Big Data Analysis using Machine Learning for Health Systems

Deepak Gupta
Department of Computer Science
Engineering College, Ajmer

Shikha Gupta
Department of Computer Science
Engineering College, Ajmer

ABSTRACT
The growth of big data in the healthcare industry presents new opportunities for data-driven decision making, but also poses significant challenges due to the complexity and heterogeneity of health data. Big data analysis itself is challenging task for real time applications and involving machine learning make it more rigid towards the solution. But believing that combining both will provide the effective and efficient solution and intelligence decision making. In this study, we review the current state-of-the-art in big data analytics and the learning model for decision support system using machine learning techniques for healthcare, and we discuss their potential impact on healthcare delivery and patient outcomes. Our framework addresses key challenges in healthcare data, such as missing data, high dimensionality, and class imbalance. We also discuss ethical and regulatory considerations for big data in healthcare, and we propose a set of best practices for ensuring patient privacy and data security. An initial survey has been carried out to define the field for big data. Hadoop tools has been used for data analysis from various hospitals as it provides us an efficient way to store data in a distributed fashion which reduce the high storage personal units like drives and others. Supervised learning with hidden markov model is suggested in this approach for intelligent behavior of proposed model.

Keywords
MapReduce, Machine learning, Q-learning, Hadoop, Parallel processing.

1. INTRODUCTION
In recent years, the healthcare industry has seen an explosion of data from a variety of sources such as electronic health records (EHRs), medical imaging, genomics, and wearable devices. This vast amount of data, commonly referred to as big data, presents both challenges and opportunities for healthcare systems. While big data has the potential to revolutionize healthcare delivery and improve patient outcomes, it also poses significant challenges due to the complexity and heterogeneity of healthcare data. In health caring system a huge volume of big data is generated every day. Parallel processing of these information may be used for several purpose to ease the society. These data can be very important in taking future decision related to medical field scenario or others. Only in India, there are about 35.5 thousands government hospitals and if we assume that 50 people visit on an average in each and every hospital on a single day, then approximately 1.77 M volume of data generated. Including all hospitals data related patients or others may in thousands of gigabytes per day. This huge data can be processed efficiently using many tools to take intelligent future decisions. Machine learning could be a good concept to apply on these big data for efficient and intelligent decision making. Big data is very difficult to process in traditional way as the volume of data is very huge [1]. Collective approach of big data along with machine learning can provide effective solution for data processing with profitable decision support system [2]. In section 2 we are describing about the big data storage technique and management using Hadoop analytical system as it provides the facility of a distributed storage system with parallel processing capabilities. Section 3 explaining the features of machine learning. Proposed approach using machine learning for big data analysis is described in section 4. Finally the conclusion of proposed system is explained in section.

2. STORING BIG DATA
The storage and management of big data in healthcare is a critical component of any big data analytics project. The volume, variety, and velocity of healthcare data generated from various sources pose significant challenges for data storage and management. In addition, healthcare data must be securely stored and protected to ensure patient privacy and data security. For enhancing the effectiveness of big data processing storage must be proper. Traditional storage systems mostly stores maximum information or data on single hardware. Due to this the chance to lose the data in case of hardware failure was more and also it was difficult to process data in comparison to the distributed storage system [3]. Read and write operations for storing and retrieving the data consumes more time as compared to copy operation. Distributed storage system reduces such problems by storing the data in multiple and distributed warehouses [4]. Data can store and retrieve as per requirement in less time form these storage units. Data warehousing involves the storage of structured data in a centralized repository, which can be easily accessed and analyzed. This approach is well-suited for healthcare data that is structured, such as EHRs and claims data. Distributed file systems, such as Hadoop Distributed File System (HDFS), allow for the storage and management of large datasets across multiple servers. This approach is particularly useful for storing unstructured data, such as medical images and sensor data. Cloud-based storage, such as Amazon Web Services (AWS) and Microsoft Azure, provides a scalable and cost-effective solution for storing and managing big data. This approach allows healthcare organizations to store and analyze large datasets without the need for significant upfront investments in infrastructure. Data lakes are a newer approach to storing and managing big data that allows for the storage of structured and unstructured data in its native format. This approach provides a flexible and scalable solution for storing and managing big data in healthcare. Hadoop is one of the big data analyzing tools that provide such kind of warehouse services. It MapReduce function is used in many real time systems. Figure 1 showing the hadoop daemons for data processing [5]. Later subsection shows the elements structure of hadoop as required for proposed approach.

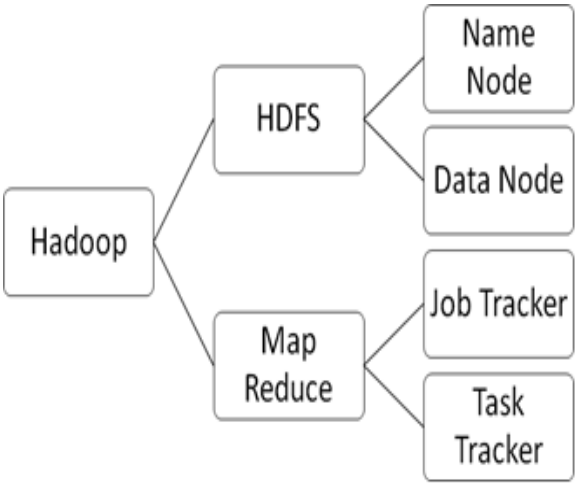

Fig. 1. Hadoop daemons

2.1 Distributed File System of Hadoop

The model of distributed file system of hadoop is designed to work efficiently and effectively on many machines and very huge volume of data sets collectively using commodity hardware [6]. This hadoop file system supports reliability, scalability and fault tolerant data storage facility. HDFS accepts data in many format of files such as text, images, videos and others regardless of system architecture. It performs automatic optimization of high bandwidth streamed data. This allows user to work on various type of data. Below are described basic node components of HDFS architecture useful from the proposed approach perspective [7].

Name Node- It runs at the master node. This node stores and manages metadata about the file system in a particular file named fsiimage and controls the daemons of slave data node for execution of input and output tasks. It also carries out memory and I/O intensive functions in a cluster.

Data Node- Data nodes run at slave nodes. In a cluster there can be more than one data node exist. These node are the basic storage element of hadoop distributed file system where the actual data resides.

Secondary Name Node- This node component read the file system in regular time interval and maintain the log record of changes to update the fsiimage file. Updation in this file is actually supporting the fast starting of name node component for next data. Below figure 2, shows the basic block diagram of hadoop cluster with nodes.

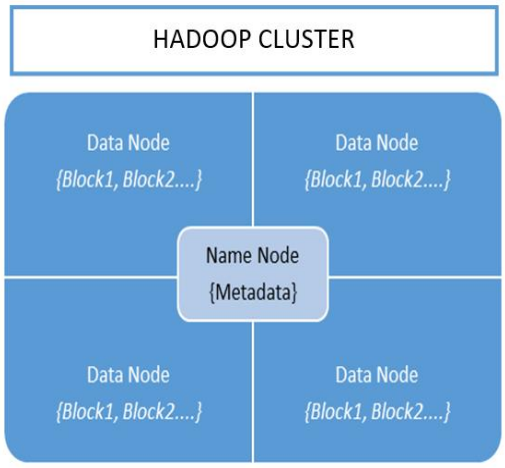

Fig. 2. Hadoop Cluster

2.2 Variety in data

With traditional system when different hospitals uses different types and categories of databases to maintain the records may results into an aggregation problems while using hadoop structure for records management such problems reduces highly. With this data can be store without the restrictions on the type of data [8]. Below are three type of structure for data has been defined in this distributed system.

Structured Data: It is a form of structured data that can be fitted in a data model like a tabular format.

Semi-Structured Data: This is the data that can be considered as a form of structured data that cannot be fitted in any type of formal structure of data models like metadata

Unstructured Data: This comprises the data that has no predefined structure and can be in any format. Like comments, dates, time and others as these cannot be constrained to a particular structure.

2.3 Data Scaling Approaches

The term data scaling means that how the system model is going to manage and accumulate data being generated. There are two types of data scaling described below [9].

Vertical Scaling: In vertical scaling the disk offering higher capacity is used and then shifting the data to that drive. This is the similar way as used in personal computer for increasing the storage capacity. This approach becomes infeasible when data grows exponential and of large value.

Horizontal Scaling: The horizontal scaling for the data storing capacity of system is done by increasing the number of disks or computing nodes [8]. This will increase the capacity of the system. This approach is more useful in the cases where growth of data is in exponential and again a very large value is accumulated.

3. MACHING LEARNING FOR BIG DATA

Collectively using both big data and machine learning for specific field is a challenging task mainly in integration of models [10]. As described in previous section there is variety of data like structured, semi-structured and unstructured required special attention for decision making using big data in health care related units as shown in figure 3.

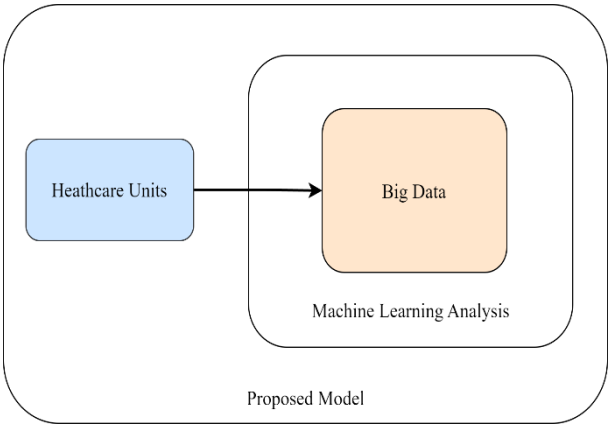

Fig. 3. Big data to machine learning.

Any system be always a leaning system. It can learn to take intelligence decision or smart processing or other works from previous work. This is machine learning. In our approach this learning is applied for intelligent behavior of hospital systems. In today's scenarios machine learning is used in almost all the fields, applications and domains as it covers many problems [11]. Based on data, training, input, output and feedback, learning is categorized majorly into three types. They are supervised, unsupervised and reinforcement learning. Machine learning can be applied in healthcare to analyse big data and extract valuable insights that can help improve patient outcomes, enhance clinical decision-making, and reduce healthcare costs. Some examples of how machine learning is used in healthcare include [12]:

Predictive modeling: Machine learning algorithms can be used to predict patient outcomes based on data from electronic health records (EHRs), medical imaging, genetic information, and other sources. For example, machine learning models can predict the likelihood of a patient developing a particular disease, the risk of readmission, or the likelihood of treatment success.

Image analysis: Machine learning can be used to analyse medical images, such as X-rays, CT scans, and MRIs, to detect abnormalities and diagnose conditions. For example, machine learning algorithms can be trained to identify signs of cancer or other diseases in medical images, potentially improving the accuracy and speed of diagnosis.

Drug discovery: Machine learning can be used to analyse large datasets of chemical compounds to identify potential drug candidates. For example, machine learning models can predict the activity of compounds against specific drug targets, helping to identify promising candidates for further development.

Personalized medicine: Machine learning can be used to analyse individual patient data to develop personalized treatment plans. For example, machine learning models can analyse patient genetics, medical history, and other data to identify the most effective treatments for specific patients.
Overall, machine learning has the potential to revolutionize healthcare by enabling more accurate diagnoses, more personalized treatments, and more efficient healthcare delivery [13, 14]. However, it is important to ensure that machine learning algorithms are properly validated and tested to ensure that they are reliable and effective before they are widely adopted in clinical practice. Based on medical recommendation, survey, and maintenance of store with the learning model the decision for end users is shown in figure 4.

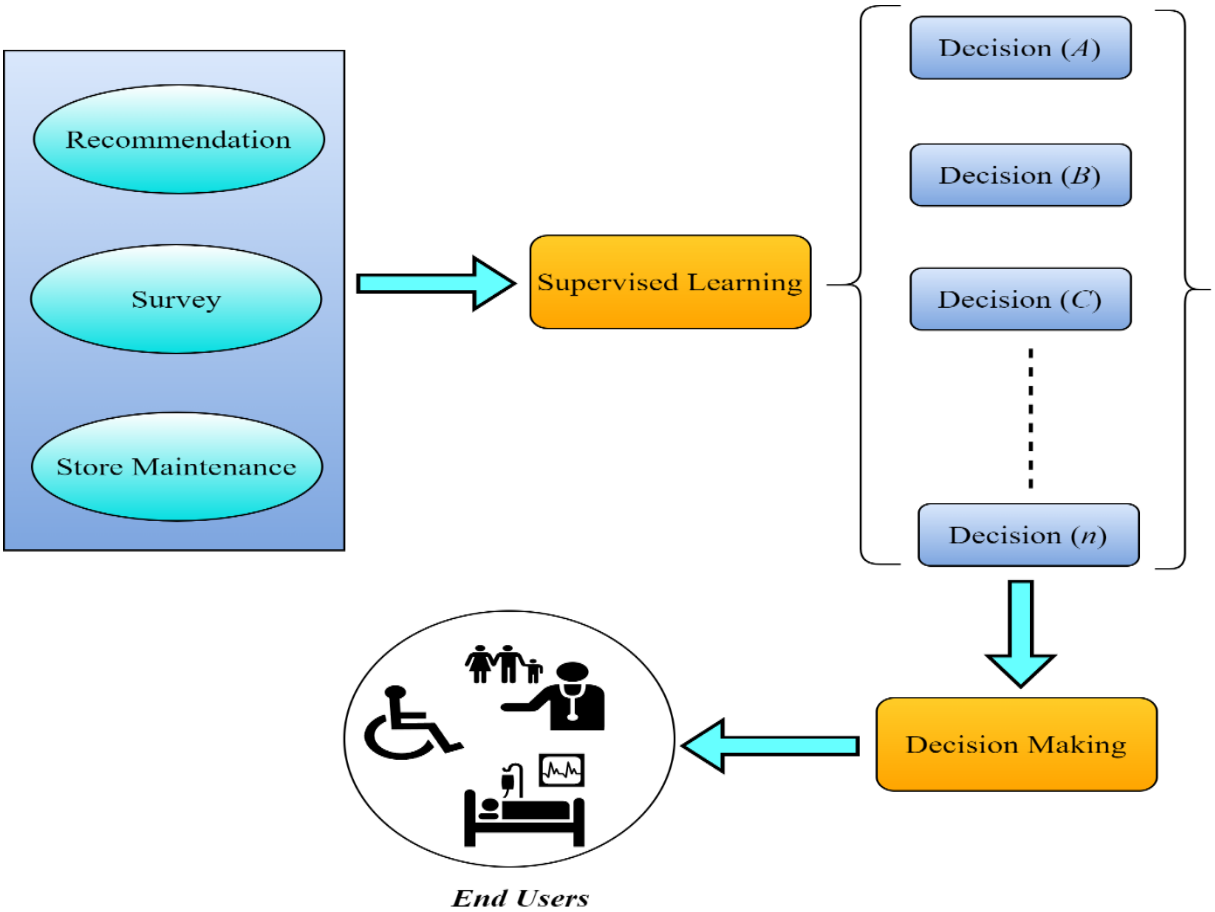

Fig. 4. Learning Model for Proposed Work

Table 1. Learning categories

Learning Type	Labelled Data	Input	Output	Feed-back
Supervised	✓	✓	✓	X
Unsupervised	X	✓	✓	X
Reinforcement	✓	✓	✓	✓

Table 1 represents the learning categorization of learning type for labelled data related to healthcare big data. There are many ongoing research efforts and proposed approaches for applying machine learning to the analysis of big data in healthcare [15-17]. Here are a few examples:

Federated learning: This is an approach where machine learning models are trained on data from multiple healthcare institutions while keeping the data local and secure. This allows for the creation of more robust and accurate models, while ensuring data privacy and security.

Graph-based learning: This approach involves representing healthcare data as graphs, where nodes represent patients, diseases, treatments, and other variables, and edges represent relationships between them. Machine learning algorithms can then be used to learn patterns and relationships within the graph, enabling more accurate predictions and diagnoses.

Reinforcement learning: This approach involves training machine learning models to make decisions and take actions based on feedback from the environment. In healthcare, this could involve training models to make treatment recommendations based on patient data, and then adjusting those recommendations based on the patient's response.

Explainable AI: This approach involves developing machine learning models that can provide explanations for their decisions and predictions. In healthcare, this could help clinicians understand why a particular treatment was recommended, and provide insights into the underlying factors that contribute to a patient's health.

4. PROPOSED APPROACH

In this approach we are providing the concept for ease of society in hospitals by using big data with the help of machine leaning. For estimation and classification of stored data a support vector machine or hidden markov model can be applied for supervised learning and for intelligent decision making on them a model free Q or R- learning can be used. The raw data from various hospitals having records of daily patient monitoring, stocks, employees management and others have been collected in distributed system. Firstly, the data is collected in our distributed system from different hospitals. As we are working only on medical representation, survey and stocks related activities therefore we are only collecting raw data related to them only. After receiving data has been aggregated into system and load in form of Resilient Distributed Dataset (RDD) for further processing. Following substructure need to understand for big data processing as shown in figure 5.

Medicine Recommendation: - This structure is related to end user requirements in which aggregated data can be analyzed using query, clean, process and result stages.

Process data: - This substructure is finally processing and analysing the filtered data to produce desired output as per used requirements and this data is processed to produce out- put. This is the step in which the analysis is actually performed.

Show result: - This displaying the desired output to user.

Survey - This field can be used for analysis of existing data sets for further decisions and survey. We used Spark SQL tool for this purpose on collected data: In survey structure following procedures and parameters are used to carry out analysis on data by analyst. Query: - In this, according to tasks need to performed on data a query is invoked by the approach model.

Clean: - Whatever the data collected needs to filter as per desired work and tasks. For this a clean procedure has been setup to filter the data which is irrelevant to our analysis requirement.

Field name: - It is the part where the analyst is going to feed the system parameters for analysis.

Clean: - In this the data is irrelevant to our analysis requirement is filtered.

Analyze: - Analyst set some query to analyze the data for survey. Here data has been process to produce output according to the query. This is the step in which the analysis is actually performed.

Result: Here we generate the graph of output. This graphical representation of survey provides better insight and supports in comparison with others fields as per requirement.

Stock maintenance: With this a requirements prediction is possible. The system can be used to provide an idea or alert to the management of the hospitals about the possible shortage of the stock of a certain medicine or some other hospital. Stock related things that will be needed to be restocked in the near future in order to avoidance the chaos due to some type of shortage. Below are suggested procedure for this.

a) Proper maintenance of data about the stock.
b) Use the prescription data to calculate the consumption of a particular medicine.
c) Set an optimal period for monitoring the consumption rate of medicines.
d) If the quantity of a medicine is gone below a set floor value of stock or consumption is increased significantly and is still increasing in some fashion.
e) Calculate the time of expected exhaustion of stock of the medicine using mathematical functions on pattern observed in data in last step.
f) Notify the management about the expected time the stock might get exhausted.

After these structure a supervised learning model has been applied in our model for various data sets.

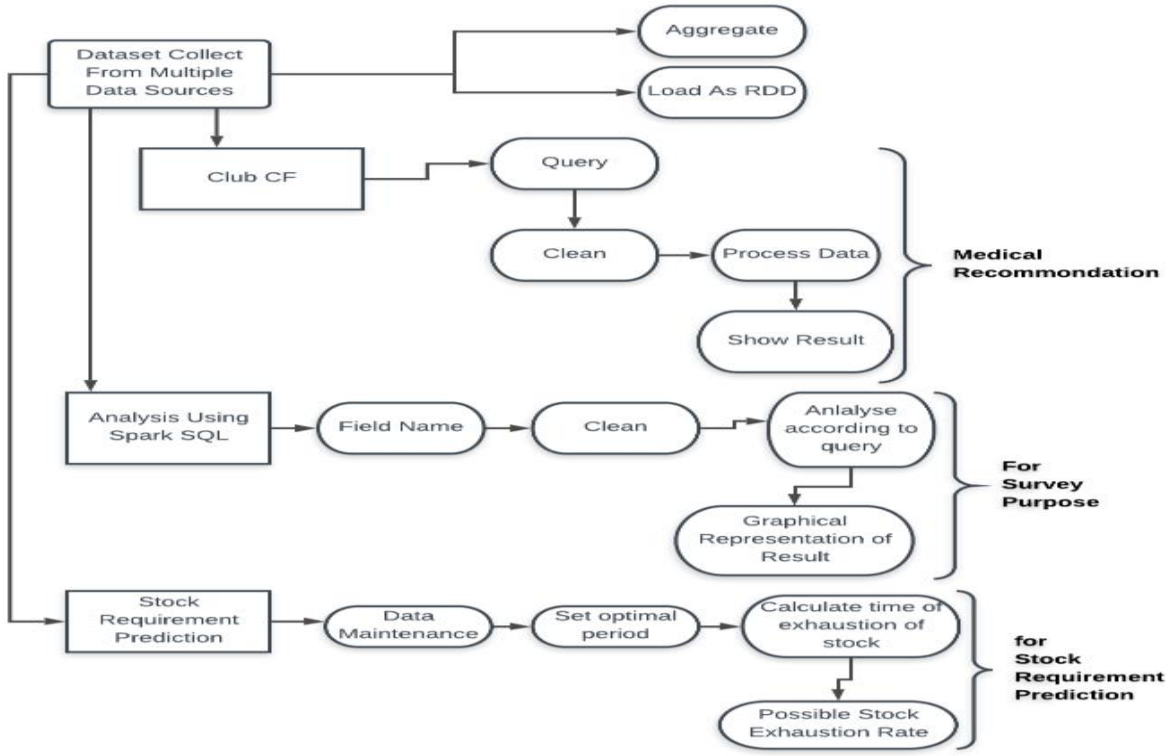

Fig. 5. Flow of Proposed Work

5. CONCLUSION

Here we get a very efficient system for analysis in parallel and storage of big data in a distributed environment. The system provides us to remove the need of very high-end system which may result in costing a lot in setting up the system as everything has to be stored on a single machine and has to be processed on a single system which will also result hi high computational cost increasing the time taken to analyze. By using this we can fulfill a lot of our management and analysis requirements on a moderate costing system. This system is going to let us perform some specific type of recommendation tasks.

6. FUTURE SCOPE

There are some work that can be include in this for future aspects such as a predictive analysis system can be embedded into this model or a streaming order like spark can be used to process real data of other applications. Other learning like's an unsupervised algorithms can be applied on this model to compare the efficiency of results.

7. REFERENCES

[1] Zhou, L.; Pan, S.; Wang, J.; Vasilakos, A.V. Machine learning on big data: Opportunities and challenges. Neurocomputing 2017, 237, 350–361.

[2] Fan, S.K.S.; Su, C.J.; Nien, H.T.; Tsai, P.F.; Cheng, C.Y. Using machine learning and big data approaches to predict travel time based on historical and real-time data from Taiwan electronic toll collection. Soft. Comput. 2018, 22, 5707–5718.

[3] Hanzelik, P.P.; Gergely, S.; Gáspár, C.; Gy″ory, L. Machine learning methods to predict solubilities of rock samples. J. Chemom. 2020, 34, 1–13.

[4] A Sandryhaila, JMF Moura, Big data analysis with signal processing on graphs: representation and processing of massive data sets with irregular structure. IEEE Signal Proc Mag 31(5), 80–90 (2014).

[5] Q Wu, G Ding, Y Xu, S Feng, Z Du, J Wang, K Long, Cognitive internet of things: a new paradigm beyond connection. IEEE Internet Things J 1(2), 129–143 (2014)

[6] Menshawy, A. Deep Learning by Example: A Hands-on Guide to Implementing Advanced Machine Learning Algorithms and Neural Networks, 1st ed.; Packt Publishing: Birmingham, UK, 2018.

[7] R. Kune, P. K. Konugurthi, A. Agarwal, R. R. Chillarige and R. Buyya, "The anatomy of big data computing", *Softw. Pract. Exper.*, vol. 46, no. 1, pp. 79-105, 2016.

[8] J. Fan, F. Han and H. Liu, "Challenges of big data analysis", Nat. Sci. Rev., vol. 1, no. 2, pp. 293-314, 2014.

[9] J. D. D. Lavaire, A. Singh, M. Yousef, S. Singh and X. Yue, "Dimensional scalability of supervised and unsupervised concept drift detection: An empirical study", Proc. IEEE Int. Conf. Big Data (Big Data), pp. 2212-2218, Oct. 2015.

[10] C Rudin, KL Wagstaff, Machine learning for science and society. Mach Learn 95(1), 1–9 (2014)

[11] CM Bishop, Pattern recognition and machine learning (Springer, New York, 2006)

[12] B Adam, IFC Smith, F Asce, Reinforcement learning for structural control. J Comput Civil Eng 22(2), 133–139 (2008)

[13] Shailaja, K., Banoth Seetharamulu, and M. A. Jabbar. "Machine learning in healthcare: A review." *In 2018 Second international conference on electronics, communication and aerospace technology (ICECA)*, pp. 910-914. IEEE, 2018.

[14] Qayyum, Adnan, Junaid Qadir, Muhammad Bilal, and Ala Al-Fuqaha. "Secure and robust machine learning for healthcare: A survey." *IEEE Reviews in Biomedical Engineering* 14 (2020): 156-180.

[15] Bhardwaj, Rohan, Ankita R. Nambiar, and Debojyoti Dutta. "A study of machine learning in healthcare." *In 2017 IEEE 41st annual computer software and applications conference (COMPSAC)*, vol. 2, pp. 236-241. IEEE, 2017.

[16] Dhillon, Arwinder, and Ashima Singh. "Machine learning in healthcare data analysis: a survey." *Journal of Biology and Today's World* 8, no. 6 (2019): 1-10.

[17] Toh, Christopher, and James P. Brody. "Applications of machine learning in healthcare." *Smart Manufacturing: When Artificial Intelligence Meets the Internet of Things* 65 (2021).