are proposed, namely; a totally centralized system in which all control and computation tasks are handled by central controller and a decentralized control system.

The patient load at a health facility is influenced by the patient rate of arrival. We define the total patient load, PL_{total} to be;

$$\begin{aligned} &PL_{\text{total}} \\ &= \text{patients relayed from other hospitals} + \text{patients from catchment area} \\ &= PL_{\text{rel}} + PL_{\text{reg}} \end{aligned} \tag{2. 6}$$

2.4.1 Centralized patient load prediction system

The centralized prediction system involves four phases, namely; data collection, training, prediction and the online training.

2.4.1.1 Data collection phase

The Cc is responsible for collecting all information of health facilities. Let PS be the patient load sequence of every facility in the last N time slots (or time intervals) recorded by the Cc. Let $T = \{t_1, t_2, \cdots, t_n\}$ be collection of time intervals over which a hospital's traffic load can be predicted. We can assume t_i to be hours over which the hospital is operational, say 7:00 - 17:00. Then let Δ be the length of each time interval. Let ps_k^i be the recorded patient load of facility i in the last time interval k. Then the past patient load PS^i of facility i is given by:

$$PS^i = \left\{ps_k^i, ps_{k-1}^i, \cdots, ps_{k-N+1}^i\right\}. \tag{2. 7}$$

Which is just a length-N vector. Where N is the number of considered past time intervals. Note that N depends on the complexity of input data and is decided according to the training performance. The controller collects all patient load series of every facility, and formats them

as a patient load matrix;

$$PS = \{PS^1, PS^2, \cdots, PS^M\}. \quad (2.\ 8)$$

This can also be expressed as;

$$PS = \{ps_k, ps_{k-1}, \cdots, ps_{k-N+1}\}. \quad (2.\ 9)$$

Which is just a time series expression of the patient loads of all health facilities in the past N time intervals.

The patient load matrix PS is the main output of the data collection phase and is then taken as the input of training data. In the next time slot, the Cc records the patient load as real future patient load as;

$$ps_{k+1} = \{ps_{k+1}^1, ps_{k+1}^2, \cdots, ps_{k+1}^M\}. \quad (2.\ 10)$$

This record is taken as output of the training data. The process is repeated several times and the Cc gathers all such labeled data for training the deep NN in the training phase.

2.4.1.2 Training phase

Consider an M deep CNNs, with each deep CNN responsible only for training the patient load of one health facility. With this structure, the central controller takes only the future patient load of a single health facility as output of corresponding deep CNN. For example, take the training data of deep CNN^i, $(x_{\text{input}}, x_{\text{output}}) = (PS, ps_{k+1}^i)$. Then, the Cc trains all the deep CNNs respectively, so as to obtain all the stable weight matrices.

2.4.1.3 Prediction and accuracy calculation phase

During this phase, the Cc forecasts future patient load and computes the prediction accuracy. The weight matrix of each deep CNN obtained during the training phase is then adopted for predicting the future patient load. For all the deep CNN, the output is recorded as;

$$PSF_{k+1} = \{psf_{k+1}^1, psf_{k+1}^2, \cdots, psf_{k+1}^M\}. \quad (2.\ 11)$$

Recall that the real future patient load of time interval $(k+1)$ is recorded as $PS_{k+1} = \{ps_{k+1}^1, ps_{k+1}^2, \cdots, ps_{k+1}^M\}$. Hence, the prediction accuracy can be computed according to;

$$\frac{1}{K \times M} \sum_{k=0}^{K-1} \sum_{i=0}^{M} \frac{\left| psf_{k+1}^{i} - pl_{k+1}^{i} \right|}{pl_{\max}^{i}}, \tag{2. 12}$$

where K is the total number of time intervals considered and $pl_{\max}^{i}$ is the maximum patient load of hospital i i.e., (daily patient capacity). See Figure 2.4.

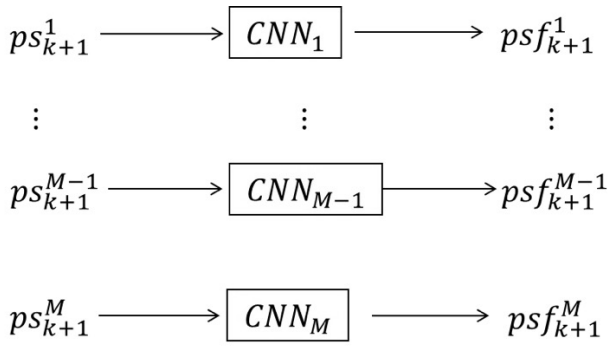

Figure 2.4: Patient load prediction phase in the centralised system

2.4.1.4 Online training phase

If the arrival pattern of patient load acts in a consistent pattern/manner, then the training and prediction processes can reasonably only be based on the existing training data. It is possible however for the arrival pattern of patients to change because of some reasons. Under such situations, it would be important that the training process be adaptive accordingly. This would necessitate the online training phase so as to allow the adjusting of the deep CNNs to adapt to the new pattern. During this phase, each health facility continuously records the patient load data, and the training phase is processed periodically with the collected new training data. Hence the weight matrices are also adjusted periodically.

2.4.2 Decentralized patient load prediction system

Unlike in the centralized system, here the Cc is granted partial computation ability, and that each health facility completes some tasks locally. Each health facility uses local information to make simple pre-forecasting. This lessens the computational burden of the Cc. The final prediction is however done by the Cc, but it integrates the global pre-forecasting information from all local healthy facility. Below we discuss only the data collection and the training phase. The prediction phase and the online phase are the same as in the centralized control system.

2.4.2.1 Data collection

In the centralized system above, the Cc predicts the patient load based on collected patient flow

patterns of all health facilities. However, in the decentralized system, the Cc has limited capabilities, such that each facility does not transfer all the raw patient flow data to the Cc. Here, each facility performs some pre-processing of the raw patient flow data and sends limited/less information to the Cc, hence reducing the computational and signaling overheads of the Cc. Each facility i captures the patient load ps_{k}^{i} of the previous time interval, and also separately captures the relayed patient load PL_{rel}^{i} and the regular patient load from its catchment area as PL_{reg}^{i} of the last N time slots. Using PL_{reg}^{i} as input, each facility i predicts the future regular patient load $psf_{\mathrm{reg}_{k+1}}^{i}$ of the next time slot. The training and prediction process is conducted in the training phase. Thereafter, each facility forwards the obtained $psf_{\mathrm{k+1}}^{i}$ and the recorded patient load of previous time slot $ps_{\mathrm{k+1}}^{i}$ to the Cc. The received information from all health facilities is then constructed by the Cc as the training data ps_k and ps_{reg_k}.

2.4.2.2 Training phase

Here the training phase is split into two steps. The first step involves each health facility training a local NNs to forecast its future integrated patient load with its past N-time-slot regular patient loads. Such that for each facility i, the training data of its local NN can be represented as;

$$\left(x_{\mathrm{input}}, y_{\mathrm{output}}\right) = \left(PL_{\mathrm{reg}}^{i}, psf_{\mathrm{reg}_{k+1}}^{i}\right). \tag{2. 13}$$

Compared with the input to the deep CNN used in the Cc, here the input is simpler such that the training can be treated as a function fitting process between inputs and outputs. Hence, the DBN can be used to perform the training process. As pointed out above (i.e., in data collection), it is the trained DBN that will be used to forecast the future integrated patient load that is denoted as; $psf_{\mathrm{reg}_{k+1}}^{i}$. The outcome of this process is periodically sent to the Cc.

Upon each facility completing self-prediction and sending the result to the Cc, the Cc performs the final prediction with last time interval's patient load ps_k and predicted regular patient load $psf_{\mathrm{reg}_{k+1}}^{i}$ of all health facilities. Since the patient load and regular load are different system features, they can be considered as two separate input data. Hence, the training data input can be formed as a matrix;

$$\left(ps_{\mathrm{k}}, psf_{\mathrm{reg}_{k+1}}^{i}\right) = (\{ps_k^1, ps_k^2, \cdots, ps_k^M\}, \{psf_k^1, psf_k^2, \cdots, psf_k^M\}). \tag{2. 14}$$

As pointed out above, the deep learning structures in the Cc are used for predicting future patient loads of all the health facilities. Just as with the central-control-based prediction, we make use of M deep CNNs to make the prediction to reduce the computational burden and guarantee the accuracy. Hence, for CNN^i, its labeled training data is formed as (see Figure 2.5);

$$\left(x_{\text{input}}, y_{\text{output}}\right) = \left(\left(ps_{\text{k}}, psf_{\text{reg}_{\text{k+1}}}\right), psf^{i}_{k+1}\right). \tag{2. 15}$$

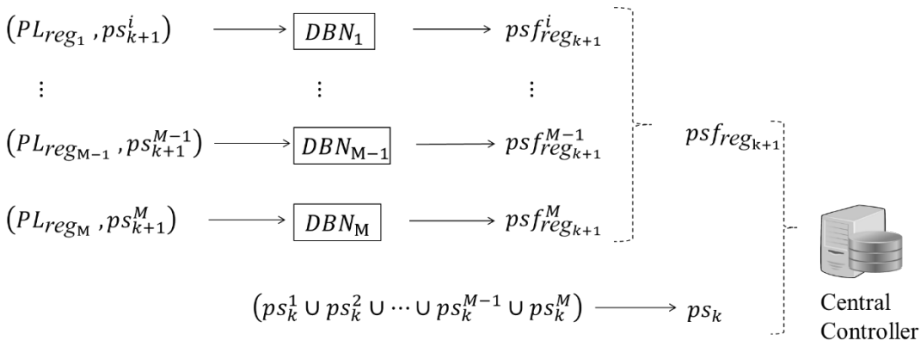

Figure 2.5: Training phase in the decentralised system

2.5 Computational considerations

A common challenge in applying machine learning techniques to problems is the exponential increase of the size of training data and the growing model complexity. It is true that the availability of large sized data requires the use of more complex machine learning models so as to discover finer structures in the data. However, more complex models are often associated with higher computational cost. Furthermore, deeper NNs also bring in more complex computational patterns [72]. Techniques such as, shrinking the model, data parallelism, and model parallelism are being explored as possible solutions to accelerate the performance of deep learning models. In [73] it was shown that about 80% of the computational cost in NNs results from the convolution operation [74]. Hence, recent studies on NN model designs have focused on how to configure convolutional layers [54]. Thus, to control the execution overhead of a NN model, in [75], the authors mathematically formulate the convolutional overhead. Without such formal overhead formulation, a NN model structure is mainly configured based on a designer's experience to balance the execution overhead and the model's accuracy. The designer manually selects a configuration (based on experience or some public models) and then trains the model.

In case of low training accuracy, a different configuration will be tested. Clearly, this approach lacks a systematic way to configure the model. On the contrary, with the execution overhead formulation, this configuration can be quantitative and effective. Denote ϑ as the convolutional overhead, such that;

$$\vartheta \leq \alpha \times \mu, \tag{2. 16}$$

where μ is the preferred or predefined resource budget and α is the percentage of computations due to convolution. Hence, the execution overhead of layer k denoted ϑ_k, needs to satisfy;

$$\vartheta_k = \alpha \times \beta_j \leq \alpha \times \mu \times \beta_j, \tag{2. 17}$$

where β_j indicates the computation percentage of each layer, $j = 1,2,3,\cdots,m$. For a detailed discussion on this method we refer the reader to [76].

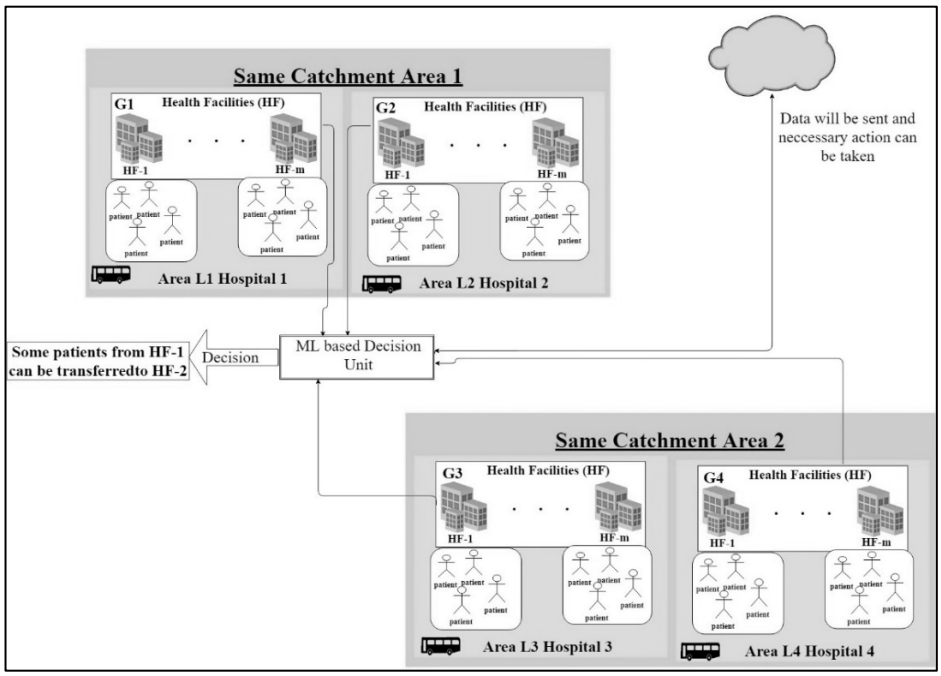

Figure 2.6: A large scale deployment scenario

2.6 Large scale implementation scenario

In Figure 2.6, a sketch for a large-scale deployment scenario is provided. In the figure, L1 represents hospitals in area/region L1. Similarly, L2 to L4. The facilities in L1 are connected via the graph G1. Similarly, the facilities in L2 by G2, L3 by G3 and L4 by G4. The ML based decision unit in conjunction with the cloud part is responsible for coordinating the decision-

making process as regards both the prediction of patient flow and also re-assigning of patients from overcrowded facilities to less crowded facilities.

2.7 Patient transfer logistics

This section presents the development of the IoT-based smart bus system which can support the timely movement of reassigned patients from a bus stop near the source facility to the destination facility. This can help prevent putting a further strain on the already fragmented ambulatory services in many clinics. Since upon arrival patients are triaged and get treated in order of urgency, then in order of arrival, the patients that are re-assigned are the least critical, such that they can manage to carry themselves to the nearest bus station.

Hospital logistics generally differentiate between three main flows: patient, information and material flows. The patient flows run in three directions: Towards the hospital (or inbound), within the hospital (or internal) and away from the hospital (or outbound). As pointed out above, the goals of the above proposed system are to predict future patient loads. Furthermore, in the case of excess patient loads, we propose that excess loads should be re-assigned to nearby facilities that have a minimal load as a way to control overcrowding and reduce the number of patients that leave health facilities without receiving medical care. However, the re-assigning of patients will imply a need for transportation for the patient to move from one facility to another. The two most commonly employed modes of transfer of patients are ground transport, with the inclusion of ambulances, and mobile intensive care units (MICUs) [77]. In this work we assume the existence of a scheduled bus transport system. A bus journey involves such stages as waiting, queuing and transferring from the origin point to the final destination. These stages are impacted by the services rendered and transportation network resources. Three stages exist at which transit services can affect passengers [78]. These stages are:

- **Origin-Point Stage-** at this level, passengers wait for the next bus. Passenger waiting time can be longer than expected due to irregular headway or technical issues that might affect the bus. Also, not all passengers might board the arriving bus due to capacity limitation. Upon being quarried, the proposed smart bus system will provide passengers with information regarding the position of the bus, travel time estimate, and information about the number of vacant seats in the bus. This information will help passengers to plan their waiting time accordingly.
- **Boarding Stage-** at this stage, the passenger services time (PST) is a function of

passenger demand. Boarding passenger's wait-time must allow passengers in the bus to alight. Here, we assume that near each health facility there is a boarding stage.

- **Arrival Stage-** arrival stage is the stage where passengers reach their final destination. Their arrival can be on time, ahead or delayed based on the deviation from timetable. For the patients, the arrival stage will be the stage near the destination facility where the patient has been referred to.

Various solutions exist that are used to track data related to departure, and arrival times of buses. Radio Frequency (RF) transceivers are installed on buses and bus stops to enable buses communicate the location to bus stops [79]. The estimated arrival time of the bus is calculated by microprocessors at the bus stops. The calculated time is displayed on screens that are installed at the stages. Some solutions employ SMS services on GSM modules to transmit bus positions to databases. GPS tracking devices are mounted on buses to provide location data. Sending of data to the database is via the Hyper Text Transfer Protocol [80] [81]. Alternatively, location data can be streamed from android devices in the bus. The bus location and travel time approximation can be accessed through android devices or web-portals. Unfortunately, some of these systems cannot handle an increase in requests. The bandwidth demands for sending the data from buses to servers is high. Hence there is need to explore scalable and efficient solutions [82].

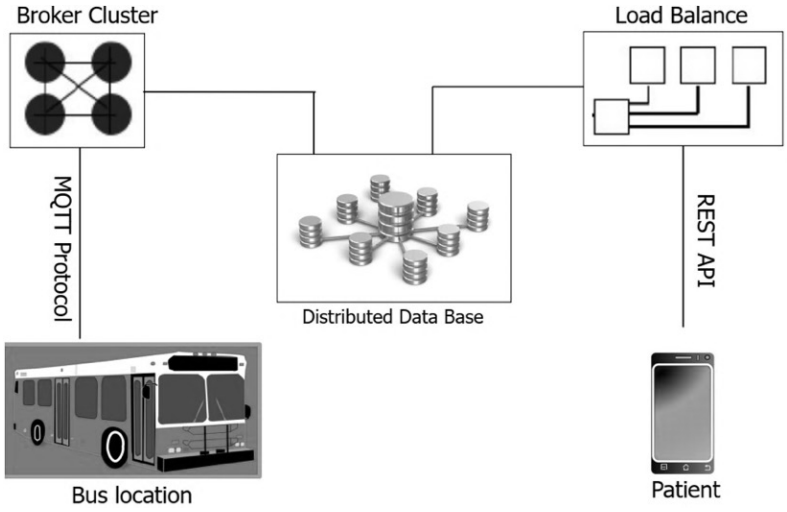

Figure 2.7: Smart bus system architecture

2.7.1 System architecture

The overall system architecture of the proposed smart transport system is shown in Figure 2.7. Each public bus is attached with an IoT smart transport kit which comprises a variety of modules

including; the GPS module, two sets of Laser light with Light Dependent Resistor (LDR) positioned at the back and front entryway of the bus respectively, and a smart phone that acts as a mobile hotspot. Both the GPS and Laser light with LDR reads and feeds real-time data to an Arduino Uno microcontroller. The Arduino Uno connects to the Internet via NodeMCU that sends the collected data to a MQTT broker on the cloud using the Wi-Fi provided by the mobile hotspot that provides Internet connection. Since the smart buses will also be servicing patients, the system will need to be reliable. The patients need to accurately know the arrival time of the bus at a bus stop and the time it will reach the destination and also the status on the capacity of the bus (whether it is full or there are vacant seats in the bus). Using the real time location of buses, it is possible to calculate the approximate time for the bus to reach a particular position. This is possible since a bus's location data can be saved on the server together with corresponding timestamps. To count the number of passengers alighting and boarding, we make use of the laser beam. The choice of the laser is motivated by the fact that it does not scatter and is invisible. Since passengers queue and enter or exit the bus one by one, with the laser, the system will not miss any passenger. Hence the counter will be reliable.

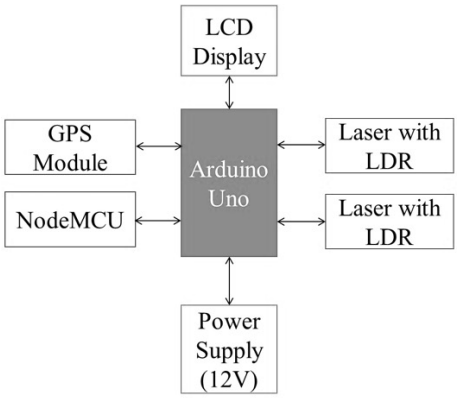

Figure 2.8: Framework of developed IoT based smart bus kit

As pointed out above, each of the public buses will be tagged with an IoT based smart bus system. Figure 2.8 shows the framework of the developed IoT kit. We briefly describe the function of each of the components of the system.

- **Arduino Uno-** This is the heart of our framework. It works as a CPU unit. Arduino Uno will generate sensor data that will have to the send to the server. This data will include, the location data of the bus, time and distance approximation of departure and arrival and also information about the capacity of the bus. Arduino will take these data and send to Nodemcu over a serial connection.

- **Laser-** A laser is a device emitting light through optical amplification based on the stimulated emission of electromagnetic radiation. The term "laser" originated as "light amplification by stimulated emission of radiation" [83]. The primary wavelengths of laser radiation for most current applications include the ultraviolet, visible, and infrared regions of the spectrum. Ultraviolet radiation for lasers consists of 180 and 400 nm wavelength. The visible region lies between 400 nm and 700 nm wavelength. The infrared region of the spectrum consists of radiation with 700 nm and 1 mm wavelength. When the intensity of the radiation is sufficiently high, damage to the absorbing tissue may happen [84] [85]. In the proposed system, the primary functionality of the laser technology is to detect motion (i.e., entry and exist of passengers) and serve as counter to determine the number of vacant seats in the bus. The counting starts when there is a discontinuity in the laser beam falling on the sensor. The discontinuity is detected at the entrance (or exit) of the door of the public buses at the instant of time where the laser beam is not being felt on the sensor. Here we utilize the laser technology whose path is invisible.
- **NodeMCU-** NodeMCU is a development board that incorporates the ESP8266 Wi-Fi chip. The ESP8266 is programmable like any other microcontroller. Since Arduino Uno does not have network capabilities, the NodeMCu interfaces the Arduino Uno microcontroller and connects the system to the Internet and drives the output for the GPS and the lasers. The Wi-Fi connection is provided by the mobile hotspot that is fixed in each bus. An Application Server implemented using Node.js collects the data from the MQTT broker through publish/subscribe mechanism and saves the data using the NSQL Couch DB database.
- **GPS Receiver-** GPS is a space-based radio-navigation system that broadcasts highly accurate navigation pulses to users on or near the Earth. This system is used to collect the real-time co-ordinates for the system. The GPS receives the satellite signals and then the position coordinates with latitude and longitude are determined by it. The location is determined with the help of GPS and transmission mechanism. After receiving the data the tracking data can be transmitted using any wireless communications systems. These units are connected at the output of the system. GPS port is serially connected with the system.

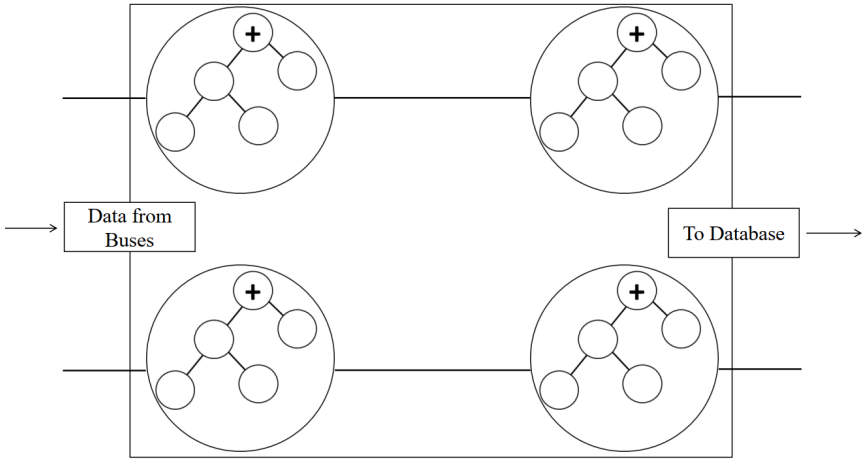

Figure 2.9: MQTT clusters- every node maintains tree of topics

2.7.2 MQTT protocol

The Message Queuing Telemetry Transport (MQTT) is a Client Server publish/subscribe messaging transport protocol. It is light-weight, open, simple, and easily implementable. MQTT is ideal for use in IoT applications that require a small code footprint and/or network bandwidth is at a premium. The protocol runs over network protocols that provide ordered, lossless, bi-directional connections, e.g., TCP/IP. In the MQTT protocol there is a broker and a receiver. The broker has 'topics' through which a recipient of data generated by a sender is determined. To receive the data, a receiver subscribes to a particular topic. A receiver receives the data published by any sender with that particular topic [86]. In the proposed system, the topic will be the route number/name, such that the published location data from buses will contain this information. The location data will comprise the current latitude and longitude of a bus and will be published at a frequency of 5 seconds. To ensure scalability of the system, the MQTT-brokers can be clustered as in Figure 2.9. Although clustered, but the brokers function as one, with each maintaining the same topic list. In case of one broker crushing, another broker automatically handles the requests, balancing within themselves. The advantage of clustering is assured system availability. For each route number, the location data is received via MQTT and stored in the database. The broker can be directly modified to write to the database upon reception of data or a MQTT client can be created to perform this function.

2.7.3 Data access by users

Various platforms can be used to achieve user access to the system including; web, smartphones or SMS. The proposed system allows users to access the data via REST APIs. HTTP requests are sent to the server and the server returns fetched data based on user context. The location

information can easily be shown to users via Google maps or an equivalent map service. Through a simple application, patients and general passengers can easily search a bus by bus number directly. In case the user is not aware of the bus number, he/she will be asked to first select the nearest bus stage (i.e., the boarding point). The process can easily be automated by keeping a database of bus stops along with their corresponding location data. To calculate distance to the system compares a user's location to that of the bus stops applies the Haversine formula as;

Lat_1: User latitude

Long_1: User longitude

Lat_2: Bus stop latitude

Long_2: Bus stop longitude

$\mathrm{d}_{\mathrm{Long}} : \mathrm{Long}_2 - \mathrm{Long}_1$

$\mathrm{d}_{\mathrm{Lat}} : \mathrm{Lat}_2 - \mathrm{Lat}_1$

r: Earth's Radius. Then

$$H = \left(\sin\left(\mathrm{d}_{\mathrm{Lat}}/2\right)\right)^2 + \cos\left(\mathrm{Lat}_1\right) \times \cos\left(\mathrm{Lat}_2\right) \times \left(\sin\left(\mathrm{d}_{\mathrm{Long}}/2\right)\right)^2$$
$$B = 2 \times \tan -1\left(\sqrt{H}, \sqrt{1-H}\right) \qquad (2.18)$$

Distance= $B \times r$.

From equation. 2.18 above, the nearest bus stop will be the one with shortest distance. The live location of a bus is queried based either on target destination or bus number. If the user enters a bus number, the user API returns the last known bus location from central buses collection. To allow mapping of the boarding stop-destination point to route number, a database of all buses on all routes can be maintained. Hence, if a user selects boarding point and destination point, he/she gets live location of all buses on route that takes him/her to the destination.

In Table 2.2 above, the first row shows the route numbers, whilst the corresponding columns lists bus stops on that route. For example, if a passenger queries data by boarding and destination bus stages, the query is then checked in the table. Take stops B and A where B is the boarding stage and A is the destination bus stage. The system simply checks the table to verify if bus stop B occurs in the table and is followed by A. However, bus Stop A need not immediately follow B. In the Table 2.2, we see that routes number 101, 201, 301 have such condition true. The query

will then fetch the live locations of buses on these three routes and return them to the user. The queries from the user module will be same across all platforms.

Table 2.2: Table of bus stops along routes

101	**201**	**301**	**401**	**501**
B	B	A	S	A
C	H	B	D	C
E	R	H	G	D
D	L	G	F	S

The output for the HTTP GET request will take the following formats.

- **Query by registration number-** the query by registration number will return a JSON response fetching data from ‘buses’ collection with the parameters; latitude, longitude, bus registration number, bus route number, direction and time stamp.
- **Query by route number-** the query by route number will return a JSON array containing objects, with each of the objects giving details of active buses running on that route. The data will be fetched from a collection with the name 'route number'.
- **Query by boarding point and destination-** This query will return a JSON array containing objects for active buses on every route taking from the boarding point to the destination. This data will be fetched from various collections based on which routes are applicable.

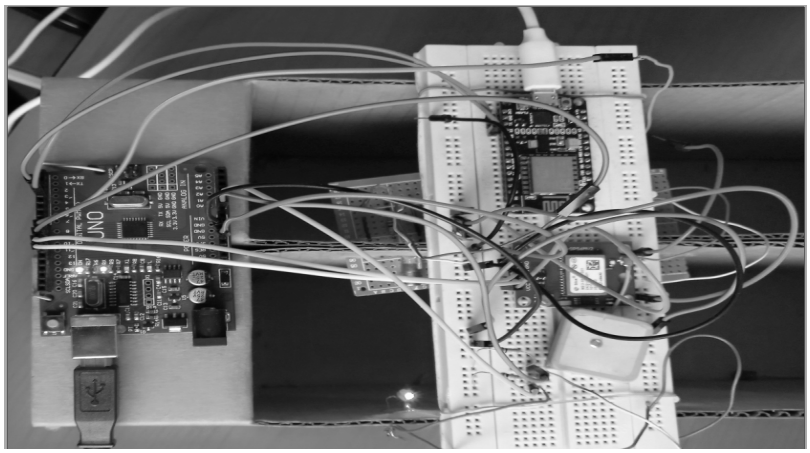

Figure 2.10: Developed IoT based smart public bus system

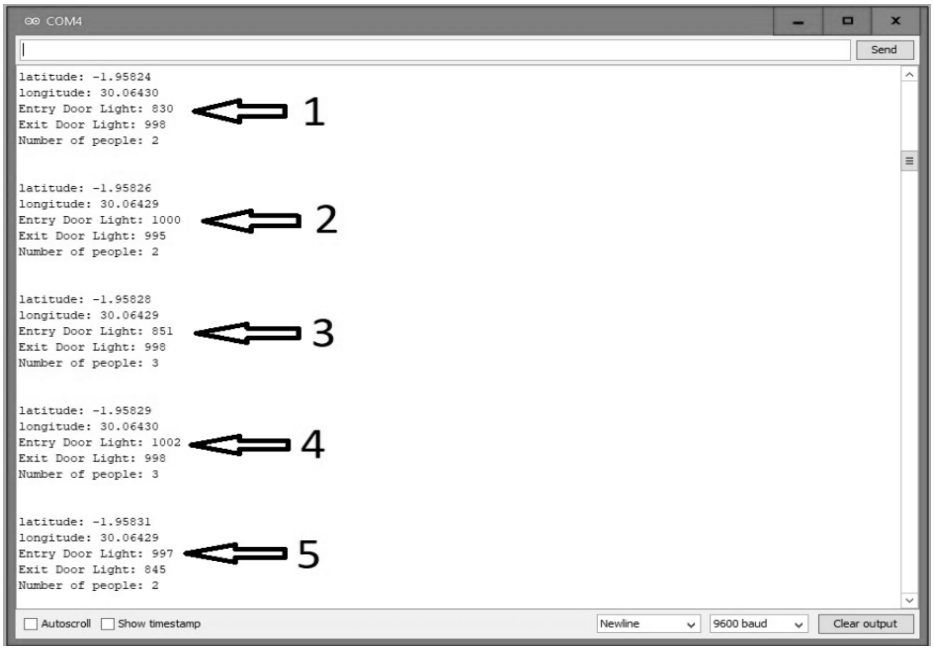

Figure 2.11: Arduino Uno serial monitor displaying output data

2.7.4 Functionality principle of the counter

Figure 2.10 shows the developed IoT based smart public bus system and Figure 2.11 is the snapshot showing the system transmitting captured data. As pointed out above, the LDR (light-dependent resistor) detects the increase in the intensity of the light if the laser is pointed directly to it. When there is an obstacle between the laser and the LDR, the LDR detects a decrease in the intensity of the light. Hence by counting each decrease in the range of the optimum level, the system is able to count passengers entering and exiting the smart bus. From the Figure 2.11, as per our simulations, the maximum light intensity of both the entry and exit doors' LDR is set at ≥ 900 such that any value of < 900 signifies an obstacle and in our case, this will signify a passenger leaving or entering the smart bus. From 2 in Figure 2.11, we see that the number of passengers in the bus is 2 and the light intensity on the LDR for the entry and exit doors is 1000 and 995 respectively. Since both are above 900, this means there is not entry or exit of passengers from the smart bus. In 3 of Figure 2.11, we observe that the light intensity for the entry door is < 900 and this decrease correspond to the increase in the number of passengers which has increased from 2 to 3. For this simulation, the smart kit was stationary, hence the fixed latitude and longitude values are; -1.95824 and 30.06430 respectively.

2.8 Conclusion

The need to access quality health services is a basic human need. However, the health systems

are experiencing many challenges that make meeting its goals of offering quality healthcare service a challenge. A mismatch between staffing ratios and service demand results in the overcrowding of patients in hospital waiting rooms leads to excessive patient waiting times, incomplete preventive service delivery and disgruntled medical staff. It is important that health facilities coordinate their efforts to ensure that all patients get the needed health care as and when they need it. In this chapter, a coordinated approach between the health facilities has been proposed as a possible solution to deal with the problem of excess patient load and overcrowding. Building on the power of deep learning, a patient load prediction model that can be used to predict future patient loads is presented. Furthermore, to avoid putting a strain on the already fragmented ambulatory service in health facilities, we propose an approach that makes use of the existing public transport system to support the healthcare delivery system. A conceptual framework for an IoT-based smart bus transport system that can support timely delivery of health services has been presented. Using off-the-shelf sensors, an IoT smart transport kit is developed. This kit can be attached to ordinary public buses allowing patients and general passengers to query bus location and time related information to enable them move from one health facility to another. In this work, we have only focused on the transfer of first time patients from one health facility to another. Our future work will focus on scheduled patient arrivals and also the transfer of patient health files since a patient's past record may have a bearing on his/her current health status.

Author Contributions: The ideas presented in this work are a product of contributions by all the authors. The conceptualization of the idea and the development of the manuscript was done by author K.M. in consultation with the supervision team comprising authors S.K., C.M., K.J. and J.N. The supervision team played a key role in providing the needed advise and direction throughout the development process of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] M. Bahadori, E. Teymourzadeh, R. Ravangard and M. Raadabadi, "Factors affecting the overcrowding in outpatient healthcare," *Journal of Educstion and Health Promotion,* vol. 6, 2017.

[2] M. Yarmohammadian, F. Rezael, A. Haghshenas and N. Tavakoli, "Overcrowding in emergency departments: a review of strategied to decrease future challenges," *J Res Med Sci.,* vol. 22, 2017.

[3] T. Tran, U. Nguyen, N. Minh and B. Tran, "Patient waiting time in the ourpatient clinic at a central surgical hospital of Vietnam: implications for resourse allocation," *F1000 Research,* vol. 6, 2017.

[4] Y. Shukla, R. Tiwari, B. Rohit and P. Kasar, "An assessment of OPD registration counter services and channelization of patients in NSCB medical collehe hospital," *Int J Med Sci Public Health,* vol. 4, pp. 1468-1472, 2015.

[5] R. Obulor and B. Eke, "Outpatient queueing model development for hospital appointment system," *IJSEAS,* vol. 2, pp. 15-22, 2016.

[6] C. Alexopoulos, D. Goldsman, J. Fontanesi, D. Kopald and J. WIlson, "Modeling patient arrivals in community clinics," *Omega,* vol. 36, no. 1, pp. 33-43, 2008.

[7] J. Fontanesi, M. Guire, J. Chiang, K. Holcomb and M. Sawyer, "Application of workflow analuysis tools in outpatient primary care settings," *Jt Comm J Qual Improv.,* vol. 26, pp. 654-660, 2000.

[8] E. Zarei, A. Daneshkohan, R. Khabiri and M. Arab, "The effect of hospital service quality on patient's trust," *Iran Red Crescent Med J.,* vol. 17, 2014.

[9] J. Sun, Q. Lin , P. Zhao, Q. Zhang, K. Xu, H. Chen, C. Hu, M. Stuntz, H. Li and Y. Li, "Reducing waiting time and raising outpatient satisfaction in a Chinese public tertiary general hospital - an interrupted time series study," *BMC Public Health,* vol. 17, p. 668, 2017.

[10] T. Dinesh, S. Singh, P. Nair and T. Remya, "Reducing waiting time in outpatient services of large university teaching hospital - a six sigma approach," *Management in Health,* vol. 17, no. 1, 2013.

[11] A. Kulshrestha and J. Singh, "Inter-hospital and intra-hospital patient transfer: recent concepts," *Indian J Anesth,* vol. 60, pp. 451-457, 2016.

[12] L. Sokol-Hessner, A. A. White, K. F. Davis, S. J. Herzig and S. F. Hohmann, "Interhospital transfer patients discharged by academic hospitalists and general internists: characteristics and outcomes," *J Hosp med.,* vol. 11, no. 4, pp. 245-250, 2016.

[13] A. A. Owad, P. Samaranayake, A. Karim and K. B. Ahsan, "An integrated lean methodology for improving patient flow in an emergency department-case studyof a Saudi Arabian hospital," *The Management of Operations,* vol. 29, no. 13, 2018.

[14] R. W. Derlet, "Overcrowding in emergency departments: increased demand and decreased capacity," *Ann Emerg Med.,* vol. 39, no. 4, pp. 430-432, 2002.

[15] A. Gajanan, "Reduced wait time prediction in hospital emergency room: lean analysis using a random forest model," Knoxville, TN, USA, 2017.

[16] P. Sarah, G. Shaun and N. H. Shah, "Predicting emergency department visits," in *AMIA Jt Summits Transl Sci Proc*, 2016.

[17] Y. Lv, Y. Duan, W. Kang, Z. Li and F.-Y. Wang, "Traffic flow prediction with bug data: a deep learning approach," *IEEE Transactions on Intelligent Transportation Systems,* vol. 16, no. 2, pp. 865-873, 2015.

[18] A. Gal, A. Mandelbaum, F. Schnitzler, A. Senderovich and M. Weidlich, "Traveling time prediction in scheduled transportation with journey segments," *Information Systems,* vol. 64, pp. 266-280, 2017.

[19] B. McCarthy, K. Propp and A. Cohen, "Learning from health infromation exchange technical archtecture and implementation in seven beacon communities," *EGEMS,* vol. 2, 2014.

[20] Google, "Google Maps," [Online]. Available: https://cloud.google.com/maps-platform/. [Accessed 3 December 2018].

[21] K. G. Srinivasa, B. J. Sowmya, A. Shikhar and A. Sign, "Data analytics assited internet of things towards building intelligent healthcare monitoring systems: IoT for healthcare," *Journal of Organizational and End User Computing,* vol. 30, no. 4, pp. 83-103, 2018.

[22] M. Chen , L. Liang, Y. Chang and W. Juang, "Emergency department overcrowding: quality improvement in a Taiwan medical center," *J. Formos. Med. Assoc.,* vol. 118, no. 1 Pt 1, pp. 186-193, 2018.

[23] D. Sethi and S. Subramanian, "When place and time matter: how to conduct safe inter-hospital transfer of patients," *Saudi J Anaesth,* vol. 8, pp. 104-113, 2014.

[24] B. Hill, "Optimization of interhospital transfer of patients to reduce emergency department overcrowding," [Online]. Available: https://scinapse.io/papers/1734556231. [Accessed 4 September 2018].

[25] D. Santillan, "Uses of satisfaction data: report on improving patient care," *Soc Sci Med.,* vol. 16, pp. 24-26, 2000.

[26] T. Dan and C. Qualls, "Time series forecasts of emergency department patient volume, lenth of stay, and acuity," *Ann Emerg Med.,* vol. 23, pp. 299-306, 1994.

[27] Z. Rotstein, R. Wilf-Miron, B. Lavi, A. Shahar, U. Gabbay and S. Noy, "The dynamics of patient visits to a public hospital ED: a statistical model," *Am J Emerg Med.,* vol. 15, pp. 596-599, 1997.

[28] H. Batal, J. Tench, S. McMillan, J. Adams and P. Mehler, "Predicting patient visits to an urgent care clinic using calendar variables," *Acad Emerg Med off J Soc.,* vol. 8, pp. 48-53, 2001.

[29] B. Reis and K. Mandi, "Time series modeling for syndromic surveillance," *BMC Med Inform Decis Mak.,* vol. 3, pp. 1-11, 2003.

[30] J. Brillman, T. Burr, D. Forslund, E. Joyce, R. Picard and E. Umland, "Modeling emergency department visit patterns for infectiousd disease complaints: results and application to disease surveillance," *BMC Med Infrom Decis Mak.,* vol. 4, pp. 1-14, 2005.

[31] T. Flottemesch, B. Gordon and S. Jones, "Advanced statistics: developing a formal model of emergency department census and defining operational efficieny," *Acad Emerg Med off J Soc Acad Emerg Med,* vol. 14, pp. 799-809, 2007.

[32] J. Boyle, M. Wallis, M. Jessup, J. Crilly, J. Lind, P. Miller and G. Fitzgerald, "Regression forecasting of patient admission data," *30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society,* pp. 3819-3822, 20-25 August 2008.

[33] S. Au-Yeung, U. Harder, E. Mccoy and W. Knottenbelt, "Predicting patient arrivals to an accident and emergency department," *Emerg Med J.,* vol. 26, pp. 241-244, 2009.

[34] H. Kam, J. Sung and R. Park, "Predicting of daily patient numbers for a regional emergency medical center using time series analysis," *Healthc Inform Res,* vol. 16, pp. 158-165, 2010.

[35] M. Capan, S. Hoover, E. Jackson, D. Paul and R. Locke, "Time series analysis for forecasting hospital census: application to the neonatal intensive care unit," *Appl Clin Inform.,* vol. 7, pp. 275-289, 2016.

[36] B. Morzuch and P. Allen, "Forecasting hospital emergency department arrivals," *26th Annual Symposium on Forecasting,* 11-14 June 2006.

[37] L. Schweigler, J. Desmond, M. McCarthy, K. Bukowski, E. Ionides and J. Younger, "Forecasting models of emergency department crowding," *Acad Emerg Med.,* vol. 16, pp. 301-308, 2009.

[38] I. Marcilio, S. Hajat and N. Gouveia, "orecasting daily emergency department visits using calendar variables and ambient temperature readings," *Acad Emerg Med.,* vol. 20, pp. 769-777, 2013.

[39] K. Kim, C. Lee, K. O'Leary, S. Rosenauer and S. Mehrotra, "Predicting patient volumes in hospital medicine: a comparative study of different time series foerecasting methods," Northwestern University, Evanston, II, USA, 2014.

[40] W. Pan, "A newer equal part linear regression model: a case study of the influence of educational input on gross national income," *Eurasia J Math Aci Technol Educ.,* vol. 13, pp. 5765-5773, 2017.

[41] S. Jones, A. Thomas, R. Evans, S. Welch, P. Haug and G. Snow, "Forecasting daily patient volumes in the emergency department," *Acad Emerg Med.,* vol. 15, pp. 159-170, 2008.

[42] C. Aladag and S. Aladag, "Forecasting the number of outpatient visits with different activation functions," *Adv Time Ser Forecast,* pp. 26-33, 2012.

[43] M. Xu, T. Wong and K. Chin, "Modeling daily patient arrivals at emergency department and quantifying the relative importance of contributing variables using artificial neural network," *Decision Support Systems,* vol. 54, no. 3, pp. 1488-1498, 2013.

[44] K. Srikanth and D. Arivazhagan, "An efficient patient inflow prediction model for hospital resource management," *IJEECS,* vol. 7, pp. 809-817, 2017.

[45] P. Malhotra, L. Vig, G. Shroff and P. Agarwal, "Long short term memory networks for anomaly detection in time series," in *23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, Bruges, 2015.

[46] J. D. Hamilton, Time series analysis, Princeton: Princeton University Press, 1994.

[47] M. E. Azoff, Neural network time series: forecasting of financial markets, New York: John Wiley & SOns, Inc., 1994.

[48] H. Lee, Y. Largman, P. T. Pham and A. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *NIPS' 09*, 22nd International Conference on Neural Information Processing Systems, 2009.

[49] K. Suzuki, T. Kiryu and T. Nakada, "Fast and precise independent component analysis for high field fMRI time series tailored using prior information on spatiotemporal structure," *Human Brain Mapping,* vol. 15, no. 1, pp. 54-66, 2002.

[50] T. Kuremoto, S. Kimura, K. Kobayashi and M. Obayashi, "Time series forecasting using a deep belief network with restricted Boltzmann machines," *Neurocomputing,* vol. 137, pp. 47-56, 2014.

[51] J. T. Turner, "Time series analysis using deep feed forward neural networks," University of Maryland, Baltimore County, 2014.

[52] M. Langkvist, "Modeling time series with deep networks," Orebro University, Orebro, Sweden, 2014.

[53] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," *Nature,* vol. 521, pp. 436-444, 2015.

[54] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM,* 2017.

[55] J. Torres, A. Troncoso, I. Koprinska, Z. Wang and F. Martinez-Alvarez, "Deep learning for big data time series forecasting applied to solar power," in *In Advanced in Intelligent Systems and Computing* , Cham, Switzerland, 2018.

[56] W. Bao, J. Yue and Y. Rao, "A deep learning framework for financial time series using stacked autoencoders and long-short term memory," *PLoS ONE,* vol. 12, p. e0180944, 2017.

[57] "Video analysis to detect suspicious activity based on deep learning," [Online]. Available: https://medium. com/@everisUS/video-analysis-to-detect-suspicious-activity-based-on-deep-learning-fee2032ea14a. [Accessed 5 may 2019].

[58] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang and G. Penn, "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition," in *2012*

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, 2012.

[59] M. Negnevitsky, P. Mandal and A. Srivastava, "Machine learning applications for load, price and wind power prediction in power systems," in *In Proceedings of the 15th International Conference on Intelligent System Applications to Power Systems*, Curitiba, Brazil, 2009.

[60] G. Fragkos, P. Apostolopoulos and E. Tsiropoulou, "ESCAPE: evacuation strategy through clustering and autonomous operation in public safety systems," *Future,* vol. 11, p. 20, 2019.

[61] Y. Li, "Deep reinforcement learning: an overview," [Online]. Available: https://arxiv.org/pdf/1701.07274.pdf. [Accessed 7 August 2019].

[62] S. Ivanov and A. D'yakonov, "Morden reinforcement learning algorithms," *ArXiv,* 2019.

[63] A. Irpan, "Deep reiforcement learning doesn't work yet," [Online]. Available: https://www.alexirpan.com/ 2018/02/14/rl-hard.Html . [Accessed 23 November 2018].

[64] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup and D. Meger, "Deep reinforcement learning that matters," in *In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, LA, USA, 2018.

[65] H. Hippert, C. Pedreira and R. Souza, "Neural networks for short-term load forecasting: A review and evaluation," *IEEE Trans Power Syst,* vol. 16, pp. 44-55, 2001.

[66] T. Senjyu, P. Mandal, K. Uezato and T. Funabeshi, "Next day load curve forecasting using hybrid correction method," *IEEE Trans Power Syst,* vol. 20, pp. 102-109, 2005.

[67] T. Jose, T. Alicia, K. Irene, Z. Wang and M. Francisco, "Deep learning for big data time series forecasting applied to solar power," in *In International Joint Conference SOCO'18-CISIS'18-ICEUTE'18*, San Sabastan, Spain, 2018.

[68] I. Goodfellow, Y. Bengio and A. Courville, Deep learning, Boston, MA, USA: MIT Press, 2016.

[69] V. Sze, Y.-H. Chen, T.-J. Yang and J. S. Emer, "Efficient processing of deep neural networks: a tutorial and survey," *Proceedings of IEEE,* vol. 105, no. 12, pp. 2295-2329, 2017.

[70] H. Richard , S. Rabul, A. Misha, J. Rodney and S. Seung, "Digital selection and analogue amplification coexist in a cortex-inspired Silicon circuit," *Nature,* vol. 405, pp. 947-951, 2000.

[71] A. F. Agarap, "Deep learning using rectified linear units (ReLU)," *Neural and Evolutionary Computing,* 2019.

[72] H. Ji, S. Yoon, E. Heo, H. Hwang and J. Kim, "Technology and policy challenges in the adoption and operation of health information exchange systems," *Healthc Inf Res.,* vol. 23, pp. 314-321, 2017.

[73] M. Li, "Scaling distributed machine learning with system and algorithm co-design," [Online]. Available: https://www.cs.cmu.edu/~muli/file/mu-thesis.pdf . [Accessed 5 August 2019].

[74] L. Huynh, Y. Lee and R. Balan, "Deepmon: mobile GPU-based deep learning framework for continuous vision applications," in *In Proceedings of the 15th Annual International Conference on Mobile Systems, APplications and Services*, Niagara Falls, NY, USA, 2017.

[75] K. SImonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv,* 2014.

[76] Yang Kang, X. Gong, Y. Liu, Z. Li, T. Xing, X. Chen and . D. Fang, "cDeeparch: A compact deep neural network archtechture for mobile sensing," in *2018 15th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, Hong Kong, 2018.

[77] "https://www.medica-tradefair.com/," Medica magazine, 01 February 2018. [Online]. Available: https://www.medica-tradefair.com/en/News/Topic_of_the_Month/Order_Topics_of_the_Month/Topics_of_the_Month_2018/Hospital_Logistics/Everything_flows_transportation_and_material_flows_in_hospital_logistics. [Accessed 30 January 2021].

[78] K. Lu, B. Han and X. Zhou, "Smart urban transit systems: from integrated framework to interdisciplinary perspective," *Urban Rail Transit,* vol. 4, pp. 49-67, 2018.

[79] G. Chheda, N. Gajra, M. Chhaya, J. Deshpande and S. Gharge, "Real time bus monitoring and passenger information system," *Internationa Journal of Soft Computing and Engineering,* vol. 1, no. 6, 2012.

[80] R. Muruthi and C. Jayakumari, "SMS based bus tracking system using open source technologies," *International Journal of Computing Applications,* vol. 86, no. 9, pp. 44-46, 2014.

[81] K. A. Salim and I. M. Idrees, "Design and implementation of web-based GPS-GPRS vehicle tracking system," *International Journal of Computet Science and Information Technology,* vol. 3, no. 12, pp. 443-448, 2013.

[82] J. Lohokare, R. Dani, S. Sontakke and R. Adhao, "Scalable tracking system for public buses using IoT technologies," in *2017 International Conference on Emerging Trends and Innovation in ICT (ICEI)*, Pune, 2017.

[83] C. Woodford, "https://www.explainthatstuff.com/," 24 September 2020. [Online]. Available: https://www.explainthatstuff.com/lasers.html. [Accessed 30 January 2021].

[84] "https://ehs.berkeley.edu/," UC Berkeley, [Online]. Available: https://ehs.berkeley.edu/laser-safety/non-ionizing-radiation-safety-manual. [Accessed 30 January 2021].

[85] K. H. Mild, R. Lundstron and J. Wilen, "Non-ionizing radiation in Swdish health care-exposure and safety," *Int. J. Environ. Res. Public Health,* vol. 16, no. 7, 2019.

[86] "http://docs.oasis-open.org/," 29 October 2014. [Online]. Available: http://docs.oasis-open.org/mqtt/mqtt/v3.1.1/os/mqtt-v3.1.1-os.html. [Accessed 30 January 2021].

Chapter 3: Adaptive Staff Scheduling at Outpatient Department of Ntaja Health Center in Malawi- A Queuing Theory Application

Accurate staff scheduling is crucial in overcoming the problem of mismatch between staffing ratios and demand for health services which can impede smooth patient flow. Patient flow is an important process towards provision of improved quality of service and also improved utilization of hospital resources. However, extensive waiting times remains a key source of dissatisfaction with the quality of health care service among patients. With rarely scheduled hospital visits, the in-balance between hospital staffing and health service demand remains a constant challenge in sub-Saharan Africa. Accurate workload predictions help anticipate financial needs and also aids in strategic planning for the health facility. Using a local health facility for a case study, we investigate problems faced by hospital management in staff scheduling. We apply queuing theory techniques to assess and evaluate the relationship between staffing ratios and waiting times at the facility. Specifically, using patient flow data for a rural clinic in Malawi, we model queue parameters and also approximate recommended staffing ratios to achieve steady state leading to reduced waiting times and consequently, improved service delivery at the clinic.

3.1 Introduction

Healthcare service provision depends on a complicated array of political, social, epidemiological, demographic and economic factors. The requirements for health care services and associated medical staff across a given country varies depending on variations in mortality, age, sex, and population density; education and wealth; patient visits patterns; terrain; and the ease of access to these services. Hence, it is important that the allocation and utilization of medical staff takes into consideration these variations within a country and not solely depend on population or institutional size [1]. Failure to consider these variations may result not only in under or over-provision of healthcare staff, but may also lead to inappropriate allocation of different cadres of staff. Already, there is an outcry among the populace concerning the decline in quality of healthcare services especially in national healthcare facilities [2]. The timely and

efficient movement of materials, information and patients, positively impacts the satisfaction of both staff and patients and consequently improves the revenue of a healthcare facility [3].

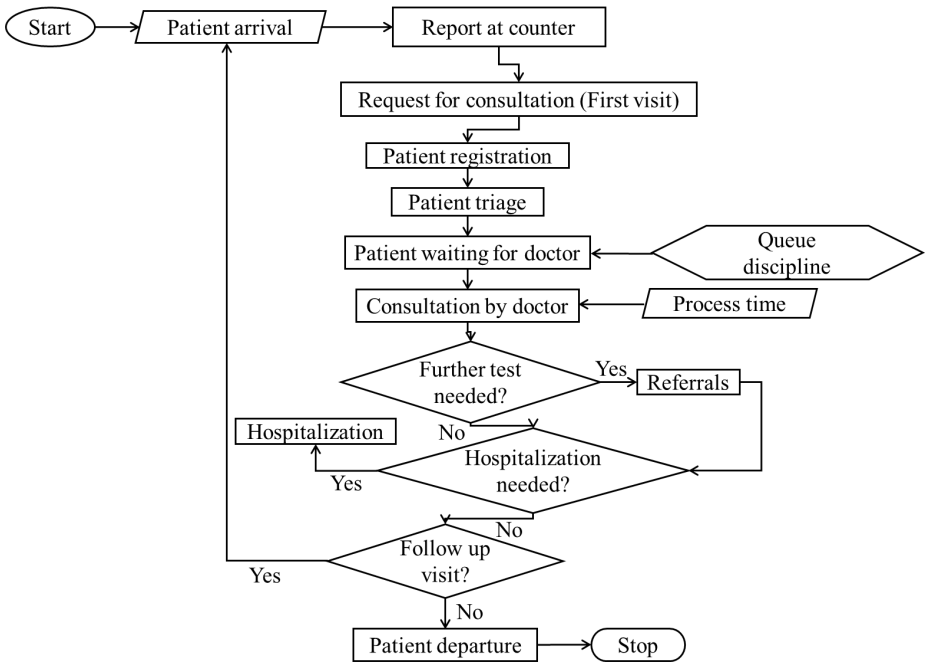

Figure 3.1. Patient flow in the outpatient clinic

In many African countries, unlike in the developed world, hospital visits by patients are rarely scheduled which makes it difficult to project daily demand for the services. Consequently, the mismatch between hospital staffing ratios and healthcare service demand remains a constant challenge. The efficient allocation and utilization of the limited healthcare staff requires knowledge of patient flow patterns. This knowledge can be generated through modeling of the arrival process of the unscheduled patients over the various intervals of the day and also seasons [4] [5]. The consequences of the in-balance between staffing ratios and service demand among others include; disgruntled medical staffs, excessive patient waiting times, and incomplete preventive service delivery. The extended waiting times may cause dissatisfaction among the patients which might manifest in the form of patients leaving the healthcare facility before the exam is complete [6] or may be forced to use personal connections with hospital staff and doctors to bypass the queues [7] [8] [9]. The major steps that a patient goes through to get medical care are summarized in Figure 3.1.

To ensure that patients get timely serviced at the facility, matching staffing ratios to patient demand is key [10]. This requires sophisticated approaches to study patient flow behaviours to

help hospital managers schedule staff according to the expected demand. Furthermore, knowledge of workload allows management anticipate the financial needs and in strategic planning for the healthcare facilities. Queuing theory can be used as a tool to understand various aspects in healthcare systems such as; appointment systems, utilization analysis, waiting time and others [11] [12] [13] [14].

There exists a dearth of research on assessing patient flow through the health clinics and also identifying various bottlenecks in the clinic structure both in Malawi and various sub-Saharan Africa countries. These time-motion studies have uncovered both deficiencies and barriers to patient flow, e.g., poor record keeping and poor allocation of health provider time [15] [16] [12]. Like many other low-income countries, Malawi faces a severe shortage of health workers, a consequence of the low output at the medical teaching institutions due to limited number of medical/health schools and brain drain. The impacts of these problems are well documented as pointed out above. The focus of this work is to demonstrate how queuing theory can be used to tackle the problem of a mismatch between staffing ratios and demand for health care services by optimizing the use of the available staff. To our knowledge, this is the first time that such a study is being carried out in Malawi. This study makes use of secondary data available from https://dx.doi.org/10.1186/s13104-016-2144-x . The data was collected for a study reported in [16].

3.2 Design and methodology

This research employs queuing theory to understand patient flow behaviour and how it relates to available human resources. This is because a healthcare system can be considered as a queue network consisting of various server types [11] [12]. The study uses secondary data that was primarily collected for a time-motion study at Ntaja health center [16]. The data was collected using a standardized questionnaire and consists of information pertaining to the patient-doctor (or health worker) consultation time, the number of medical staff(s) that attended to each patient, and the total time spent at the center. As part of the same study, an exit survey was conducted with an aim of collecting demographic information and data on patient's perception on quality of care during the visit to the health center. In terms of patients, the study identified two kinds of patients from the data namely; children who constituted 42.3% of the total patients and adults who constituted 57.7% of the patients surveyed. During the study period, all patients visiting the health center were invited to take part in the study. Upon receiving an informed consent,

each participant was assigned a unique identification number. As regards the exit survey, all participants were interviewed about whether they sought care for themselves or someone else. Since children are escorted by an adult, the adult was responsible for responding to the interview.

Table 3.1. A summary of relevant findings from the study [16]

Characteristic	**Waiting time mean (SD)**	**Contact time mean (SD)**	**Total time mean (SD)**
Patient type			
Adults	108.4 (67.6)	2.3 (6.0)	110.7 (67.9)
Children	133.2 (65.4)	1.7 (4.3)	134.9 (65.5)
Healthcare cadre			
Registry clerk	68.8 (55.3)	0.6 (3.0)	69.4 (55.3)
Hospital attendant	34.7 (34.5)	1.2 (6.0)	35.9 (34.6)
Medical attendant	59.9 (52.7)	1.1 (3.7)	61.0 (52.8)
Pharmacy attendant	13.5 (20.7)	0.4 (2.7)	13.9 (20.7)
Time of day			
06:00-08:00	156.6 (58.1)	1.8 (5.3)	158.4 (58.3)
08:00-10:00	133.4 (61.7)	2.0 (3.9)	135.4 (61.8)
10:00-12:00	73.4 (50.8)	1.5 (2.0)	74.9 (51.0)
12:00-14:00	65.6 (35.3)	1.8 (4.3)	67.4 (35.6)
14:00-16:00	52.7 (26.7)	2.0 (5.0)	54.7 (27.5)
Patient satisfaction			
Excellent	92 (71.0)	1.9 (2.5)	93.9 (71.3)
Very good	110.2 (66.1)	1.7 (5.1)	111.9 (66.2)
Good	122.0 (68.7)	1.8 (4.3)	123.8 (68.8)
Poor	129.0 (63.1)	1.8 (5.0)	130.8 (63.3)
Very poor	121.0	2.6 (4.0)	123.6 (76.2)

Table 3.1 is a summary of the relevant parameters extracted from the secondary data that have been used to develop and validate the model. Note that SD stands for standard deviation. From the table, the total average time spent by patients at the health center was 123 minutes (2-366 minutes), with variations observed between time spent by children and that spent by adults. Health worker contact time was 2.3 minutes for adults and 1.7 minutes for children. Furthermore, short patient waiting times are associated with higher perceptions of quality of service. Hence, this study is very important since matching demand for healthcare service and

staffing rations can lead to short access times. Since adults and children queue on different queues and are attended by different physicians, the two queues can be considered to be independent. Note that, although the registration might be carried by the same registration officer, however, in this work, the queues considered are those in the waiting room, which are queues of patients waiting to be attended by the physician. The two service times are treated independently and are used to drive two queueing models governing the two independent queues.

3.3 Patient flow and queuing theory

Patient flow involves systematic steps that patients go through from the time they arrive at a healthcare facility to the time they leave or get discharged. The process involves interaction of both patients' behaviors and also medical activities [14]. Patient flow is an important process towards provision of improved quality of service and also improved utilization of hospital resources [14]. In this work we consider patient flow as a closed network in that, upon completion of a sequence of hospital activities, all arriving patients return to the beginning (or starting point). Hence, a closed patient flow network is capable of reaching steady state after a considerable period of time.

In practice, it is unrealistic to expect 100% utilization of a healthcare system. Hence, for improved health outcomes, healthcare facility managers strive to keep the total waiting and capacity costs to a minimum. One way through which waiting and capacity costs can be kept to a minimum is by matching staffing ratios to patient demand for health care services, i.e., ensuring that patients arrival rates and service rates must be stable. To evaluate the utilization of a healthcare system, it is important to consider such aspects as: average number of patients and average time the patients wait in the queue, the capacity utilization, costs of a given level of capacity and the probability for an arriving patient to wait for service (computed using other relevant variables). These elements make the application of queuing theory to study patient queue behavior a natural approach. A queuing model is characterized by [14]:

- Number of servers (in this case servers means physicians): associated with each server is its capacity. Experience and exposure can have an impact on the capacity of a particular physician. In turn the capacity of a server directly influences the capacity of the queue system.

- Population source: in this work we consider an infinite population source, whereby patient arrivals are unrestricted hence at any time the system's capacity can be exceeded.
- Arrival patterns: patient arrival patterns are not uniform. As such a system may be overloaded (temporarily) due to the variability in both the arrival and service patterns.
- Service patterns: patients have varying nature of sickness such that the times for treatment are also varied.
- Queue discipline: in this work we assume that the patients are treated on the First-Come-First-Serve (FCFS) policy.

A queue can be modeled by using the extended Kendall's notation, A/B/C/D/E/F where A: is arrival time distribution, B: the service time distribution, C: the number of servers, D: capacity of the system (i.e., the number of customers in the system), E: the calling population and F: the queue discipline. The goal of this work is to demonstrate the application of queuing theory in determining the minimum number of physicians needed to achieve stable state of the healthcare delivery system. The input and output to our queuing model is the patient's arrival and the patient's discharge from the outpatient clinic respectively. It is important to note that, in Malawi, a rural health center is normally run by a medical assistant who sometimes is supported by other health workers, such as; registry clerks, nurse-midwives, attendants and health surveillance assistants. In the case of Ntaja health center, patients were first registered by the registry clerk, then proceeded to the medical assistant for consultation and diagnosis. Finally, patients were sent to the pharmacy for collection of prescribed medicine.

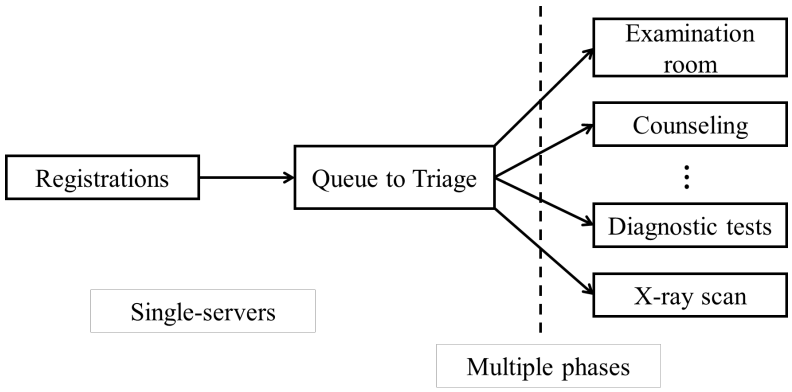

Figure 3.2. Single queue-multiple phases model environment depiction

In this work we consider a single queue, multiple phase model depicted in Figure 3.2. To achieve the goal of estimating the number of servers, an M/M/c queuing model is considered. An M/M/c model is a multi channel queueing system with Poisson arrival pattern and exponential service

pattern. Hence, note that where M/M/c is used, then it is assumed that E and F are clearly defined. In our case, E is infinite and F is FCFS queue discipline. As alluded to above, in this chapter, A takes up a Poisson arrival pattern and B an exponential service pattern. The Poisson arrival distribution allows us to compute the probability of arrivals over a given time period. We find these assumptions sensible since we are considering arrival of unscheduled patients in the outpatient clinic. Let:

$$\frac{\lambda}{c\mu} < 1, \tag{3.18}$$

where λ is the patient arrival rate, μ is the patient service rate and c is the number of servers. We further define; ρ as the system utilization, $\frac{1}{\mu}$ as the service time, P_0 as the probability of having zero units (or patients) in the system and P_θ as the probability of having θ patients in the system. Hence, for an optimized process, we wish to find the probability P_θ, i.e., the probability that an arriving patient queues for treatment. This means the probability that all the servers are busy. The relations below are applied to aid the calculation of the target probabilities:

$$P_\theta = P_0 \left(\frac{\lambda}{\mu}\right)^\theta \frac{1}{\theta!}, \qquad \theta < c \tag{3.2}$$

$$P_\theta = P_0 \left(\frac{\lambda}{\mu}\right)^\theta \frac{1}{c!\, c^{\theta - c}}, \qquad \theta \geq c \tag{3.3}$$

Considering that the basic property, $\sum_{\theta=1}^{\infty} P_\theta = 1$ is satisfied, then it is possible to calculate P_θ and all P_i for any number of patients i . Note that, a queue will only exist if $\theta \geq c$ (i.e., when the number of patients arriving is greater than the number of physicians). Hence, unless the number of arrivals surpasses that of servers, there is no need for optimization. Here, the aim of optimizing is to reduce the patient waiting time. However, the use of the model may vary depending on the problem at hand.

3.4 Results

As summarized in Table 3.1, our analysis of the secondary data revealed that the data was collected over a one-week period starting from 6am and ending at 4pm daily. Furthermore, we noted that a total of 1189 patients were registered during this period at Ntaja health center with

503 of these being children and 686 being adults. Hence, taking 10hrs to equal a day, the calculated hourly (60 min) arrival rate of adult patients and children is 10 and 14 patients respectively. Figure 3.3 and Figure 3.4 display the hourly patient flow over the five days of the survey. Using this information, below we drive two queuing models characterizing the movement of adult and also children patients in the system. These models are important in that they can inform management on the expected demand for service, making it possible for advance planning and resource scheduling.

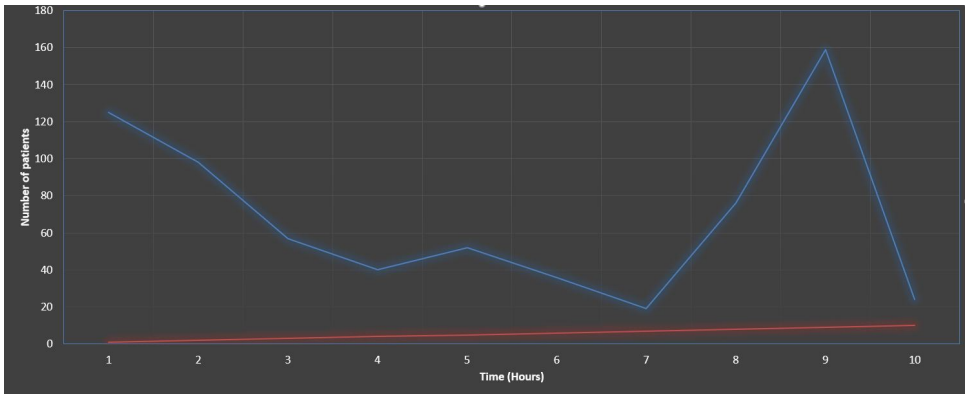

Figure 3.3. Average arrival of adult patients at Ntaja health center over 1 week period in 2016

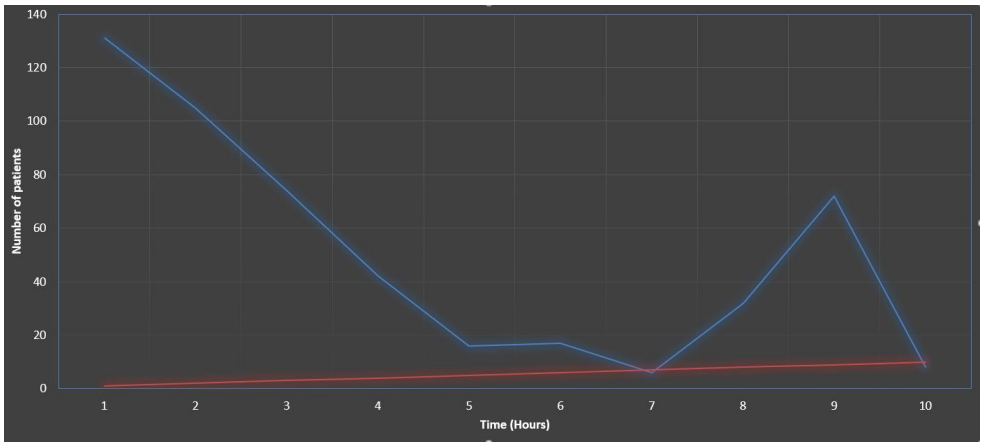

Figure 3.4. Average arrival of children patients at Ntaja health center over 1 week period in 2016

3.4.1 The adult patients queue model

Let σ_i be the weekly average arrival of patients during the time interval $[t_{i-1}, t_i]$, with $t_0 = 0$ means the zero hour, i.e., 6am in this case. Considering the limited amount of the data set, we consider the interval $[t_{i-1}, t_i]$ as the interval of hours of the day over which the clinic is

operational. Let $P_{\theta}^{h,i}$ denote the number of patients that arrived over an hour h in the interval $[t_{i-1}, t_i]$. Therefore, the average number of patients arriving/hour is given by:

$$\sigma_i^h = \frac{\sum_{d=1}^{5}\sum_{h=1}^{10} P_{\theta}^{h,i}}{5}, \tag{3.4}$$

where d is the day of the week and h is the number of hours of the day (i.e., working hours of the day). Hence, the average number of patients that arrived over the interval $[t_{i-1}, t_i]$ can be obtained from:

$$\sigma_i = \frac{\sigma_i^h}{10}, \tag{3.5}$$

Hence the average adult patients arrival will be given by:

$$\lambda_{ad} = \frac{\sum \sigma_i}{10}, \tag{3.6}$$

By applying equation 3.6 and information extracted from Figure 3.3 we get $\lambda_{ad} = 1$, i.e., adult patients arrive at a rate of 1 patient/minute and joins a queue. From the data we get a weekly service rate of: $\mu_{ad} = 0.4$ for adult patients. Recall that for a system to be in steady state the relationship $\frac{\lambda}{c\mu} < 1$ must be satisfied, where c is number of servers i.e., medical personnel. Hence, upon evaluating $\frac{\lambda_{ad}}{c\mu_{ad}} < 1$ the minimum number of servers needed for the system to achieve steady state is: $c = 3$. This means our queuing model is an M/M/3, which can be used to understand the various characteristics of the queue. Below we derive the probability that the adult queue is empty i.e., that no adult patient is in the OPD.

The probability that the queue is empty can be expressed as:

$$\sum_{\boldsymbol{\theta=1}}^{\infty} P_{\theta} = P_0 + P_1 + P_2 + P_3 + \cdots = 1 \tag{3.7}$$

For c=3 we get

$$P_0 + P_0\frac{\lambda_{ad}}{\mu_{ad}} + P_0\left(\frac{\lambda_{ad}}{\mu_{ad}}\right)^2\frac{1}{2!} + P_0\left(\frac{\lambda_{ad}}{\mu_{ad}}\right)^3\frac{1}{3!\,3^0} + P_0\left(\frac{\lambda_{ad}}{\mu_{ad}}\right)^4\frac{1}{3!\,3^1} + \cdots = 1 \tag{3.8}$$

$$P_0 + P_0\frac{\lambda_{ad}}{\mu_{ad}} + P_0\left(\frac{\lambda_{ad}}{\mu_{ad}}\right)^2\frac{1}{2!} + P_0\left(\frac{\lambda_{ad}}{\mu_{ad}}\right)^3\frac{1}{3!}\left[\frac{1}{3^0} + \left(\frac{\lambda_{ad}}{\mu_{ad}}\right)\frac{1}{3^1} + \cdots\right] = 1 \tag{3.9}$$

Consider the 4th term of equation (3.9) above.

$$\begin{aligned} P_0\left(\frac{\lambda_{ad}}{\mu_{ad}}\right)^3\frac{1}{3!}\left[1 + \left(\frac{\lambda_{ad}}{\mu_{ad}}\right)\frac{1}{3^1} + \cdots\right] &= P_0\left(\frac{\lambda_{ad}}{\mu_{ad}}\right)^3\frac{1}{3!}\sum_{n=0}^{\infty}\left[\left(\frac{\lambda_{ad}}{\mu_{ad}}\right)\frac{1}{3}\right]^c \\ &= P_0\left(\frac{\lambda_{ad}}{\mu_{ad}}\right)^3\frac{1}{3!}\left(\frac{1}{1-\frac{\lambda_{ad}}{3\mu_{ad}}}\right) \end{aligned} \tag{3.10}$$

Equation (3.10) holds, since the expression is clearly a geometric series with $|r| < 1$, hence, it is convergent as $c \to \infty$.

Substituting the term back into the equation (3.9) we get:

$$P_0 + P_0\frac{\lambda_{ad}}{\mu_{ad}} + P_0\left(\frac{\lambda_{ad}}{\mu_{ad}}\right)^2\frac{1}{2!} + P_0\left(\frac{\lambda_{ad}}{\mu_{ad}}\right)^3\frac{1}{3!}\left[\frac{1}{1-\frac{\lambda_{ad}}{3\mu_{ad}}}\right] = 1. \tag{3.11}$$

Which reduces to:

$$P_0\left(1 + \frac{\lambda_{ad}}{\mu_{ad}} + \left(\frac{\lambda_{ad}}{\mu_{ad}}\right)^2\frac{1}{2!} + \left(\frac{\lambda_{ad}}{\mu_{ad}}\right)^3\frac{1}{3!}\left[\frac{1}{1-\frac{\lambda_{ad}}{3\mu_{ad}}}\right]\right) = 1. \tag{3.12}$$

Which implies;

$$P_0 = \frac{1}{1 + \frac{\lambda_{ad}}{\mu_{ad}} + \frac{1}{2!}\left(\frac{\lambda_{ad}}{\mu_{ad}}\right)^2 + \left(\frac{\lambda_{ad}}{\mu_{ad}}\right)^3\frac{1}{3!}\left[\frac{3\mu_{ad}}{3\mu_{ad} - \lambda_{ad}}\right]} \tag{3.13}$$

Hence, the probability that there is no patient in the adult queue is: $P_{0_{ad}} = 0.04494$.

The above computed parameters for the adult patient's M/M/3 queueing model, including arrival rate and service rate were utilized to generate other steady state probabilities that were used to plot the steady state distribution curve for the adult patient's queue shown in

Figure *3.5.* Steady state probabilities distribution curve for the adult queue at Ntaja health center

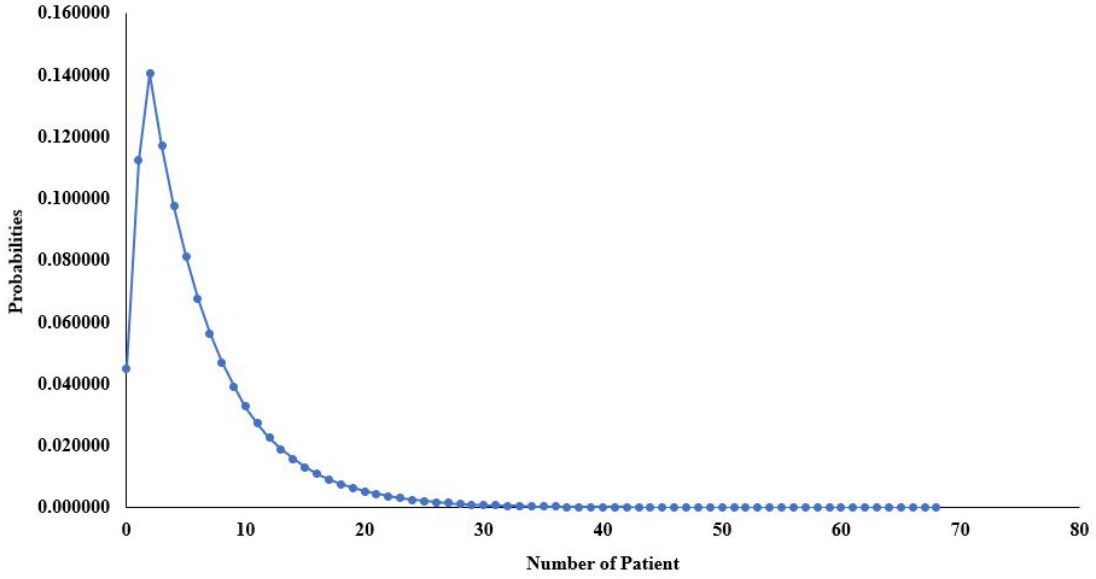

Figure 3.5. Steady state probabilities distribution curve for the adult queue at Ntaja health center

3.4.2 The children patients queue model

The same parameters as in the adult patient's queue analysis are adopted. Our focus in this section is calculation of: λ_{ch}, μ_{ch}, and n. From the data we get: $\lambda_{ch} = 1$, $\mu_{ch} = 0.6$ and $c = 2$.

Following similar approach as above, the probability of that the children's' queue is empty will be given by:

$$P_0 = \frac{1}{1 + \frac{\lambda_{ch}}{\mu_{ch}} + \frac{1}{2!}\left(\frac{\lambda_{ch}}{\mu_{ch}}\right)^2 \left[\frac{2\mu_{ch}}{2\mu_{ch} - \lambda_{ch}}\right]}. \tag{3.14}$$

Hence, the probability that the children's queue is empty is: $P_{0_{ch}} = 0.09090$.

3.5 Discussion

In this study, a queueing theory model approach is applied to discover hidden crucial queue parameters at Ntaja health center, a rural outpatient clinic in Malawi. Secondary data is used to validate the queuing models. Two queuing models have been driven, an M/M/3 model for adult patients and an M/M/2 model for children. This means that, to achieve steady state (i.e., that an arriving adult patient will not queue), the number of servers (i.e., medical assistant) needed is 3. Whilst for children, the needed number of servers is 2. The computed arrival rate for adult patients is found to be $\lambda_{ad} = 1$ with corresponding service rate of $\mu_{ad} = 0.4$. Since we

considered a multi-server scenario, the health center utilization rate by adult patients for the period is: $\rho_{ad} = \frac{\lambda_{ad}}{c\mu_{ad}} = 0.8333$. This means the server was busy 83% of the time. Similarly, other queue parameters such as; average queue length, average number of patients in the system, average waiting time of an arriving patient in the system, they all can be expressed in terms of arrival rate and service rate. Hence, the computed queueing parameters in this work are important in guiding managers to make informed decisions on staffing ratios for improved patient flow. This is crucial, since in clinical environments, any changes made in the system need to lead to improved experience for the patient. Similar computations can equally be made for the children queue.

Although similar works on the application of queuing theory in healthcare exists, however, most of these works focus on the application of queuing theory in the emergency department. In [17], the authors applied queuing theory to study the general performance of an emergency department located in the Mures County, Romania. In [18], Haghighinejad HA. et al. applied queuing theory to determine patient numbers waiting for emergency services and how it related to waiting times for an Iranian emergency department. The work reported in [8] is similar to this work as it demonstrates the application of queuing theory in an outpatient clinic. However, a few contrasts exists; e.g., the work in [8] uses primary data collected from a tertiary hospital whilst the results reported in this work are driven form secondary data for a rural outpatient clinic. Furthermore, the above referred works specifically focus on either resource utilization or patient waiting time and how it impacts patient's perception of quality of care. The work reported herein has shown how queuing theory can be used to match a clinic's staffing ratio and demand for healthcare service.

As indirectly alluded to above, this study has limitations. Two of these limitations include; 1) the use of secondary data, which may not reflect the current status of the study environment, and 2) the data covered a short period of time, hence, may not have captured the system behavior which may vary with weather, seasons and other variables in the environment. Despite these limits however, the modelling approach in this research shows that given data, it is possible to carry out experiments on different patient flow management policies.

3.6 Conclusion

This work has demonstrated how queuing theory as a tool can be used to analyze and understand queue related parameters at healthcare facilities. Managers of healthcare facilities can use queuing theory to plan scheduling of medical staff. Proper staffing is key to reducing lengthy waiting times consequently leading to improved quality of healthcare service delivery. Our future work will consider a large-scale analysis such as multiple phase queues which can be applied to central hospitals.

Funding: this work did not receive any external funding.

Contribution: The study conception and design were done by KM. KM, AG and MN participated in data analysis and interpretation. Manuscript drafting was done by KM and KJ. All authors critiqued the manuscript for intellectual content and approved it for publication.

Conflict of interest: The authors declare no potential conflict of interest. Availability of data and materials: The data that support the findings of this study are openly available at: https://dx.doi.org/10.1186/s13104-016-2144-x.

Ethics approval and consent to participate: Not applicable.

Patient consent for publication: Not applicable, since the study uses secondary data.

Significance for Public Health: This study has shown that queueing theory can be used to discover hidden crucial queue parameters that can guide hospital managers to make informed decision on staffing ratios for improved patient flow leading to a better experience for both staff and patients. Specifically, the study shows that; 1) There is need to regularly evaluate patient hospital visits patterns since the findings would help hospital administrators to apply dynamic staff allocation strategies tailored to patients flow patterns leading to optimized use of the limited staff. 2) The requirements for the provision of quality health services across the country and the related staff who go with these services varies depending, among other things the utilisation patterns of health services. Hence it is important that the allocation and utilization of medical staff (which are limited) takes into consideration of these variations within a country and not solely depend on population or institutional size.

References

[1] C. Alexopoulos, D. Goldsman, J. Fontanesi, D. Kopald and J. R. WIlson, "Modeling patient arrivals in community clinics," *Int. J. of Management Sci.,* vol. 36, pp. 33-43, 2008.

[2] W. Maphumulo and B. Bhengu, "Challenges of quality improvement in the healthcare of South Africa post-apartheid: A critical review," *Curationis,* vol. 42, no. 1, pp. e1-e9, 2019.

[3] A. Dennis, W. David, A. Joy and K. Edward, "Reducing patient waiting times in Rwadan hospital outpatient services," MSH, Kigali, 2019.

[4] H. Chu, R. Westbrook, S. Njue-Marendes, T. Giordano and B. Dang, "The psychology of the waiting time experience-what clinics can do to manage the waiting experience for patients: a longitudinal qualitative study," *BMC Health Serv Res,* vol. 19, no. 1, p. 459, 2019.

[5] J. Altena, K. Bien-Aime, M. Roger and W. Blaise, "Causes of long waiti time in health consultation services and strategies to reduce them: An observational study in rural Haiti," *The Lancet,* vol. 8, 2020.

[6] S. Creemers and M. Lambrecht, "Modeling a healthcare system as a queuing network: The case of Belgian hospital," *SSRN Electronic Journal,* pp. 1-54, 2007.

[7] E. Alenany and M. El-Baz, "Modeling a hospital as a queuing network analysis for improving performance," *World Academy of Science, Engineering and Technology, International Journal of Social, Behavioral, Educational, Economics, Business and Industrial Engineering,* vol. 11, pp. 1181-1187, 2017.

[8] N. Ameh, M. Oyefabi and B. Sabo, "Application of queuing theory to patient satisfaction at a tertiary hospital in Nigeria," *Nigeria Medical Journal,* pp. 54-64, 2013.

[9] L. Mayhew and D. Smith, "Using queueing theory to analyse the government's 4-h completion time target in accident and emergency departments," *Health Care Manag Sci,* vol. 11, pp. 11-21, 2007.

[10] F. Vericourt and O. Jennings, "Nurse staffing in medical units: A queuing perspective," *Operations Research,* vol. 59, pp. 1320-1331, 2011.

[11] B. Lantz and P. Rosen, "Using queuing models to estimate system capacity," *Production Planning & Control,* Vols. 1037-1046, p. 28, 2017.

[12] R. Wanyenze , G. Wagner, S. Alamo, G. Amanyire, J. Ouma, D. Kwarisima, P. Sunday, F. Wabwire-Mangen and M. Kamya, "Evaluation of the efficiency of patient flow at three HIV clinics in Uganda," *AIDS Patient Care STDS,* vol. 24, no. 7, pp. 441-6, 2010.

[13] R. Colebunders, T. Bukenya, N. Pakker, O. Smith, V. Boeynaems, J. Waldron, A. Muganga, C. Twijukye, K. McAdam and E. Katabira, "Assessment of the patient flow at the infectious diseases institute out-patient clinic, Kampala, Uganda," *AIDS Care,* vol. 19, no. 2, pp. 149-51, 2007.

[14] M. Chong, M. Wang, X. Lai, B. Zee, F. Hong, E. Yeoh, E. Wong, C. Yam, P. Chau, K. Tsoi and C. Graham, "Patient flow evaluation with system dynamic model in an emergency department: Data analytics on daily hospital records," in *2015 IEEE International Congress on Big Data*, New York, USA, 2015.

[15] J. Berry, D. Hall, D. Kuo, E. Cohen, R. Agrawal, C. Feudtner, M. Hall, J. Kueser, W. Kaplan and J. Neff, "Hospital utilization and characteristics of patients experiencing recurrent readmissions within children's hospitals," *JAMA,* vol. 305, no. 7, pp. 682-90, 2011.

[16] M. Jafry, A. Jenny, S. Lubinga, E. Larsen-Cooper, J. Crawford, C. Matemba and J. Babigumira, "Examination of patient flow in a rural health center in Malawi," *BMC Res Notes,* vol. 9, 2016.

[17] V. Hajnal and K. Zsuzsanna, "Application of queuing model to patient flow in emergenency department: Case study," *Procedia Exonomics and Finance,* vol. 32, pp. 479-487, 2015.

[18] H. Haghighinejad, E. Kharazmi, N. Hatam, S. Yousefi, S. Hesami, M. Danei and M. Askarian, "Using queuing theory and simulation to reduce waiting times in an emergency department," *Int J Community Based Nurs Midwifery,* vol. 4, no. 1, pp. 11-26, 2016.

Chapter 4: Technology for Improved Operating Room Scheduling-A Case of Kilimanjaro Christian Medical Center of Tanzania

Optimal scheduling of surgeries can result in efficient utilisation of the limited available operating rooms. Considering that operation theatres are an expensive investment, the optimal usage of these facilities will ensure that the dividend from the investment on the facilities is maximised. The integration of proper technology in the scheduling process can aid the creation of optimal schedules of patients who want to undergo surgeries leading to minimisation of patient waiting times under conditions of limited available operating facilities and specialised equipment. This work seeks to model the scheduling of surgeries at Kilimanjaro Christian Medical Center, a referral hospital located on the northern corridor of the Republic of Tanzania. We model the operation room scheduling problem as an integer linear programming problem. The model is solved using Torsche toolbox with the help of MATLAB routines and functions by considering appropriate configurations like resources, task parameters and optimisation criterion.

4.1 Introduction

Operating theater (OT) planning and scheduling remains an important subject of research [1] [2]. Considering that many patients undergo surgical interventions in their care pathway, the importance of the OT cannot be overemphasized. Despite their importance, the OT also remains one of the most expensive departments within the health facility, hence planning and scheduling ensures optimal use of this resource. Furthermore, it is important that the waiting time of patients on the waiting list has to be minimized [3]. Decisions involving planning can be classified into; strategic, tactical and operational [4]. For healthcare settings, the operational level planning is split into an online and offline operational level, with the former involving the controlling and monitoring of the process in real time and the latter involves advance short decision making [5]. Tactical level planning involves using patient demand (e.g., surgery appointment requests) to address resources usage. Tactical level planning focuses beyond the surgery sequencing, rather the level verifies whether the planned surgeries will cause resource conflicts for the operating theater and the subsequent hospital departments (e.g., wards) or for medical instruments.

Creation of the master surgical schedule is done at this level [6].

Operating theater scheduling is classified into three types of strategies [7], namely; open scheduling, block scheduling and modified block scheduling. Under open scheduling strategy, surgical cases are assigned to an operating room at the convenience of the surgeon; whilst under block scheduling, specific surgeons or groups of surgeons are assigned a set of time blocks, either for some weeks or months into which they can schedule their surgical cases [8]. Unlike the block scheduling where surgeons own time blocks such that they cannot be released, in the modified block scheduling, the scheduling is flexible, i.e., some time is blocked and some is left open, or unused block time is released at an agreed-upon time before surgery.

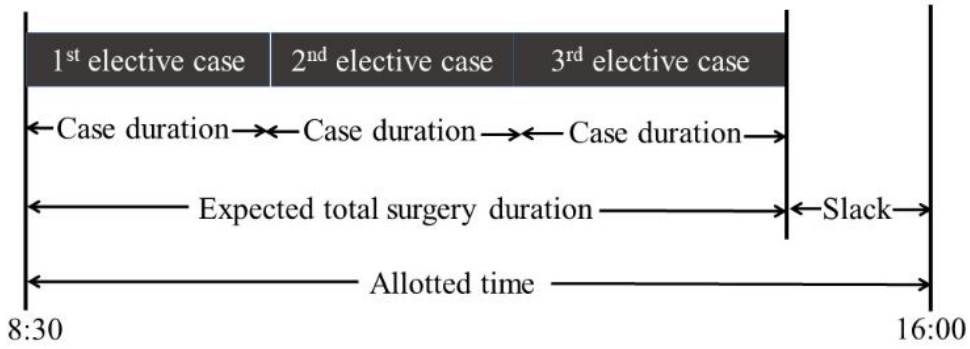

Figure 4.1: Elective surgical time line

The scheduling of operating theater for emergency surgery is complex due to the different patient arrival patterns and a further need for the patient to wait the minimum possible time before proceeding to OT. Furthermore, for many hospitals, there usually is a reserved operating room on standby for emergency surgery. Hence, in this work we consider the scheduling of operating theaters for elective surgery. Normally, elective surgery patients will make surgery appointments and will arrive as per the appointment time [9]. Upon arrival, the nurse prepares the patient for surgery, after which he/she is taken to the operating table. Once the surgery is successfully performed, the patient is taken to the recovery ward. The length of stay in the recovery ward will vary from patient to patient depending on varied factors. Figure 4.1 shows a general outlay surgical time line for elective surgeries. Each surgery will have a duration time and the slack time (i.e., the reserved capacity) which ensures that overtime is accommodated [10].

The problem of operating room scheduling has received considerable attention from researchers [11] [12] [13] [14] [15] [16], with approaches such as; queuing models, simulation models,

heuristic approaches, deterministic and stochastic mathematical programming models widely used to investigate this problem [17] [18] [19] [20] [21]. However, as highlighted by Rym et al. in their work titled "The planning and scheduling of operating rooms: A simulation approach", some of these models are very complex such that their application requires advanced mathematical programming knowledge; a factor that hinders their adoption by health care professionals [22]. Furthermore, in Sub-Saharan Africa, despite the existing challenge of limited availability of surgical facilities [23], much of the research on surgical services has focused on social aspects of the situation rather than on optimized use of the existing infrastructure. Some of the focus of recent research include; assessment of factors driving cancellation of surgeries [24], the assessment of cost and emotional impact of cancellation of elective surgeries [25], studies into how to scale up safe surgeries [26], and assessment of factors affecting starting delays in operating theaters [27]. This research seeks to bridge this gap by presenting a simple OT optimization model and further introduce technological tools that healthcare professionals and hospital administrators can use to ensure the optimal use of the limited, yet costly and important surgical facilities.

As pointed out, despite being costly, operation theaters play a crucial role in the provision of healthcare to patients since it is here where the surgical solutions are tried. A typical OT comprises several operating rooms and one or more recovery rooms where recuperating patients will be moved upon completion of the surgery. There is also waiting rooms where patients will be prepared for surgeries. The movement of patients from one section to the other during the entire process needs to be as seamless as possible. This work seeks to study the scheduling problem of operation theaters at the Kilimanjaro Christian Medical Centre (KCMC), an academic tertiary referral hospital [28]. We formulate the OT theater allocation as an optimization problem and use the Torsche-a Time Optimization, Resource SCHEduling toolbox for MatLab to solve the problem, where solving implies finding an optimal schedule [29]. Below we give an overview of the healthcare system in Tanzania with a focus on the situation at KCMC.

4.2 Surgery service delivery in Tanzania

In many developing countries surgical care is extremely limited. Tanzania is one such country where millions are in-need of surgical care but can hardly access it [30]. According to the World Health Organization survey of Tanzanian's primary healthcare services, it was discovered that

suturing was routinely available at all medical facilities, against the 35 listed basic interventions. Unfortunately, these medical facilities serve a population of over 23 million with only 64 surgeons, across multiple sub-specialties. These facilities clearly lack the ability to provide a full complement of surgical services [31]. This puts a burden on a few referral hospitals in the country that have the infrastructure and equipment available and hence have a capacity to deliver emergency care. It is therefore these tertiary referral facilities that are more likely to provide surgical care in its entirety to all patients in need [32].

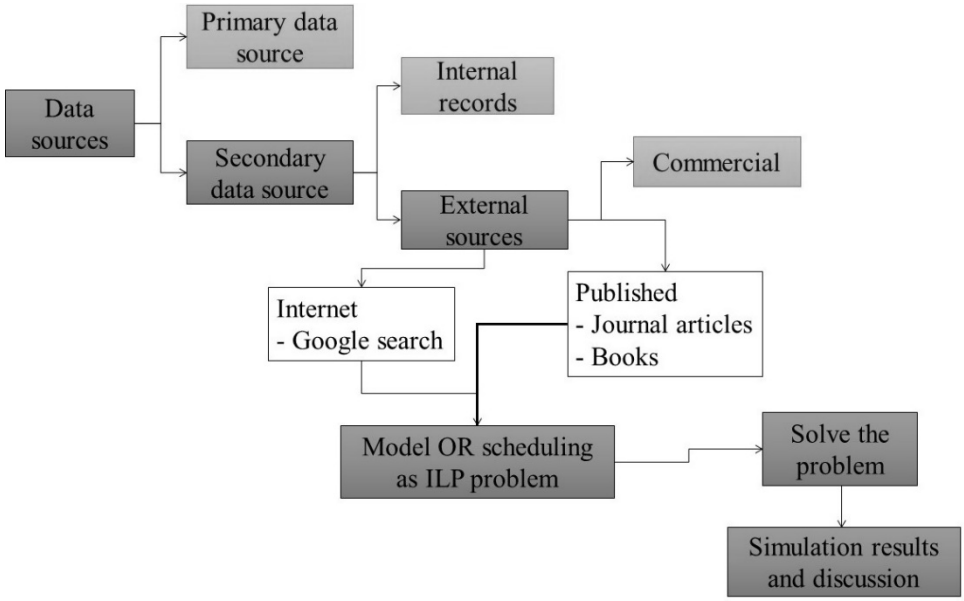

Figure 4.2: Data source and general study flow

This research uses secondary data of the Kilimanjaro Christian Medical Centre (KCMC), an academic tertiary referral hospital located on the northern corridor of the Republic of Tanzania. Figure 4.2 shows the data source and general flow of this study. The rural areas of the northern corridor of Tanzania has a population of over 11 million and comprises provinces of Kilimanjaro, Tanga, Arusha, Manyara and Singida. Despite availability of primary medical facilities and other smaller providers of care in the region, this area is primarily served by a single major academic tertiary referral hospital, the Kilimanjaro Christian Medical Centre. Usually, the facility is overwhelmed by the enormous demand for surgical services in the region, such that for example; 90 of the population in need of orthopaedic surgery are unable to access the service. The orthopaedic department at KCMC alone sees over 11,000 admitted and outpatient patients per year. Of the annual inpatient census, over 95 require surgical intervention of which less than 60 actually receive surgical care, with the average time to surgical

intervention being greater than 10 days [33]. This clearly demonstrates the existing discrepancy between supply and demand. This mismatch between demand and supply can be attributed to, among other factors; limited physical resources, work-flow issues, operating theater case mix and patient financial burden [34] [35].

Table 4.1: Five months analysis of payment methods, operating days, case volume and cancellations for KCMC

Surgery type	Payment methods			Operating days	Case volume	Cancellation
	Cash	Insurance	NL/Others			
General	104	11	124	104	244	64
Gynecology	70	71	3	74	144	33
Orthopaedics	161	132	39	122	323	76
Septic	105	91	12	110	208	65

At Kilimanjaro Christian Medical Centre tertiary referral hospital, there are five main operating suites, namely; multi-disciplinary emergency surgery, general surgery, orthopaedic surgery, multi-disciplinary septic surgery and gynecology. At the end of each case, a surgical staff manually records activities of the operating theater and the records are stored in the administrative rooms in the surgical wards. In this work we use secondary data to demonstrate how technology on scheduling can be used to improve efficiency in the scheduling of the operating rooms at KCMC. In the work titled "Understanding surgical care delivery in Sub-Saharan Africa: a cross-sectional analysis of surgical volume, operations, and financing at a tertiary referral hospital in rural Tanzania", Rajaguru et al. assessed the operations and financing of the main operating theaters at KCMC through a retrospective review of operating report books. Table 4.1 summarizes the findings from their five months analysis of payment methods, operating days, case volume and cancellations for KCMC selective surgeries [36].

4.3 Operating theater scheduling and torsche toolbox

Scheduling of operating theaters can aid in addressing specific objectives that can lead to improved service delivery. Some possible objectives for scheduling may include; minimizing patient waiting time, minimizing total idle times of operating rooms or minimizing $C_{\max}$, i.e., time of completion of last surgery. It is well known that the planning and scheduling of the processes of an operating room area is a very complex task. The complexity is a consequence of the many constraints that need to be met, some of which are opposite in their objectives. These constraints includes; the availability of the surgeon, anesthetists, and supporting staff. To

minimize the patient waiting time and also ensure that there is no overload on any of the resources, it is important that all operating stations function nearly the same duration.

Developed by the Czech Technical University in Prague, TORSCHE (Time Optimization, Resources, Scheduling) Toolbox for MATLAB supports solving integer linear programming problems. In TORSCHE, a task is a data structure, defined with all scheduling process parameters such as; processing time, arrival time, starting time, release time, deadline, due date, etc. Figure 4.3 is a graphical representation of task parameters in TORSCHE. Below we briefly discuss each of these task parameters.

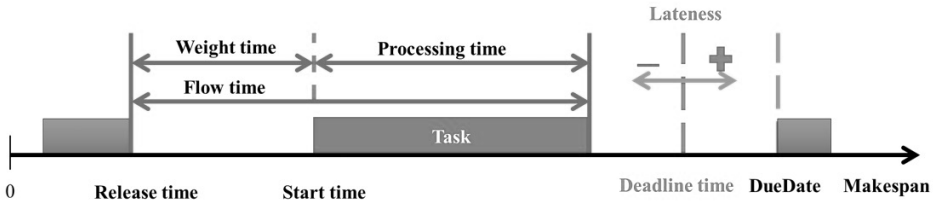

Figure 4.3: TORsche task parameters description

- **Release time (or Ready time):** this is time at which a task becomes ready for execution represented as; r_{Task_j}. If $r_{Task_j} = 0 \; \forall \; Task_j$, it implies that all tasks have the same release time.
- **Start time:** this is time when the execution of $Task_j$ is started, represented as S_{Task_j}.
- **Due date:** this is a time limit $\left(d_{Task_j}\right)$ by which the task should be completed. Associated with the due date are penalty functions that are defined.
- **Deadline time:** this is a hard real-time limit $\left(\bar{d}_{Task_j}\right)$by which a task becomes ready for execution and must be completed.
- **Completion time:** this is time $\left(C_{Task_j}\right)$ when the execution of a task is finished.
- **Lateness:** this is the time difference in executing a certain task and earliness finishing time operations before meeting the deadline, defined as; $L_{Task_j} = C_{Task_j} - d_{Task_j}$.
- **Makespan:** this is time at which the task is finished and it is calculated as; $C_{\max} = \max C_{Task_j}$.
- **Weight (or Priority):** this is the priority $\left(\omega_{Task_j}\right)$ of the task.

- **Flowtime:** this is the period required for completing the task. It is the sum of waiting and processing times, i.e., $F_{Task_j} = C_{Task_j} - r_{Task_j}$.

The performance optimality criterion below is used to evaluate schedules;

$$\text{Schedulelength(Makespan)} C_{\max} = \max C_l, \tag{4.1}$$

Where mean flow time is defined as;

$$\bar{F} = \frac{1}{n}\sum_{j=1}^{n} F_j, \tag{4.2}$$

With maximum lateness given by;

$$L_{\max} = \max L_l. \tag{4.3}$$

Our goal is to demonstrate how this tool can be used to solve the problem of scheduling surgeries in the operating rooms of the operating theater at KCMC so as to optimize time. We formulate the operation room assigning problem as an integer linear programming problem and solve it using the $P \parallel C_{\max}$ scheduling algorithm. The objective of $P \parallel C_{\max}$ scheduling algorithm is to assign a set of independent tasks to parallel identical processors in order to minimize schedule length and preemption is not allowed. The algorithm finds optimal schedule using Integer Linear Programming (ILP). This problem is known to be NP hard in the strong sense and is called the $P \parallel C_{\max}$ problem. Recall that at KCMC the OT has five operating rooms (or theaters). Based on data reported in [36] we drive Table 4.2. Note that, although cases are treated as independent, however, in practice this may not be the case [37].

Table 4.2: Details of procedures carried out in 5 months at KCMC

Theater	Operating days	No. of cases	Complete procedure mean	Operational day length	Mean duration/procedure
General	104	244	2.35	5:52hrs	2.35hrs/proc
Gynacology	74	144	1.95	5:02hrs	2.6hrs/proc
Orthopaedics	122	323	2.65	6:12hrs	2.3hrs/proc
Septic	110	208	1.89	4:20hrs	2.22hrs/proc

For a typical operating theater scheduling, the basic requirement is that the total duration of surgeries on each of the operating beds is nearly the same with no priority for any of the surgeries. All the operating beds are considered to be equal, in the sense that any surgery could be assigned to any of the beds. However, in real scenarios, the scheduling of the entire surgical process might produce an overhead when different surgeries are scheduled in a surgical bed.

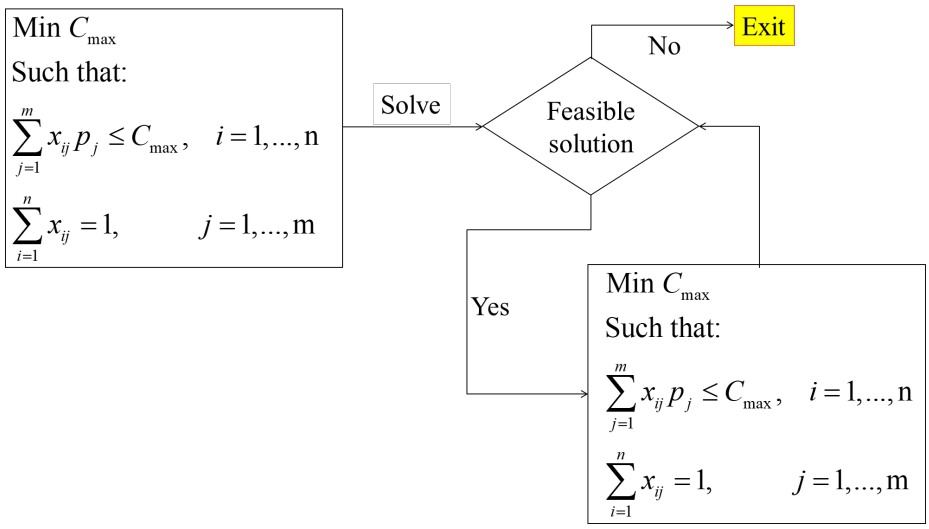

Figure 4.4: Flow of how the scheduling problem is solved

Disregarding this above however, the operating theater scheduling problem can be mathematically, formulated as follows:

$$\min C_{\max} \tag{4.4}$$

Subject to:

$$\sum_{j=1}^{n} x_{ij} p_j \le C_{\max}, \quad i = 1, \cdots, m \tag{4.5}$$

$$\sum_{i=1}^{m} x_{ij} = 1, \quad j = 1, \cdots, n, \tag{4.6}$$

where p_j is the duration for which j operating room is in use, n represents operating rooms and m represents surgeries to be scheduled such that if surgery j is scheduled in room i, x_{ij}=1 else x_{ij}=0. Here, the goal is to minimize the schedule length by assigning a set of surgeries (i.e., independent tasks) to operating rooms (i.e., parallel identical processors). Since the tasks are independent, the execution of a particular task will never get precedence over others. The problem is solved using integer linear programming implemented in TORSCHE Toolbox for MATLAB. Figure 4.4 below show the flow of how the problem is solved in TORSCHE. Recall that the problem is solvable if a feasible solution (i.e., schedule) exists.

4.4 Simulation results and discussion

We simulate the scheduling of five operating rooms of an OT. Table 4.3 summarizes the simulation environment.

Table 4.3: A summary of simulation environment

Simulation tools	**Description**
Laptop	Intel(R) Core(TM) i5-5300U CPU @2.30GHz 2.29GHz
Operating system	Windows 10 Pro 64-bit
MatLab	R2018b
Torsche	Release 0.4.4

Recall that the problem $P \parallel C_{\max}$ is known to be NP hard in the strongest sense. However, the right hand side $C_{\max}$ is found by some approximation algorithms. Below we briefly describe some of these algorithms [38].

- Longest Processing Time (LPT): LPT is a List Scheduling (LS) algorithm's strategy that requires an arrangement of tasks in order of non-increasing processing time p_{j}.
- Shortest Processing Time (SPT): With SPT, the processes are arranged in the ascending order with respect to their processing times and are assigned to the next available processor in that order.
- Integer Linear Programming (ILP): Note that in equation 5.5 and equation 5.6, for the unknowns m, n; x_{ij} can take only integer values namely; 1 or 0 and is subjected to $m +$

n linear constraints. In addition to the above unknowns, $C_{\max}$ which is also an integer is an unknown. Hence, this problem is an Integer Linear Programming (ILP) problem solved by using branch and bound algorithm.

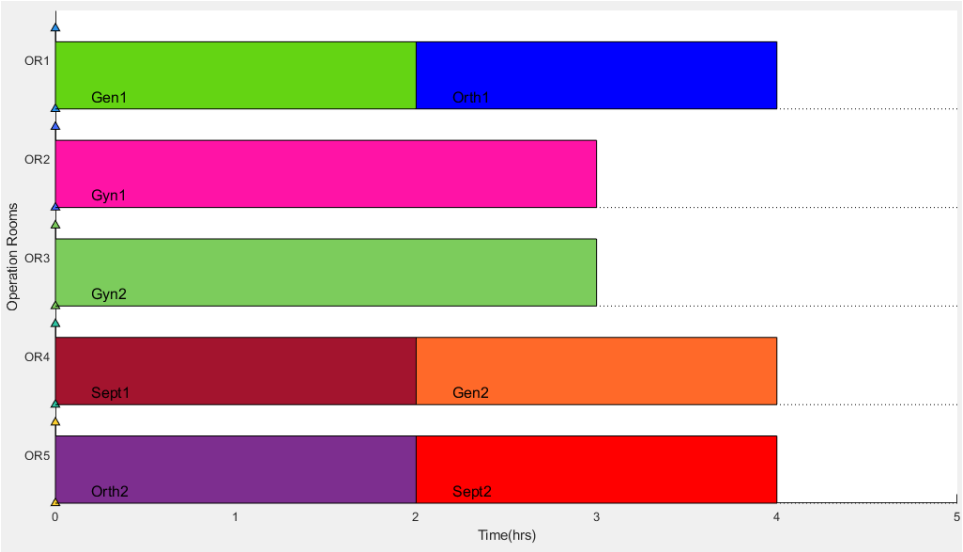

Figure 4.5: TORsche scheduling of 8 surgeries in 5 ORs

The "mean duration/procedure" column in Table 4.2

Table *4.3* shows the mean time (in hours) taken to complete a procedure. Comparing the "operating day length" for each surgery type, we see that each operation room carried out an average of 2 surgeries per operation day. To produce meaningful simulation results as depicted in Figure 4.5, we considered the scheduling of 2 surgeries for each surgery type. The surgeries were named as: SurgeriesToSchedule={'Gen1' 'Gen2' 'Gyn1' 'Gyn2' 'Orth1' 'Orth2' 'Sept1' 'Sept2'}, where mean duration for 'Gen1' and 'Gen2' is the same as in Table 4.2, similarly for 'Gyn2', 'Orth1', 'Orth2', 'Sept1' and 'Sept2'. Hence, Figure 4.5 shows the scheduling of 8 surgeries in the 5 operation rooms. From the generated schedule, it is found that Operating rooms 1,4 and 5 are utilized for 4 hours while 2 and 3 are utilized for 3 hours. From the figure it is noted that roughly, there is a 1 hour difference in total operation time between the operation rooms which is far less than the mean completion time of any type of the procedures. Hence, this shows that the surgeries were evenly distributed between the operation rooms.

4.5 Conclusion

Sub-Saharan Africa healthcare systems are stretched by demand for surgical services. This work highlights the situation at KCMC, an academic tertiary referral hospital in northern Tanzania.

While highlighting the importance of surgical interventions in the provision of quality healthcare, this work has also highlighted the challenges faced with surgical facilities. Since surgical department represents an expensive investment, it is important that they are used wisely so as to harvest optimal dividends from the investment on the facilities. Optimal scheduling of surgeries remains one of the way through which efficient utilisation of the limited available operating rooms can be achieved. Optimal scheduling of surgeries can further lead to reduced waiting times for patients on the waiting list. This work has demonstrated how the use of appropriate technologies can be used to achieve scheduling of limited operation rooms. We modelled the operation room scheduling problem as an ILP problem and used the TORSCHE toolbox in MATLAB to solve the problem. The toolbox managed to solve the problem by producing a schedule of patients in the limited operation rooms at the health facility. Our future work will consider using primary data to develop a procedure completion time prediction system using machine learning techniques. Such a system can be used by hospital managers to plan and schedule surgeries and achieve optimal usage of surgical facilities.

References

[1] Z. Shuwan, F. Wenjuan, L. Tongzhu, Y. Shanlin and M. Panos, "Dynamic three stage operating room schduling considering patient waiting time and surgical overtime costs," *Journal of Combinatorial Optimization,* vol. 39, no. 1, pp. 185-215, 2020.

[2] Z. Shuwan, F. Wenjuan, Y. Shanlin, P. Jun and M. Panos, "Operating room planning and surgical case scheduling: A review of literature," *Journal of Combinatorial Optimization,* vol. 37, no. 3, pp. 757-805, 2019.

[3] F. Dexter, R. Epstein, R. Traub and Y. Xiao, "Making management decisions on the day of surgery based on operating room efficiency and patient waiting times," *Anesthesiology,* vol. 101, pp. 1444-1453, 2004.

[4] R. Anthony, Planning and control systems: a framework for analysis, Boston: Harvad Business School Division of Research, 1965.

[5] E. Hans, M. Van Houdenhoven and P. Hulshof, "A framework for health care planning and control," *International Series in Operations Research & management,* vol. 168, pp. 303-320, 2011.

[6] E. Ruth and D. Franklin, "tactical increase in operating room block time for capacity planing should not be based on utilization," *Anesthesia & Analgesia,* vol. 106, no. 1, pp. 215-226, 2008.

[7] F. Hongying, M. Nadine and C. Chengbin, "An operating theater planning and scheduling problem in the case of block scheduling strategy," in *International Conference on Service Systems and Service Management*, Troys, 2006.

[8] L. Ya, C. Chengbin and W. Kanliang, "A new heuristic algorithm for the operating room scheduling," *Computer & Industry Engineering,* vol. 61, no. 3, pp. 865-871, 2011.

[9] F. Dexter and R. Traub, "How to schedule elective surgical cases into specific operating rooms to maximize the efficiency of use of operating room time," *Anesth. Analg.,* vol. 94, pp. 933-942, 2002.

[10] W. Erwin and T. Peter, "Operating theater planning and scheduling," in *In book: handbook of Healthcare System Scheduling*, 2012.

[11] L. Medhi, G. Fredric and X. Xiaolan, "Optimization methods for a stochastic surgery planning problem," *International Journal of Production Economics,* vol. 120, no. 2, 2009.

[12] C. Sangdo and E. Wilbert, "On capacity allocation for operating rooms," *Computers & Operations Research,* vol. 44, 2014.

[13] F. Dexter, A. Macario and R. Traub, "Which algorithm for scheduling add-on elective cases maximizes operating room utilization? Use of bin packing algorithms and fuzzy constraints in operating room management," *Anesthesiology,* vol. 91, pp. 1491-1500, 1999.

[14] F. Dexter, P. Shi and R. Epstein, "Descriptive study of case scheduling and cancellations within 1 week of the day of surgery," *Anesth. Analg.,* vol. 115, no. 5, pp. 1188-1195, 2012.

[15] F. Dexter , R. Wachtel, R. Epstein , J. Ledolter and M. Todd, "Analysis of operating room allocations to optimize scheduling of specialty rotations for anesthesia trainees," *Anesth. Analg.,* vol. 111, no. 2, pp. 520-524, 2010.

[16] C. McIntosh, F. Dexter and R. Epstein, "The impact of service specific staffing, case scheduling, turnovers, and first-case starts on anesthesia group and operating room productivity: a tutorial using data from ana Australian hosptal," *Anesth. Analg.,* vol. 103, pp. 1499-1516, 2006.

[17] T. Jean-Sebastien, R. Benoit, C. Jean-Philip and R. Fouad, "Assessing the impact of stochasticity for operating theater sizing," *Decision Support Systems,* vol. 55, pp. 616-628, 2012.

[18] H. Liu, T. Zhang, S. Luo and D. Xu, "Operating room scheduling and surgeon assignment problem under surgery durations uncertainty," *technol Health Care,* vol. 26, no. 2, pp. 297-304, 2018.

[19] M. Panos, G. Pando, P. Petraq and N. Britta, "Systems analysis tools for better health care delivery," in *Optimization and Its Applications*, Springer, 2013.

[20] S. Kavitha and P. Venkumar, "Ant-based job shop scheduling with genetic algorithm for makespan minimization on identical machines," *Int. J. Computer Aided Engineering and Technology,* vol. 9, no. 2, pp. 199-206, 2017.

[21] A. Muthiah and R. Rajkumar, "Scheduling problem solving using genetic and greedy algorithms," *Int. J. Computer Aided Engineering and Technology,* vol. 9, no. 2, pp. 207-217, 2017.

[22] H. Anwar and M. Rym, "The planning and scheduling of operating rooms: A simulation approach," *Computers & Industirial Engineering,* vol. 76, pp. 235-248, 2014.

[23] L. Sam, S. Macfarlane, J. Von Schreeb, M. Kruk, M. Cherian, S. Bergstrom and et al., "Increasing access to surgical services in sub-saharan Africa: priorities for national and international agencies recommended by the Bellagio Essential Surgery Group," *PLos Med,* vol. 6, no. 2, p. e100200, 2009.

[24] L. Martin, B. Papougnezambo, K. Bertille, F. Armel and et al., "Economic and psychological burden of scheduling surgery cancellation in a Sub-Saharan country (Burkina Faso)," *South African Journal of Anaesthesia and Analgesia,* vol. 23, no. 6, pp. 145-151, 2017.

[25] C. Pittalis, R. Brugha, G. Crispino and et al., "Evaluation of a surgical supervision model in three African countries-protocol for a prospective mixed-methods controlled pilot," *Plot Feasibility Study,* vol. 5, no. 25, 2019.

[26] M. Prin, J. Eaton, O. Mtalimanja and A. Charles, "High elective surgery cancellation rate in Malawi primarily due to infrustructural limitations," *World Journal of Surgery,* vol. 42, no. 6, pp. 1597-602, 2018.

[27] y. Assumpta, "Starting time delay in operating theater at University Teaching Hospital of Kigali (UTHK)," College of Medicine and Health Sciences, Kigali, 2017.

[28] "Kilimanjaro Christian Medical Centre," KCMC, [Online]. Available: http://www.kcmc.ac.tz/Services/general_surgery.php. [Accessed 20 February 2020].

[29] P. Ed and E. Leite, Matlab Modeling, Programming and simulations, Rijeka, Croatia: Sciyo, 2010.

[30] T. Weiser, S. Regenbogen and K. Thompson, "An estimation of the global volume of surgery: Amodeling strategy based on available data," *Lancet,* vol. 372, no. 9633, pp. 139-144, 2008.

[31] P. Tom, C. Hillary, P. Kibatala, A. Magoda, G. Saguti, L. Noel, S. Groth, D. Mwakyusa and M. Cherian, "Emergency and surgery services of primary hospitals in the United Republic of Tanzania," *BMJ Open,* vol. 2, no. 1, p. e000369, 2011.

[32] B. Tim, L. Edwin, E. Jaran, M. Victor, I. Lars and K. David, "Emergency and critical care services in tanzania: a survey of ten hospitals," *BMC health Services,* vol. 13, no. 1, 2013.

[33] P. Ajay, M. Honest, J. David, R. Jared, Y. Xiahan and P. Neil, "The burden of orthopaedic disease presenting to a referral hospital in northen Tanzania," *Global Surgery,* vol. 2, no. 1, pp. 70-75, 2016.

[34] M. Leshabari, E. Muhondwa, M. Mwangu and N. Mbembati, "Motivation of health care workers in Tanzania: A case study of muhimbili national hospital," *East African Joiurnal of Public Health,* vol. 5, no. 1, pp. 32-37, 2008.

[35] A. Geoffrey, I. Lenka, K. Peter, A. Lenard, P. Noralis, N. Joseph, R. Mayanja and G. Mark, "Out-of-pocket payment for surgery in Uganda: The rate of impoverishment and catastrophic expenditure at a government hospital," *PLoSOnE,* vol. 12, no. 10, p. e1087293, 2017.

[36] P. Rajaguru, M. Jusabani, H. Massawe, R. Temu and N. Sheth, "Understanding surgical care delivery in Sub-Saharan Africa: a cross sectional analysis of surgical volume, operations and financing at a tertiary referral hospital in rural Tanzania," *Global Health Research and Policy,* vol. 4, no. 1, pp. 1-9, 2019.

[37] T. Austin, H. Lam, N. Shin, B. Daily, P. Dunn and W. Sandberg, "Elective change of surgeon during the OR day has an operationally negligible impact on turnover time," *Journal of Clinical Anesthesia,* vol. 26, no. 5, pp. 343-349, 2014.

[38] B. Jacek, H. Klaus, P. Erwin, S. guenter and W. jan, Scheduling computer and manufacturing processes, Boston: Harvard Business School Division of Research, 2001.

Chapter 5: Conclusions

5.1 Summary and contributions

In this thesis, the outcomes of a study to optimize patient flow by reducing patient waiting time have been presented. Instead of applying firefighting measures to overcome these issues, this work has demonstrated that modeling-based solutions can go a long way in improving the QoS characterized by reduced queue abandonment and optimized utilization of surgical resources. Various aspects of the healthcare service provision and access have been modeled. Some of the models have been verified with secondary data.

As a way to help hospital managers to plan and make an informed assessment of needs, the knowledge of anticipated patient load is key. In Chapter 1:, a deep learning-based patient flow prediction model is presented. The proposed model takes past patient load as input and outputs future patient load. The advantage of deep learning is in its ability to learn such features by themselves, reducing the need for human experts. Since we propose patients to move from crowded health facilities to less crowded ones, the timely movement of these patients becomes a critical issue. To avoid putting a strain on the already fragmented ambulatory service in health facilities, a conceptual framework for an IoT based smart bus transport system that can support timely delivery of healthcare services has been presented. Using off the shelf sensors, an IoT smart transport kit is developed. This kit can be attached to ordinary public buses allowing patients and general passengers to query bus location and time related information to enable them move from one health facility to another.

Chapter 3: presents a queuing theory model that demonstrates how queuing theory as a tool can be used to analyze and understand queue related parameters at healthcare facilities. The results demonstrate that queuing theory is a simple but powerful tool that managers of healthcare facilities can use to plan scheduling of medical staff. Proper staffing is key to reducing lengthy waiting times consequently leading to improved quality of healthcare service delivery.

Chapter 1: highlights the importance of surgical interventions in the provision of quality health care and the challenges faced with surgical facilities. Since surgical department represents an expensive investment, it is important that they are used wisely so as to harvest optimal dividends

from the investment on the facilities. Optimal scheduling of surgeries remains one of the ways through which efficient utilization of the limited available operating rooms can be achieved. Optimal scheduling of surgeries can further lead to reduced waiting times for patients on the waiting list.

With relation to Section 1.3, the main contributions of this thesis are:

1. A formal machine learning-based patient flow prediction model for the outpatient departments. The model assumes the existence of a centralized hospital management system, that would allow hospitals to timely exchange patient load information thereby allowing excess patient load from an overcrowded health center to be timely re-assigned to the nearest health centers.
2. A queueing theory based model that relates patient waiting time and staffing ratios. The model has been verified on real data.
3. An integer linear programming model for operation room scheduling verified on real data.

5.2 Future research directions

Several aspects of our study need to be revisited and tested. Among our future works: 1) we intend to revisit the proposed patient flow prediction model presented in Chapter 1: and verify it against real data. Furthermore, we intend to study the arrival patterns of scheduled patients and also the transfer of patients health files since a patient's past record may have a bearing on his/her current health status. Technologies such as blockchain will be explored in this regard. 2) With regards to study findings in Chapter 3:, in our future work we will consider a large-scale analysis such as multiple phase queues which can be applied to central hospitals. Finally, 3) building on the model presented in Chapter 1:, we intend to use primary data in the future work to develop a procedure completion time prediction system using machine learning techniques. Such a system can be used by hospital managers to plan and schedule surgeries and achieve optimal usage of surgical facilities.

Author's Publications List

Publications published in international journals

1. **K. Mtonga,** S. Kumaran, K. Jayavel, C. Mikeka, Technology for improved operating room scheduling- A case of Kilimanjaro Christian Medical Center: Int. Journal of Comp. Aided Eng. & Tech, vol. 16, No. 1, 2021, doi: 10.15.04/IJCAET.2022.119588.
2. **K. Mtonga,** G. Antoine, K. Jayavel, M. Nyirenda, S. Kumaran, Adaptive Staff Scheduling at Outpatient Department of Ntaja Health Center in Malawi- A Queuing Theory Application**:** Journal of Public Health Research, 2022, 11, 2347. doi: 10.4081/jphr.2021.2347.
3. **K. Mtonga**, E. Twahirwa, K. Jayavel, Modeling Classroom Space Allocation at University of Rwanda-A Linear Programming Approach: Applications and Applied Mathematics: An International Journal (AAM), Vol. 16, Iss. 1, Article 40.
4. **K. Mtonga, S.** Kumaran, C. Mikeka, K. Jayavel, J. Nsenga. Machine Learning-based Patient Load Prediction and IoT Integrated Intelligent Patient Transfer Systems: Future Internet 2019, 11, 236.

Other publications

1. **K. Mtonga,** S. Kumaran, K. Jayavel, O. Gatera and W. Kasakula, "An Integrated Patient Triage and Capacity Recommender System for Robust Outpatient Department Service Delivery," *2022 4th International Conference on Computer Communication and the Internet (ICCCI)*, 2022, pp. 162-169, doi: 10.1109/ICCCI55554.2022.9850267.
2. **K. Mtonga**, S. Kumaran, K. Jayavel, O. Gatera, W. Kasakula, An Integrated Patient Triage and Capacity Recommender System for Robust Outpatient Department Service Delivery, Accepted at the the 4th International Conference on Computer Communication and the Internet (ICCCI 2022).
3. S.A. Prakash, S. Dhruvil, K. Jayavel, **K. Mtonga.** Hydropower Energy Generation Prediction Model – A Machine Learning Approach. 2022 International Conference on Computer Communication and Informatics (ICCCI), 2022, pp. 01-04, doi. 10.1109/ICCCI54379.2022.9740915.

4. E. Twahirwa, **K. Mtonga,** K. Jayavel, W. Kasakula, P. Bamurigire. Assessment of the Impact of COVID-19 on Operations of Local Businesses and Level of Enforcement of Public Health Safety Measure within Business Premises: A Quantitative Study of Businesses in Huye-Rwanda. *Sustainability*. 2021; 13(23):13013. https://doi.org/10.3390/su132313013.
5. E. Mwakilama, P. Ali, P. Chidzalo, **K. Mtonga,** E. Eneya. On Average Distance of Neighborhood Graphs and Its Applications: IntechOpen, doi:10.5772/intechopen.98986.
6. E. Twahirwa, **K. Mtonga,** D. Ngabo, S. Kumaran. A LoRa Enabled IoT-based Air Quality Monitoring System for Smart City: 2021 IEEE World AI IoT Congress (AIIoT), 2021, pp. 0379-0385, doi: 10.1109/AIIoT52608.2021.9454232.
7. K. Jayavel, K. Meenakshi, G. Lavanya, M. Suriyah, **K. Mtonga.** Virtual Aided Teacher-Learner Relationship-A Text Analytical Approach for Improved Behavior Analysis: 2020 Sixth International Conference on e-Learning (econf), 2020, pp. 104-109, doi: 10.1109/econf51404.2020.9385503.
8. **Mtonga K.**, Kasakula W., Kumaran S., Jayavel K., Nsenga J., Mikeka C. (2020), A Distance Integrated Triage System for Crowded Health Centers. In: Serrhini M., Silva C., Aljahdali S. (eds) Innovation in Information Systems and Technologies to Support Learning Research. EMENA-ISTL 2019. Learning and Analytics in Intelligent Systems, vol 7. Springer, Cham.