Leveraging Big Data and Machine Learning in Healthcare Systems for Disease Diagnosis

P. Rajyalakshmi
Assistant Professor
Department of CSE
Guru Nanak Institute of Technology (A)
Hyderabad, India
pannangiraji@gmail.com

Chinnala Balakrishna
Assistant Professor
Department of CSE
Guru Nanak Institute of Technology (A)
Hyderabad, India
balu5804@gmail.com

E. Swarnalatha
Assistant Professor
Department of ECE
Guru Nanak Institute of Technology (A)
Hyderabad, India
swarnalatha.ecegnit@gniindia.org

B. S. Swapna Shanthi
Assistant Professor
Department of CSE
Sri Indu Institute of Engineering & Technology
Hyderabad, India
swapnashanthi45@gmail.com

K. Aravind Kumar
Assistant Professor
Department of ECE
Guru Nanak Institute of Technology (A)
Hyderabad, India
aravindkumarkyrika@gmail.com

2022 3rd International Conference on Intelligent Engineering and Management (ICIEM) | 978-1-6654-6756-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICIEM54221.2022.9853149

***Abstract*— Substantial technology advancements in the medicare industry have resulted in a slew of advances in therapeutic interventions, patient health assistance programmes, identifying trends in medical consequences, and so on. This also contributes to a plethora of information resources that can provide a range of forecasts on a lot of illnesses. The paper mentions technology advances in health sector, as well as the intricacy of systems and data quantities that may be utilized to make sophisticated clinical forecasts. It illustrates the opportunities that Big Data (BD) and Machine Learning (ML) might offer to this profession by employing a system that uses a Matlab/Simulink predictive model of a person's health and AzureML to identify potential cardiac issues.**

Keywords*— Analytics; medical care; health sector; Prediction; AzureML.

I. Introduction

Technical and socioeconomic changes, as well as autoimmune disorders on the skyrocket of rapid urbanization and greater longevity, diverse illnesses, and unexplained mortality grounds as a result of a combination of variables, have raised the demand for excellent medical services. The increasing need for better, more effective, and tailored clinical care facilities has necessitated the incorporation of informational and communications technology with medical equipment into a system of numerous dispersed datasets, including consumers, patients, physiological, and clinical evidence. By gathering, preserving, and analyzing health information, this developmental phase decides large amounts of data from which interesting information may be derived in providing answers to health-related concerns.

This paper is divided into five segments that provide a comprehensive view of medical assimilation with communications and information exchange innovations, as well as the developmental move into medical prognostication that goes with it, and is deduced with a test case established with AzureML and the Matlab/Simulink simulations.

II. Assimilation Of Information Technology And Health Care

Each person has a distinct DNA identity that is detected using a mix of criteria such as behavior, health records, and paternal ancestry. Certain metrics (for example, EMG, EEG, BLE, physical measure) can be observed and reviewed using a network of instruments, including bio sensing devices, nano-sensors, and embedded systems [1,3]. Sensing devices emerged in response to the necessity to monitor actions or variations when employed in conjunction with electronic-based mechanisms in order to effectively communicate with it via controllers. These sensors detect tangible variables (such as sunlight, humidity, and warmth) and transform them into a changeable electrical impulse. Monitoring systems were integrated into the field as a result of advancements in semiconductors, information and communication techniques, computing, and mechatronics, as well as the need to upgrade to a more accurate, personalized health service depending on a person.

Internet infrastructural facilities, info, machine, and communications networks have enabled the connectivity of many machines to share data, opening the way for the tendency in health care services to incorporate sensing devices with the person to enable vital tracking for a more precise personal health prognostication – to anticipate malady outcomes results – utilizing the appropriate equipment. Among the tested variables that are gathered is prospective data that can be derived utilizing sophisticated ML methods like "deep learning (DL)," by implementing methodologies that attempt to imitate biomedical neural pathways into orchestrated process framework to enable machines to perceive the universe in terms of a ranking of notions and to gain knowledge from perception the relationship between each notion [2,4].

978-1-6654-6756-8/22/$31.00 ©2022 IEEE

The medical field is primarily data driven, perhaps one of the greatest major obstacles for healthcare implementations and systems is the computation of vast quantities of information that must be gathered from numerous references, in multiple settings, and analyzed in order to derive styles that could represent solutions in diagnosing illness, personalized therapies, population - based patterns, personalized and optimized hospital assistance, and so on. The phrase "BD" refers to very massive and complicated data volumes that need sophisticated analytics for investigation, translation, and computational capacity to reveal unknown data trends and knowledge. BD in health care system is discussed from various viewpoints in paper [1]: one way utilizes clinical information to enhance medical care services, while the alternative leverages knowledge as a significant resource for digital safety. Because the purpose of this paper is to examine healthcare concerns through the lens of business intelligence, BD is discussed in connection to ML approaches such as Artificial Neural Networks (ANN). Computational modeling is used to offer results and create recommendations in order to transform data into thoughts and chances for making quicker and more effective judgments. AzureML is a technology utilized in this paper for generating advanced analytical answers derived from datasets, and it is used in the research scenario in Section IV.

III. THREAT ELEMENTS FOR HEALTH

Medical care system generates massive volumes of data, which may be converted into statistics for essential medical discoveries utilizing ever-evolving technology. The demand for enhanced healthcare treatments and specialized assistance for people stems from a variety of factors such as environmental variation, an ageing urbanization, impoverishment, and behavioral health illnesses as a result of many of the consequences of urbanization expansion.

As per [3], carcinoma, hypertension, coronary artery illness, obstructive pulmonary illnesses, and psychiatric illnesses contribute for 86 percent of mortality in the European Community. Hazardous indicators such as cigarettes can raise the odds of early mortality by up to 50% between the age group of 60 and 75 for less energetic people, and this proportion reduces for energetic people. [4] lists the most important risk variables that are currently recognized. The relationship connecting "potential cause" and "illness" has been demonstrated by arbitrary research of the pathogenic activity of particular chemicals over age, evolutionary factors, and culture, with the expanding epidemiological knowledge repository concerning the infectious illnesses.

Following close examination, it is discovered that Table I comprises susceptibility variables for the major determinants of generalized fatality and hospitalization [5-7] worldwide.

TABLE I. PRIMARY HEALTH THREAT ELEMENTS

Hypertension;
Metabolic abnormalities;
Extreme dietary lipids intake;
The use of nicotine;
High liquor intake;
High sodium intake;
High intake of sucrose;
Inadequate liquids intake;
Changes in dextrose resistance;
No regular exercise;
Prolonged contact to the sunshine
Inadequate immunisation;
unsafe pornographic encounters;
Excess weight and being obese;
Changes to the sleeping schedule;
High use of chemical additives in food;

With the upsurge of a wide range of nutritional goods (elevated in sodium, lipids, sucrose content, flavours, and other materials) and the downgrading of a healthier daily lives in favour of one controlled by physiological and psychological anxiety as a result of productivity expansion, these illness communities were considered "illnesses of the medieval world" unless it was discovered that these started to progressively impact sections of younger age, in addition to other illnesses with rising prevalence (immunological sicknesses, diabetic neuropathy, etc.). Because of such characteristics, this has becoming a severe medical concern that necessitates a systematic strategy to, at the very minimum, reduce their result. The chapter that follows discusses the significance of BD analytics and displays a methodology for medical forecasting in Matlab/Simulink, as well as a cardiovascular issue detection in AzureML.

TABLE II. POTENTIAL REASON AND THREAT ELEMENTS FOR HEALTH

Potential Reason	Threat element

Heart Disease
(e.g. Aortic arterial dysfunction, sudden myocardium infarction)
	Hypertension;
	High sodium intake;
	The use of nicotine;
	High cholesterol;
	High blood sugar;
	Excess weight and being obese;
	genealogy of the ancestors;

Carcinoma (e.g. Adenocarcinoma)	The use of nicotine; Bad oxygen purity or irritating contact on a regular basis; a compromised immunological systems; genealogy of the ancestors; As a youngster, you had a background of chest diseases; Pneumonia, flu, sinusitis, TB, and HIV;
Long-term breathing Illness	The use of nicotine; Contact to allergens in the lungs for an extended period of time; A hereditary background of progressive pulmonary disease; Lung disorders;

	As a youngster, you had a background of chest diseases;
Diabetes	overindulgence in processed sweetener; hypertension; blood sugar intolerance problems; sedentary behavior; Excess weight and being obese;

Amnesia
(e.g. Alzheimer)
	Brain injuries;
	Ineffective interaction with others (isolation);
	Unhealthful way of living;
	Families genealogy – acquired ancestors' genetics;
	Age greater than 65 years;
	Moderate neurological dysfunction;

Other – Liver Disease	Long-term liquor Intake;
	Long-term infections hepatitis;
	Fatty lipids buildup around the liver;

IV. Big Data Analytics Architecture In Health Prognosis

As knowledge and data sharing technique is integrated and developed in the clinical scheme, an IT architecture is decided to include for accumulating, storage, and analyzing info via comprehensive solution of datasets (cell phones, wearable innovations, clinical image processing, diagnostic tests, epigenetic and fingerprint scanners sensing details, telehealth devices, and quantities) to achieve performance requirements shipment for primary health care. This move against a more intellectual strategy in the healthcare field generates vast volumes of data that may be handled by specialist programmes. BD refers to huge and complicated information volumes that need extensive analytics and reporting using computational methods in order to uncover underlying trends and correlations [2]. Sophisticated computing data may be utilized to create tailored individual histories and better accurate therapies. The offered information's capability for detecting healthy and pathological behaviour patterns associated with genetic background, chronic diseases, diet, and actual information (EKG, warmth, blood stiffness, and so on) is contributing to the emergence of information repositories. Figure 1 depicts a paradigm for illness prognosis based on BD and ML.

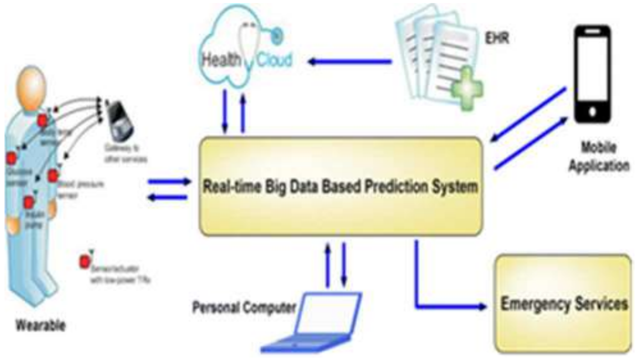


Figure 1. Architecture for medical prognosis by employing ML

ML is a subtype of artificial intelligence (AI) that simplifies the building of an empirical framework by evaluating acquired data for pattern classification, connections, and abnormalities. In medical, ML is used to [8-11, 13]:

i. Detect disorders and make diagnoses – earlier tumor detection. Through intelligent systems, IBM Watson Genomics assists in the diagnosis of cancers.

ii. Pharmaceutical designing and fabrication include upgrading or developing newer and better accurate medicinal treatments for a variety of ailments. Microsoft's Project Hanover, which is centered on machine learning, offers tailored malignancy medicine therapy.

iii. Determine behaviour adjustment using a smartphone application focused on ML that observes individual motions.

iv. Forecast outbreaks by analyzing massive amounts of data over the web and a notification network called ProMED-mail.

v. Use IBM Watson Cancer to personalize therapy by offering a variety of therapeutic choices.

vi. Evaluate clinical photos using an AI-powered tool called Inner Eye developed by Microsoft.

ML algorithms had improved significantly, and the capacity to manage massive amounts of info in a reasonable fashion via advanced statistical computations is a significant invention. The issue of ML is created as a remedy in the prognosis and assessment of clinical situation of patient in the research scenario part.

V. Research Scenario-Health Prognosis System

The objective of the whole scenario is to highlight the possible outcomes that BD and ML offer to the medical industry via a modelling a group of statistics to calculate the probability of forecasting a cardiovascular issues and analyzing the pattern of a clinical condition of patients; the modelling and advancement of the structure was done in the MATLAB/Simulink setting and AzureML utilizing computational library functions related to mathematics.

Picking the ML strategy is difficult since it is not necessarily feasible to determine which is the perfect suited ahead of time. To make this procedure easier in this scenario, the essential factors have been recognized:

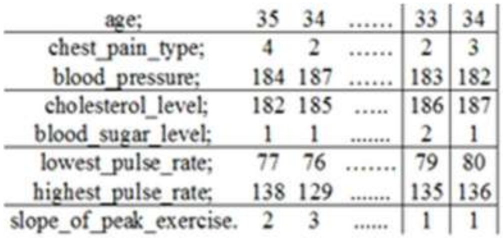
i. type of data, volume, and reliability;
ii. intended result;
iii. the amount of effort spent on training or calculation;
iv. the intended degree of precision of the outcome

The training dataset was taken from the UCI ML Repository [12, 14] and tailored to the research scenario, with the relevant characteristics:

TABLE III. RESEARCH SCENARIO FEATURES

Age;
Chest_pain_type;
Blood_pressure;
Cholesterol_level;
Blood_sugar_level;
Lowest_pulse_rate:
Highest_pulse_rate;
Slope_of_peak_exercise

A multivariate extrapolation method is developed to forecast the quantity of a parameter depending on the measurements of additional parameters, and the relationship between information may be defined by a straight equation with a specific level of precision [15]. For this research scenario, the threat rating is computed by evaluating the relationship between the data packet components. Due to the obvious inter-relationships among the predictor factors, the information is strongly linked, resulting in a multi - collinearity issue [16]. This was accomplished in AzureML modeling by utilizing a Python package called "pandas" and the Boolean technique to examine the association between the parameters (uninterrupted or distinct) [17-24].



age;	35	34	......	33	34
chest_pain_type;	4	2	......	2	3
blood_pressure;	184	187	......	183	182
cholesterol_level;	182	185	.....	186	187
blood_sugar_level;	1	1	....	2	1
lowest_pulse_rate;	77	76	....	79	80
highest_pulse_rate;	138	129	....	135	136
slope_of_peak_exercise.	2	3	......	1	1

Figure 2. Health data frame

Figure 2 is a snapshot of the medical perception, which was gathered each week for 19 years. This methodology was constructed for the Matlab/Simulink projection modelling utilizing ANN – vector auto regression neural networks – as ML techniques that were developed on the previous given dataset required to produce the forecast of a 5 years' time frame. Figure 3 depicts a modeled condition of the clinical condition of patient depending on the computed threat variable for cardiovascular illness:

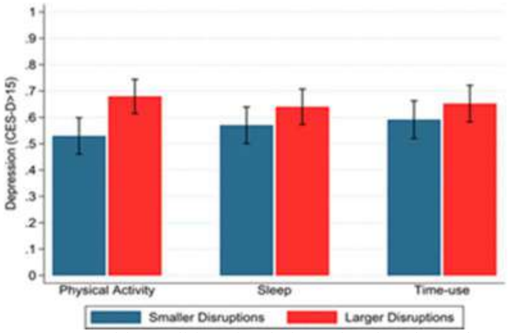


Figure 3. Health condition simulation related to threat factors

The NAR framework employed is a seven-layer system with the Levenberg-Marquardt learning mechanism. This structure was generated in MATLAB for 600 iterations with a learning target of less than 0.001. Figure 4 depicts the approximated situation throughout a 15-month span in red:

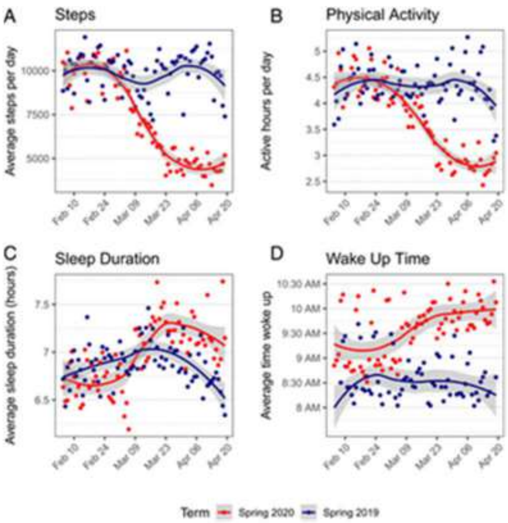


Figure 4. Modelled health condition relied on threat parameter – blue, estimated condition using ANN – red

Figure 5 highlights the depiction of the disparity between the anticipated and actual states, with the comment that the outcome may be enhanced by employing a larger set of statistics and more in-depth learning.

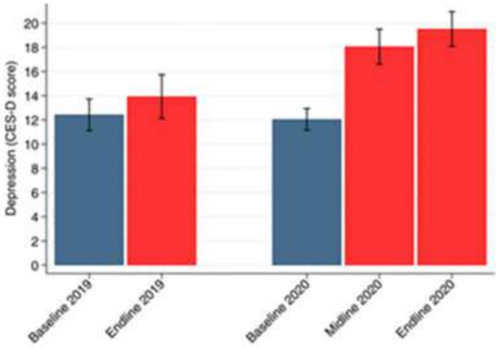


Figure 5. Delta proportions between the modelled and actual health states

The benchmark for raising an alert indicating a severe threat cardiac condition will be determined by clinical competence. This research demonstrates the contribution that sets of data utilized in conjunction with ML techniques and technologies offer the clinical industry. The findings may be enhanced and more accurate inferences drawn by continual learning and broad use of varied databases [25][26].

VI. CONCLUSIONS

The existing medical system transformation is based on the assimilation of data and messaging mechanisms with software and hardware to provide an interlocking network for letting meaningful data to be transferred, significantly improved tracking of the recipient's clinical condition, and overarching reliability in medical services. This paper presented and give an outline of the opportunities that BD can help the health sector for information gathering in order to use the perspectives it can provide to construct an experience and understanding framework that allows patient personal health prognostications for a more accurate and appropriate medical system.

REFERENCES

[1] T. T. Chhowa, M. A. Rahman, A. K. Paul and R. Ahmmed, "A Narrative Analysis on Deep Learning in IoT based Medical Big Data Analysis with Future Perspectives," 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), 2019, pp. 1-6, doi: 10.1109/ECACE.2019.8679200.

[2] A. Mir and S. N. Dhage, "Diabetes Disease Prediction Using Machine Learning on Big Data of Healthcare," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018, pp. 1-6, doi: 10.1109/ICCUBEA.2018.8697439.

[3] M. A. Sarwar, N. Kamal, W. Hamid and M. A. Shah, "Prediction of Diabetes Using Machine Learning Algorithms in Healthcare," 2018 24th International Conference on Automation and Computing (ICAC), 2018, pp. 1-6, doi: 10.23919/IConAC.2018.8748992.

[4] P. S. Kumar and S. Pranavi, "Performance analysis of machine learning algorithms on diabetes dataset using big data analytics," 2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS), 2017, pp. 508-513, doi: 10.1109/ICTUS.2017.8286062.

[5] P. Bide and A. Padalkar, "Survey on Diabetes Mellitus and incorporation of Big data, Machine Learning and IoT to mitigate it," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 2020, pp. 1-10, doi: 10.1109/ICACCS48705.2020.9074202.

[6] A. I. Ebada, S. Abdelrazek and I. Elhenawy, "Applying Cloud Based Machine Learning on Biosensors Streaming Data for Health Status Prediction," 2020 11th International Conference on Information, Intelligence, Systems and Applications (IISA, 2020, pp. 1-8, doi: 10.1109/IISA50023.2020.9284349.

[7] A. Alahmar, E. Mohammed and R. Benlamri, "Application of Data Mining Techniques to Predict the Length of Stay of Hospitalized Patients with Diabetes," 2018 4th International Conference on Big Data Innovations and Applications (Innovate-Data), 2018, pp. 38-43, doi: 10.1109/Innovate-Data.2018.00013.

[8] U. Ahmed and C. Li, "Machine Learning for Diabetes Prediction," 2021 International Conference on Information and Communication Technology Convergence (ICTC), 2021, pp. 16-19, doi: 10.1109/ICTC52510.2021.9621066.

[9] S. N. Induja and C. G. Raji, "Computational Methods for Predicting Chronic Disease in Healthcare Communities," 2019 International Conference on Data Science and Communication (IconDSC), 2019, pp. 1-6, doi: 10.1109/IconDSC.2019.8817044.

[10] V. Kedia, S. R. Regmi, K. Jha, A. Bhatia, S. Dugar and B. K. Shah, "Time Efficient IOS Application For CardioVascular Disease Prediction Using Machine Learning," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021, pp. 869-874, doi: 10.1109/ICCMC51019.2021.9418453.

[11] V. Vijayaganth; M. Naveenkumar, "4 Smart Sensor Based Prognostication of Cardiac Disease Prediction Using Machine Learning Techniques," in Applications of Machine Learning in Big-Data Analytics and Cloud Computing , River Publishers, 2021, pp.63-80.

[12] W. Wang, B. Jiang, S. Ye and L. Qian, "Risk Factor Analysis of Bone Mineral Density Based on Feature Selection in Type 2 Diabetes," 2018 IEEE International Conference on Big Knowledge (ICBK), 2018, pp. 221-226, doi: 10.1109/ICBK.2018.00037.

[13] G. Sasubilli and A. Kumar, "Machine Learning and Big Data Implementation on Health Care data," 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), 2020, pp. 859-864, doi: 10.1109/ICICCS48265.2020.9120906.

[14] S. Pitoglou et al., "MODELHealth: Facilitating Machine Learning on Big Health Data Networks," 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2019, pp. 2174-2177, doi: 10.1109/EMBC.2019.8857394.

[15] A. Ed-Daoudy and K. Maalmi, "Real-time machine learning for early detection of heart disease using big data approach," 2019 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS), 2019, pp. 1-5, doi: 10.1109/WITS.2019.8723839.

[16] A. Sisodia and R. Jindal, "Exploring the Application of Big Data Analysis in Healthcare Sector," 2017 International Conference on Computational Science and Computational Intelligence (CSCI), 2017, pp. 1455-1458, doi: 10.1109/CSCI.2017.254.

[17] S. Kumar and M. Singh, "Big data analytics for healthcare industry: impact, applications, and tools," in Big Data Mining and Analytics, vol. 2, no. 1, pp. 48-57, March 2019, doi: 10.26599/BDMA.2018.9020031.

[18] S. Athmaja, M. Hanumanthappa and V. Kavitha, "A survey of machine learning algorithms for big data analytics," 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2017, pp. 1-4, doi: 10.1109/ICIIECS.2017.8276028.

[19] M. Sughasiny and J. Rajeshwari, "Application Of Machine Learning Techniques, Big Data Analytics In Health Care Sector – A Literature Survey," 2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2018 2nd International Conference on, 2018, pp. 741-749, doi: 10.1109/I-SMAC.2018.8653654.

[20] P. S. Mung and S. Phyu, "Effective Analytics on Healthcare Big Data Using Ensemble Learning," 2020 IEEE Conference on Computer Applications(ICCA), 2020, pp. 1-4, doi: 10.1109/ICCA49400.2020.9022853.

[21] M. Bochicchio, A. Cuzzocrea and L. Vaira, "A Big Data Analytics Framework for Supporting Multidimensional Mining over Big Healthcare Data," 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), 2016, pp. 508-513, doi: 10.1109/ICMLA.2016.0090.

[22] D. I. Dogaru and I. Dumitrache, "Big Data and Machine Learning Framework in Healthcare," 2019 E-Health and Bioengineering Conference (EHB), 2019, pp. 1-4, doi: 10.1109/EHB47216.2019.8969944.

[23] S. Juddoo and C. George, "A Qualitative Assessment of Machine Learning Support for Detecting Data Completeness and Accuracy Issues to Improve Data Analytics in Big Data for the Healthcare Industry," 2020 3rd International Conference on Emerging Trends in Electrical, Electronic and Communications Engineering (ELECOM), 2020, pp. 58-66, doi: 10.1109/ELECOM49001.2020.9297009.

[24] M. A. Lambay and S. Pakkir Mohideen, "Big Data Analytics for Healthcare Recommendation Systems," 2020 International Conference on System, Computation, Automation and Networking (ICSCAN), 2020, pp. 1-6, doi: 10.1109/ICSCAN49426.2020.9262304.

[25] A. H. A. Balushi, S. I. A. Kazmi, J. Pandey, A. V. Singh and A. Rana, "The Intelligent Control of Street Light System in Oman through Internet of Things Technology," in IEEE 8th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO' 2020), Noida, India, 2020, pp. 694-698, doi: 10.1109/ICRITO48877.2020.9197855.

[26] B. Al-Bahri, H. Noronha, J. Pandey, A. V. Singh and A. Rana, "Evaluate the Role of Big Data in Enhancing Strategic Decision Making for E-governance in E-Oman Portal," in IEEE 8th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO' 2020), Noida, India, 2020, pp. 914-917, doi: 10.1109/ICRITO48877.2020.9197808.