Machine Learning in Modern Healthcare: Leveraging Big Data for Early Disease Detection and Patient Monitoring

Karthik Chava

DevOps Engineer
Email: *kkarthikchava[at]gmail.com*
ORCID ID: 0009-0007-6001-4526

Abstract: ***Healthcare has witnessed significant transformations in recent years, fueled by developments in smart technologies and the Internet of Things. Modern healthcare is distinguished by remote monitoring, wearable biosensors, and timely personalized healthcare services in a rapid feedback loop. Recently, smart healthcare has opened applications in disease diagnosis and disease treatment, such as health basic assessment and detection of daily human activities. This has motivated healthcare organizations and different organizations to invest in the new healthcare sector of big data analytics. Big data in smart healthcare has unique and influential characteristics, which give rise to specific and challenging issues, and also requires new techniques, technologies, and platforms for outbreak area monitoring and emerging disease detection. In addition to the above challenges, a newly raised challenge for sensing-based healthcare applications is disease diagnosis. Rapid detection of human activities and its associated illness is helpful for taking timely actions for patients' safety assurance. In certain cases, the human activity is too complicated to be solely defined by experts, or operate according to experts' rules. New techniques relying on the concept of computational epidemiology are broadly applied, such as correlation measure, spread people simulation process involving parameter estimation, and state-space model-based and distance model-based spread event forecasting. All such techniques rely on event categorization on the very first step. However, outbreak detection in a new geographical area hardly needs human-doing-based categorization processes because of matching the geography and available data set's inference before consulting experts. These challenges make the design of a big data analytical platform for fault data storage management, rapid emergency detection, and disease diagnosis in exactly the new area difficult and also necessary. This paper proposes an efficient and robust big data analytical platform for processing real-time and sensing-based healthcare applications. The platform deploys big data and related technologies to addressing the challenges raised by irregular and large quantity data streams from biosensors. The platform consists of many layers, which include storage management, real-time control, decision-making, patient classification, disease diagnosis, and data retrieval. Each layer is analyzed in detail, and case scenarios are taken to demonstrate the platform's applicability and plausibility on real health data.***.

Keywords: Machine learning, modern healthcare, big data analytics, early disease detection, patient monitoring, predictive modeling, healthcare informatics, real-time health data, clinical decision support, medical diagnostics, AI in healthcare, health data mining, personalized healthcare, electronic health records (EHR), anomaly detection, health prediction models, remote patient monitoring, chronic disease management, supervised learning, healthcare data integration

1. Introduction

Recent advancements in sensor technologies, information processing, and data transmission have opened new horizons to a vast number of applications in the health monitoring and disease diagnosis domains. Such applications include heart diseases detection and monitoring, blood glucose monitoring, and monitoring Parkinson's disease symptoms. On the other hand, smart technologies, sensing devices, and low-cost wireless networks are the main factors fostering the development of smart health applications. As a result, healthcare organizations and hospitals are starting to invest in big data analytics to obtain business value from their data. Besides this growth, sensing-based healthcare applications face many challenges. The first challenge is big data collection and storage. With the emergence of thousands of new devices every day, the number of sensed parameters and the number of patients is expected to grow massively over the coming few years. This massive data collection is accompanied by high speed generation and a heterogeneous nature. Presently available traditional technologies for data storage cannot efficiently handle such data. The second and very important challenge of such applications is that dangerous medical emergencies can occur any time, and rapid detection of them is very crucial for the safety of patients. Thus, medical records outside the normal health ranges need to be accurately detected and reported to the medical staff in a very short time. The third challenge of healthcare applications is patient classification and disease diagnosis. The final output of healthcare applications needs to be a diagnosis of the existing diseases and predicting the type of the patient to aid medical staff make correct decisions.

Today, health monitoring and disease diagnostics are considered as prevalent applications of smart technologies. In these applications, drowning patients data is a crucial means to increase accuracy in classifying patient diseases. However, analyzing such data can be considered a risky operation, as any wrong conclusion or interpretation can cause death or disability of a patient. Consequently, adequate mechanisms are needed to conduct such analysis, speedily and accurately to reach higher degrees of precision in health diagnosis. In this regard, a big data analytical platform is proposed. The developed platform can assist to reach this goal, by analyzing patients' parameters, symptoms, and diseases to extract the correlation between patients' symptoms and the diseases and classify them.

2. The Role of Big Data in Healthcare

The Obesity Epidemic is transforming healthcare as we know it. A sea change in patient treatment is occurring as physicians who have long depended on scans, blood tests, and personal expertise are set to receive assistance like never before: the ability to analyze a patient's health data and diagnose diseases up to years before any symptoms present themselves. This problem has given way to a new era of "Big Data" in which scientists are exploring new ways to understand the large amounts of unstructured data generated by modern technologies. Despite these advances, we see few examples of Big Data being leveraged in healthcare despite the opportunities it presents for creating personalized treatments. The increasing number of patients and new diseases makes health monitoring a complicated task. The processing of big and heterogeneous data collected by biomedical sensors, along with the need for patient classification and disease diagnosis, are major challenges. Combining remote sensing devices with big data technologies has proven to be an efficient and low-cost solution for healthcare applications. Despite these advancements, few examples of this kind of patient monitoring and disease diagnosis are found in the public healthcare landscape.

Effective use of Big Data in Healthcare is enabled by the development and deployment of machine learning (ML) approaches. ML approaches are often interchangeably used with artificial intelligence (AI) approaches. An AI approach describes the broader sphere of machines that are made to "think" or "act" like humans, while ML specifically denotes a subfield of AI where machines are able to adapt their behaviour through experience. The latter definition is more apt in the context of healthcare as it describes the types of algorithms that are more commonly employed. It is well established that diseases and their progression are manifestations of the interactions of a multitude of factors including genomic, epigenomic, proteomic, transcriptomic, exposure, and environmental factors. Application of ML tools is supplemented by the widespread adoption of Electronic Health Records (EHRs). EHRs allow patient data to become more accessible to both patients and a variety of physicians, but also researchers by allowing for remote electronic access and easy data manipulation.

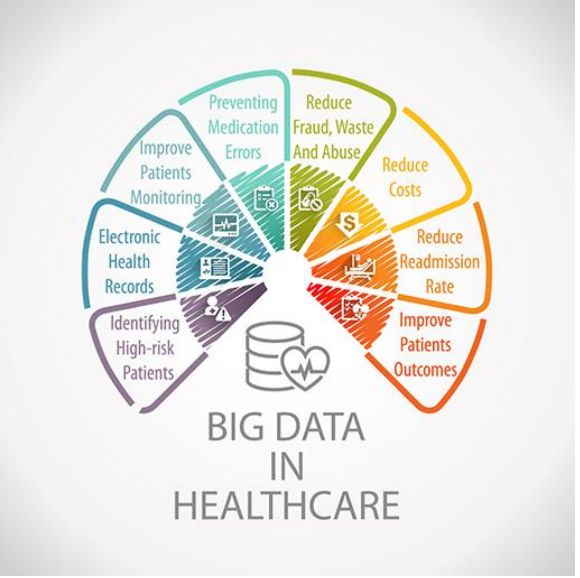

Figure 1: Use of Big Data in Healthcare

3. Early Disease Detection Techniques

The early detection of diseases is achieved through an array of different methods based on careful radiological inspection. New diseases are frequently discovered and spread in large populations, impairing further advances in healthcare fields. On the other hand, a condensed and comprehensive detail set of the individual's health contingency can further enhance precision in prediction about an impending or existing spectrum of diseases. On this front, it has become evident in recent years that multiplex matrix-affinity encodings of the body's biophysical signals forms a 'deep' and intrinsic pattern of the patients' state. One overwhelming advantage of this deep understanding is that they are capable of inferring the trajectory of the healthy individual towards different pathological conditions without the reliance on the detailed understanding of the biophysics of each case.

Machine learning has become a synonym of big data. It provides tools to customize well-founded generative statistical models to leverage unseen data, thereby extracting its hidden representations and information. The far larger and diverse scope of possible pathologies and the more fundamental challenge of careful filtering to identify features, which will become sources of predictive information have limited the possible harnessing of this approach in the context of health signal monitoring. However, advances in algorithms and education have resulted in a renaissance of interest to apply machine learning in this front. The methodological advances have provided means to define a 'universal' methodology, which can flexibly incorporate differences in the sensors and encoded signals. Thereby turning their scalar and disconnected waveforms into an intrinsically compact metric space of their biophysical patterns and the cohort they belong to.

3.1. Predictive Analytics

Predictive analysis techniques are utilized in the healthcare sector to aid in early disease forecasts. Healthcare practitioners employ predictive models to anticipate the beginning, worsening, or advancement of ailments before they manifest in patients. For instance, predictive models established on Electronic Health Records (EHR) are designed to detect changes in patient health status and provide alerts to clinicians in advance. Because EHRs are probably the most significant data source used in these predictive models, an early warning system was proposed using the bidirectional gated recurrent units and attention-based architecture to identify COVID-19-positive cases based on EHR data exhaustively [4]. A system for identifying the onset of diseases, such as diabetes, hypertension, and COPD, using EHR and a hybrid of medical embedding and GRU architectures has also been proposed. The recommended system not only identifies diseases with a high F1 score but also predicts diseases among patients for whom models are not constructed. Furthermore, a robust and explainable model elucidated based on Self-explanatory/SHapley Additive exPlanations principles confirmed the critical role of diabetes-related medications. Existing predictive analysis models concentrate on personal health records. However, there is a lack of models that can be applied to comprehend the broader reasons affecting health at the population level.

Hand-crafted health features, such as physical health, mental health, and healthcare utilization, should be encoded with deep learning models over time to evaluate chronic disease prediction. Nonetheless, few predictors permit simultaneous mining of extensive primary records of numerous individuals, and none aim to ascertain the onset of diseases. The memory-based remembrance of global health status and changes in population health conditions is crucial for health planning. It is acceptable to ponder how the general population's health would look after a period given the present health status of the population. It is also a necessity that the described health features abide by privacy regulations. These assessments are particularly useful for monitoring the alteration in the health conditions of a population with large amounts of heterogeneous EHR data and finding a feasible way to encode population health status and timestamped changes with deep learning models, for now, remains unexplored.

3.2 Image Recognition

A large class of algorithms can be found under the umbrella term of Machine Learning (ML), which covers methods that learn from data. The objective of ML is to find a mapping, called the hypothesis, that transforms a vector in the input space into a vector in the output domain. A dataset consists of data points (vector x, label vector y). Each data point xi can be seen as an input space point xi ϵ Rn to which a label yi ϵ Rm is assigned by the expert. Each label vector can represent primitive data or a classification. ML can be seen as an optimization problem. Data points comprise the training set, and the goal is to find a model that can generalize and have a low prediction error on unseen data. The model is required to store the training data in the form of parameters that are often called weights [6].

Given that the training set is often influenced by noise, the goal becomes one of balancing generalization and modeling capacity. If the model has fewer parameters than the training set dimensions, it is likely to obtain a low training error, whilst a large model is more capable of obtaining a low training error due to overfitting the noise present in the data. The change in weight parameters after observing input data points follows the Stochastic Gradient Descent (SGD) rule and iteratively minimizes the objective loss function over a fixed number of iterations or until convergence. Convergence is reached when the norm of the gradient is lower than a fixed threshold or when the relative change of cost is below a value. The hyperparameters are also estimated.

The methods to learn from data can typically be categorized as 1) supervised and monitored on labeled data, 2) unsupervised on unlabeled data, 3) weakly supervised or weakly labeled on partially labeled data, and 4) semi-supervised on small amounts of labeled data. While significant research has been dedicated to supervised/unsupervised learning, the categorization of semi-supervised ML is still an open area, often defined as any combination of weakly/unpartially labeled/limited datasets for supervised learning.

3.3 Genomic Data Analysis

The experience of developing AI based products using genomic data has been pleasant for the authors. Genomic data is primarily of two types DNA sequence and annotation information of genes and gene products. The objective of Genomics is to take the DNA code from sequencing instruments and transform it into biology by getting it interpreted in terms of health and disease. However, the harder task than sequence acquisition is its interpretation. Machines can infer the sequence, but they cannot ascertain if a sequence means disease or health, as such information is in the realm of biology. Internally or by observing marker effects, genomic variants influence several pathways and systems in the body that ultimately affect human health. These genetic mechanisms are complex. Data obtained from the SNP microarrays are signals from different blood collections and substances in the blood. Processes of modelling events at evolutionary scale, the body's anatomy and its components, physiology, and technology that manufactures chips and components and designs platform, probes, reagents, etc., all contribute to underlying mechanisms.

Genomic data acquired either from microarrays or sequencers as table information consists of a lot of information. With the availability of diverse data types – gene expression, SNP genotypes, demographics, health history, and laboratory findings – machine learning algorithms have become the obvious choice for accurate prediction of disease risk and personalized treatment. Numerous machine learning methods, such as support vector machines, random forests, and Bayesian networks, are being used successfully in genomics research and applications. Deep learning and machine learning algorithms, which come under the umbrella term Artificial Intelligence (AI), are being used in clinical practice through numerous commercial applications involving clinical and genomics data. High throughput data acquisition technologies such as Next Generation Sequencing are generating a wide variety of genomic data with increasing volume. A well known personal genomics company uses machine learning algorithms in disease risk prediction. Many startups are increasingly using the combination of machine learning algorithms and genomics in creating tools and processes that enhance healthcare systems.

In India, very few organizations use machine learning algorithms in clinical genomics, with the reasons being lack of awareness and lack of expertise in research and application of AI. Some of the Indian pharmaceutical and genomic organizations that are using AI include Innoplexus, Lantern Pharma, Manipal Group of Hospitals, TCS Innovation Labs, BioXcel Therapeutics Inc., Mapmygenome, OncoStem, and PierianDx. Main technical challenges in the application of ML algorithms are data curation and data pre-processing. Different hospitals and laboratories adopt different terminologies to record a disease or a health condition and use different reference ranges. In India, Electronic Health Records Standards were released by the Ministry of Health and Family Welfare in 2016. Data sets used in training the machine learning algorithms should clearly represent the target data for which risk predictions are made.

4. Patient Monitoring Systems

Patient Monitoring Systems (PMS) provide continuous patient data collection and analysis using wearable or non-wearable sensors to improve patient care and medication safety. With normal condition monitoring via wearables, symptoms can be alleviated before they worsen, and medication can be administered on demand. A smart and portable patient monitoring system is thus developed using different types of sensors and hardware modules, and the sensed data is processed at the cloud. Another PMS uses a wireless body area sensor network to monitor patient data such as body temperature, heart beats, and ECG signals. If the patient status exceeds the predefined ranges, the patients get alarmed, and physicians can monitor their patients via a web page. Moreover, wearable health monitoring devices have gained attention in recent years due to their low costs and improved performance and accuracy. They can continuously capture data in real-time and allow easy assessment of a patient's health remotely. Monitoring devices are used to communicate patients' health status with clinics. Any sudden increase in measured values can push immediate alerts to physicians for timely action. Pervasive sensing and processing in data-driven healthcare are promising for better proactive prevention, real-time intelligence, and optimized treatment. Such systems lead to personalized healthcare that improves patient outcomes while dramatically lowering costs. A monitoring system also provides an integrated cloud-based platform for data acquisition, storage, and analysis where the machine learning model learns sequence actions for decisions.

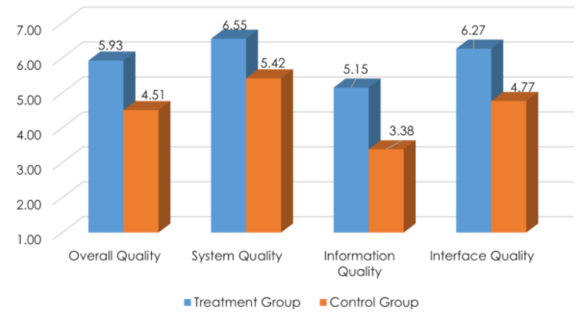

Figure: Visualization of Remote Patient Monitoring System Based on Internet of Medical Things

Researchers developed an intelligent IoT-based sensing inflexible healthcare system to monitor diabetes patients. The system consists of different types of smart sensors that are connected to wearable devices to collect patient health condition data. These devices are designed flexibly to fit in various clinical environments. The acquired data are processed locally at the edge level and transferred to the cloud server for storage. Data analytics based on the deep neural network can provide two types of detection decisions: real-time and batch mode. The system can also support an anomaly detection method for patient health status diagnosis.

4.1. Wearable Technologies

Wearable Sensors to Monitor Physiological Parameters and Activities. Wearable sensors have been researched and developed by researchers and technology companies over the last decade to monitor physiological parameters and activities. In this light, wearable sensors have undergone a significant evolution over the past decade; the interest from the scientific community, technology companies, and healthcare professionals has increased [10]. These sensors can record real-time information from the user. The wearable can be a usable gadget and clothing. Wearable sensors can register physiological signals such as heart rate monitoring, body temperature, pulse oximetry, breathing rate, and body movement monitoring. Body motion monitoring has different approaches such as 3-axes accelerometer and gyroscope sensors in a single unit, or just accelerometer sensors since it is widely used in phones. The majority of the coordinated models require real-time feedback with a medical approach. However, real-time feedback is useful for both patients and health professionals. For patients, they need to convert the information to better understand their disease and see the immediate and objective results from the actions taken. This will help patients to improve their behavior and give them the tools to be empowered to make decisions. Health professionals need access to patients' data for the same reasons but have to observe, assess, and provide personalized advice. Also, improve adherence, predict events, prevent disease, early diagnosis and chronic control conditions. There are still challenges facing the way sensors and systems are developed. As these continue to decrease in size and cost, and have enhanced capacities, there is an enormous potential to increase sensors and wearables in clinical trials, accelerating this path for new knowledge and new treatments. This path relies on tools to collect, store and manage data, allowing trusted results on population groups. This info could arrive at the patients' phones providing direct feedback to what is monitored, and sending alerts and which information is needed to provide even more significant and rapid results for users. This path also includes education on using this technology, translating data to knowledge, and waiting for the readiness of healthcare professionals to include new routines and best practices. Above all this, there is essential concern about less controllable events such as fraud and theft, data safety, and government regulation. This is vital to improve the systems' acceptance within different responsibility firms, companies, and health institutions.

4.2. Remote Patient Monitoring

Recent advancements in big data and data acquisition technology, such as wearable health sensors, which can monitor a patient's physical signals, have prompted a considerable rise of interest in remote patient monitoring (RPM), a telemedicine branch. It addresses the necessity to everyday monitor patients in an affordable manner using automatic equipment and aims to reduce the cost of surveillance. Patience's heartbeat, blood pressure, oxygen level, and other physical signs are monitored using wearable smart sensors, and then patients' conditions can be periodically evaluated based on the records. In the early stage of RPM, sensed signals such as received sound and light from instruments were analyzed to detect certain diseases based on applying filters and detecting the pulse of signals or the variation of stimulated signals. However, due to the complexity of the analysis and the noise of the sensing process, it is easy to receive inconsistent glimpses of information or even miss fatal cases, yielding a lower precision and recall performance. Beyond analyzing a limited kind of signals, more advanced RPM systems have been recently proposed, where RPM resolution was leveraged to address the complexity challenged beyond the human ability to process the near infinite nodes of RPM sensors by exploiting the wisdom of machine learning. In this way, a great battle for health care knowledge discovery has commenced. The potential and values of monitoring and understanding patients with either type of signals is enormous, and many recently proposed modes, such as monitoring personal mobility using devices on the body, pocket or hand, exhibit larger promising solution spaces for developing novel sophisticated RPM techniques.

5. Data Privacy and Security in Healthcare

Advances in data collection and information technology have driven the emergence of data-driven healthcare systems, which collect data using various data collection devices and use it to provide services. Nowadays, many patients with chronic diseases monitor their health status using wearable sensors equipped with a wide variety of data collection devices, such as 1-D and 3-D accelerometers, ECG sensors, glucose sensors, and more. Alternatively, patients can now subscribe to monitoring services or emergency detection platforms that use wearable data to detect health anomalies and trigger hospitals' visit. The emergence of these services is expected to save many lives, but it saves a cost beyond the subscription fee - ways must be established to protect patients' privacy before these types of services are widely accepted and deployed. The rapid growth of data-driven data gathering and processing systems goes hand-in-hand with a growth of privacy concern, where institutions that collect this data are required to consider the protection of data rights and govern this data responsibly. But protecting privacy in data-driven healthcare systems comes with unique challenges and requirements.

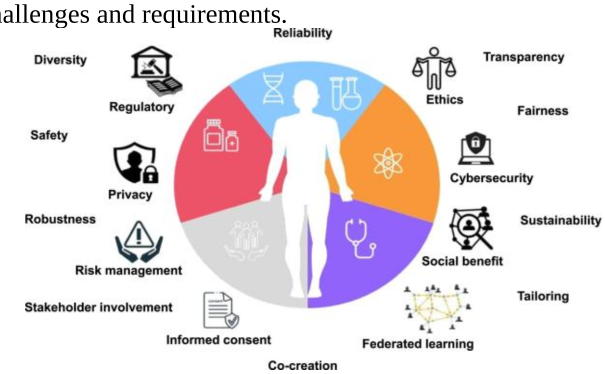

Figure 2: Computational Nuclear Oncology Toward Precision Radiopharmaceutical Therapies

The advantages of utilizing data to construct data-driven healthcare systems come with an effort to protect patients' privacy without jeopardizing the utility of the data. Protecting privacy in data-driven healthcare systems is both urgent and challenging due to the high sensitivity of the data, whether it is in those health data's characteristics or in their identification. In the United States, three essential laws protect patient health data privacy from abuse and unauthorized exposure, but exposure of sensitive healthcare data violates HIPAA and can change a patient's life forever. Privacy must be preserved throughout all data usage purposes, including storing, processing, and analyzing data safety. Healthcare teams, who treat their patients, need to protect their data while processing it. Before treating them, healthcare specialists would like to check if anything serious can prevent them from treating patients safely, or if some heartbeats can be ignored with a constant sign off. These checkups wouldn't require the data itself, but the detailed analytics instead.

5.1 Regulatory Frameworks

The existing regulatory frameworks relevant to building, launching, and marketing AI systems in the regulated health space, including in day-to-day clinical settings and at relevant stages of integration or use. There is an overview of the proposed European Health Data Space, including mechanisms for access to public and private health data. The overall impact of proposed legislation on the regulatory oversight of AI models for EHR segmentation, dimension reduction, and imaging applications is discussed. Then, summarizing existing regulations on these applications in other geographies, including the USA, England and Wales (UK), and Canada.

The following summarizes the machinery determining the approval of AI-based medical devices in the USA. Generally, premeditated use of AI tools must integrate into clinical practices should proceed with the premarket (510(k)) notice. FDA Center for Devices and Radiological Health (CDRH) staff typically provide preliminary feedback on pivotal issues affecting the regulatory strategy within 60 days of this submission. Key concerns include definitions of target population, modalities, clinical tasks, performance metrics, reference AI results, and gold standards. Considerations of AI specifics are incorporated into data generation, performance quantification, and all AI-related terminology. Due to the modest and tentative nature of biases in preprocessing, additional data augmentation techniques and hyperparameter tuning have been proposed to avoid possible performance decrease when generalizing the independent test set. Two tools are now available to support similar analyses routinely conducted for radiomic feature extractions or post hoc provisions of text or DNA variations.

Considerable efforts are being devoted worldwide to designing an ongoing regulatory framework addressing challenges posed by widespread adoption of AI in healthcare. Since 2009, proposed AI systems, including deep learning and other machine learning technologies, subsumed under the term Software as a Medical Device (SaMD), are assigned to a unique classification based on a combination of risk categorization and a reference to intended use [14]. Acquisitions and intended uses relevant to health information technology are all under the preview of the US Congress-established Federal Trade Commission (FTC). European Commission formulation steps are detailed to ensure compliance with Assembly Act on Digitized Health Data as access to Europe's health data is democratized, decentralizing and federated systems being established to empower data holders and communities on data ownership. Analyses of health regulation in England/Ireland and Canada closely follow similar templates.

5.2 Data Encryption Techniques

AI holds great promise for improving human productivity and enhancing quality of life. Today, applications of AI are increasingly impacting a greater diversity of society, including transportation, retail, logistics, finance, oil and gas, industry, healthcare, and telecommunications. AI is impacting all these industries by finding a place in revelation of things, automated reasoning, robotic process automation, natural language processing and digitally connecting things among others. In healthcare, it can aid in early disease detection, patient monitoring and population health analytics among others.

Equation 1 : Ciphertext Expansion Ratio (CER)

$$CER = \frac{S_c}{S_p}$$

- S_c: Size of the ciphertext
- S_p: Size of the plaintext
- $CER > 1$ indicates overhead due to padding, headers, etc.

Healthcare is one of the most complex and fastest growing fields in society, whose constant evolution results in the obtaining of immense amounts of data in the participants' daily activities. The continuous gathering of data in the healthcare sector is produced by various sources, such as the Internet of Medical Things, Electronic Health Records or any service related to patient health monitoring. All new devices and services produce an enormous amount of informative data in the form of Medical Data dedicated to the description of the patients' state of health, which further can be transformed into valuable Information. The data gathering process is accelerated by a faster, better and cheaper infrastructure and by open source analysis tools, allowing the necessary processing of big data, which otherwise cannot be solved. A commonly held belief is that big data is one of the greatest assets that a person can own, and this belief also applies in the medical field.

The analysis of the expansive amount of data produced by either patients or hospitals can lead to the extraction of important insights regarding the patients' state of health or regarding a particular treatment procedure. Abundant information, insights are valuable in providing better healthcare services to the patients. Not all existing data sources are being monitored and a malfunction in one of the monitoring components might affect the processing of information due to the lack of another possible data source. The primary objective is to create a framework for knowledge extraction from all the Medical Data by designing dedicated tools that will both provide useful Information regarding the patients' state of health and will consider and preserve the integrity of the data. The proposed architecture consists of data acquisition, analysis and visualization premises, thus covering all the healthcare aspects from the patients' point of view. This framework aims to support clinicians in making timely, data-driven decisions by integrating diverse data sources such as electronic health records, wearable devices, and diagnostic imaging. Additionally, strong emphasis is placed on compliance with privacy regulations like HIPAA, ensuring that patient confidentiality is maintained throughout the data lifecycle.

6. Challenges in Implementing Machine Learning

While ML-based applications exploring the potential and thriving opportunities in health care are available, unique risks and challenges come along. First, and foremost, the roll-out and upscaling of health care applications incur costs. The investments required to run these technologies may not be feasible for institutions without sufficient funding. Most emerging countries will not meet these capital costs, thus hindering access to health care technologies. The pricing of these applications is also uncertain as the technology is so novel and unique to each company. Application purchases also incur additional costs for maintenance and training. The severely limited information on costs for the maintenance and running of devices could lead to unforeseen expenses that strain the budgets and finances of institutions. A second risk lies in the probability of error in prediction and its potential effects. Slips in prediction that lead to wrong or late diagnosis and treatment would adversely impact health care. Physicians generally regard a false negative instance of a test as more severe than a false positive instance, since there is lethal potential in a missed diagnosis. Thereby, the integration of machine learning (ML) and artificial intelligence in health care requires an initial "trust" step in which the ML approach should convince the physicians and the health care system that it is "safe" to use. Training datasets are generally based on probabilistic distribution, on which a probability distribution function is fitted to Joseph's din investment models by the ML-based approach. The main consequence of this is the probability of error of diagnosis/prediction that creates doubt on faith in and acceptance of ML in health care. The impact of wrong prediction in health care is measured in years of life lost, money wasted, and, not least, potential human casualties. The grind of lethal ML-based oncology systems would steer numerous studies to scrutinize all aspects from selection bias to software specification. The implication of a false positive prediction in an ML-based screening for diabetes would trigger awareness of the potential of widespread testing down the line. A third risk factor concerns the availability of data, particularly high-quality training/testing data of large-enough size for reproducible results. It is so singleplistic and self-evident as to betray (at least) two analyst's blind spots when discussing the feasibility of largescale roll-out of machine learning in health care: box thinkers assume that requisite data takes forms and coding that is stable enough to be interchangeable.

6.1 Data Quality Issues

Data accuracy & data quality issues have attracted the attention of many scholars in different areas as the importance of how to achieve & maintain data quality increases. The emerging mobile health (mHealth) sector attracts a significant amount of investment and population attention. Mobile phones, with the potential of a good mobile technology and access to internet, can collect health information and deliver high quality information anytime even in remote areas. With this data, diagnostic services can be provided effectively & efficiently to find previously undiagnosed patients through a data driven approach, and diseases can be monitored automatically over time using patients' mobile phones, which can save patients' time and the burden on hospitals. Unfortunately, enabling this progress relies heavily on the assumption that the data collected on patients' health via mobile phones is accurate. In consideration of the huge progress that could be achieved, the consequences of inaccurate data on patient's health can be dire. Common medical errors are those found during investigation errors, diagnosis errors, treatment errors, communication errors, and office administrations errors. Process issues and data quality issues could be sources of those common errors. Constraining these errors would require learning about the data collected and application of analysis techniques to find sources of inaccurate data. Data quality issues need to be addressed as the choice of the mobile technology to collect data and the design of forms to be presented on patients' mobile phones are both properly executed. Machine learning (ML) algorithms are thus considered in multiple stages of the mHealth. They play a pivotal role in acquiring accurate data by deploying pre-trained algorithms in mHealth solutions. With respect to the collected data, ML algorithms could provide reasoning for detecting inaccurate data which could be of crucial importance in the quality of the data collection stage of the mHealth. Assurance of high data quality, either by removing inaccuracy or by prohibiting it from entering the data collection phase would yield mHealth solutions which with their lower costs than conventional health services could manage diseases better and reduce health care costs.

6.2 Integration with Existing Systems

Building healthy responsive predictive healthcare systems requires more than just smartphones and smart devices. Access to technology must be combined with access to patients' big data. To store, maintain, analyze, and utilize this big data in a timely manner, an enormous computational infrastructure is required. However, due to the fast evolving technologies and formats of patient data, a significant amount of data reside in the remote storage clouds. In this case, analyzing personalized patient data must be moved based on a thorough analysis to the maintained medical data cloud of the healthcare system of the patient. It is at this medical data cloud where in advance technique understanding and data are concentrated. This make it very promising for intelligently evaluating the data. However, the performance of big data analysts is decided by other data approaches involved in it, i.e., by data moving to a distant data analyst rather than using the existing levels of data approach or by opting for live period analysis algorithms so the amount of data can immediately be analyzed. In the case of a limited time window, near-data processing solutions will be desired.

Equation 2: System Integration Efficiency (SIE)

$$SIE = \frac{F_s \cdot C}{T \cdot R}$$

Where:

- F_s = Number of successfully integrated functions or modules
- C = Compatibility factor (range: 0 to 1, higher = more compatible)
- T = Time required for integration
- R = Resources used (e.g., cost, personnel, processing units)

The proposed architecture entails data challenge flows instead of the collection or initial storage of big data. This framework tackles all challenges kindly while supporting a broad mixture of static data and live data big data cloud services. It partially leverages the strengths of existing MDA solutions while introducing novel ideas to address the limitations of the existing solutions. The proposed scalable architecture include detailed measures of consideration which the suggested solutions need to observe. There is no other shared open-source healthcare big data repository has also been illustrated. The viability of the suggested MDA architecture is implemented in a cloud environment, deployed on real-world high disease diagnosis data, and tested with simulated patient data delays. The existing several TPC benchmark both for static and for continuous streams crashes have been employed to test the system response time while being scalable and with the presence of heavy bird-type of data bursts.

6.3 Ethical Considerations

With the increase in the volume of available data, it becomes ‘big data’. The data from multiple streaming sensors and the data imported from other service providers gets accumulated. In recent times, the availability and affordability of high internet bandwidth and stream processing make it happen. The previous streaming data is caused by the past statistics and is in the form of entries of pools containing lists of field records. As the usage of sensors increases the number of illegal health monitoring devices, such as smart health care watches, medical IOT devices, environment monitoring devices, and 2D/3D virtual reality devices, the monitoring data and data to train classifiers may generate from one or multiple illegal devices.

Using the them of modern health care, ML technologies help consumers to early detect health unlikely disease cases during daily working times and then hint them to see specialists. Using the proposed methods, ML models can be updated to the new data obtained from authorized service providers instead of authorized one on likely disease detection. Meanwhile, privacy of health monitoring data is preserved. The proposed approaches have been evaluated end-to-end system from three different aspects including privacy preservation from both sensors’ data and consumers’ identification, lossless model training on non-trusting cloud server in the proposed privacifying case, and finally consumer execration performance improvement on several pharmetric testing datasets. As the earlier health, consumers ideally want to improve the whole process of early health change detection.

Chronic diseases, such as hypertension, diabetes, cardio disorders etc., are the major causes of morbidity in developing world where resources for primary health care are scarce. Conducting a screening for hypertension to detect the beginning stage of the disease among public would be much more beneficial. ML techniques have recently emerged as a powerful predictive tool in healthcare domains. Several parsimonious analytical methods have been proposed for predictive modelling of early hypertension screening. As the increasing in the screening population number, screening completion of hypertension can assist resource allocation at diagnosis.By leveraging machine learning models, healthcare providers can identify high-risk individuals earlier, enabling timely interventions that could prevent the progression of hypertension. Moreover, the integration of real-time data from wearable devices could further enhance the accuracy of predictions, making it possible to monitor patients remotely and provide continuous care.

7. Case Studies of Successful Implementations

Identifying patients at risk of clinical deterioration is crucial for improving care quality and preventing adverse events in hospitals. Inspired by this, this work presents an intelligent autonomous early warning system that uses machine learning (ML) to generate interpretable alerts about patient outcomes. The solution, called Monitora, is a cloud-connected intelligent platform deployed in 27 Brazilian hospitals, which worked with different electronic health record (EHR) systems. Each hospital group deploys medical alerts powered by an AI-derived score, calculated using input derived from different EHRs [18]. It was found that the alerts decreased both 30-day mortality and length of hospital stay (LOS). A predefined score from the input data predicts outcomes in the following process. First, the score and threshold for generating alerts are determined, which involves balancing fast delivery of alerts and a low fall rate. From training to deployment, two models are trained according to patients' time histories. Breach of threshold determines alert presence, with generated alert times saved in a database. The retrieval process includes a user query that limits the time window. Sara retrieves alerts from the database for compute-intensive models to derive a report. The report and alerts are forwarded to hospital-specific channels, such as Telegram and WhatsApp. This framework provides a flexible and general approach for transforming ML into an operational tool while enabling a team to analyze a region or population of hospitals.

Equation 3: Implementation Success Index (ISI)

$$ISI = \frac{(I_b - I_a) + (E_b - E_a) + (Q_b - Q_a)}{3}$$

Where:

- I_b, I_a = Key performance **indicator before and after** implementation (e.g., revenue, uptime, diagnostic accuracy)
- E_b, E_a = **Efficiency** gain (e.g., time saved, reduced processing load)
- Q_b, Q_a = **Quality** of service or user satisfaction (e.g., survey score, error rate reduction)

Acute kidney injury (AKI) is a significant health concern affecting the kidneys. Early detection, classification, and risk assessment of AKI remain challenging issues, particularly in developing countries. This work presents a new, comprehensive framework for deep learning that deploys three deep learning models (DL) for the early detection of AKI patients in real-life data that considers the time factor for a 3-hour readmission prediction window. Considering the guidelines, the framework employs a combination of text-preprocessing methods, feature-embeddings, CNN architectures, and modified encoder-decoder. This robust framework achieved excellent results when deployed in the Brazilian Ministry of Health Group of Hospitals model. Future works involve evaluating new predictive models for more effective early detection of AKI patients.

7.1. Early Detection of Diabetes

Diabetes mellitus is a chronic ailment with high prevalence in today's world and is expected to affect 642 million individuals by the year 2040. Diabetes is inevitable in the modern world due to hectic lifestyle, junk food intake, lack of exercise, and hereditary reasons. Diagnosis of diabetes at an early stage can take the correct remedial measures so that the patient's life does not get affected to a very large extent. The advantages of analyzing data based on machine learning and data mining techniques are that all the survey forms filled by individuals can be studied and evaluated. Diabetes can be detected at an early stage, and remedial measures can be suggested.

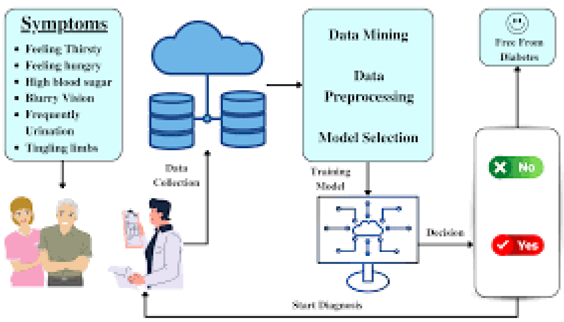

Figure 3: Block diagram representation of precision medicine employing various data analytics and machine learning algorithms.

Using precision and the number of true-positive diabetic readings by utilizing machine-learning algorithms Logistic Regression, Support Vector Machine, Naïve Bayes, K-Nearest Neighbors, and Decision Tree. Data pre-processing is done to eliminate duplicated and irrelevant entries. Data normalization is done to convert data to one unit interval. Necessary important modules are imported, and each algorithm is processed. The model is tested on divided train and test data, and parameters such as precision, accuracy, and score are displayed. The study concludes with which algorithm was best in predicting the condition of the data.

The aim of this work to observe and aim progress in the diagnosis of diabetes using newly invented methodologies in knowledge and intelligence realization which will help judicious decision-making and operational potholes. The urgency of diabetes monitoring and step-in processes has become even more urgent due to the advancements in computing and informatics in healthcare. This paper examines some representative early works as expert systems for diabetes diagnosis depicting practical applications of knowledge-based methodologies. The works are broadly categorized into risk evaluation systems aimed at estimating the likelihood of an individual developing diabetes given risk factors and appetite regression models for estimating appetite levels and subsequently blood glucose concentration given activity-related events and accompanying context information.

7.2. Heart Disease Monitoring

Coronary artery disease (CAD) is the leading cause of death in adults over the age of 35. The Morbidity data says that CAD kills 7.4 million people per annum; this number is so huge it requires a big data perspective. Heart disease detection systems based on heart valve movement simulation have been an active research topic for almost two decades. With recent advancements in medical imaging, statistical modeling, and machine learning, detection accuracy and system productivity have increased drastically. Accuracy is no longer an issue thanks to ensemble methods, which combine single models to exploit their strengths [3]. The multicentric research on heart disease detection with various medical images and datasets has opened up new avenues for detection. For example, echocardiograms are now routinely completed during patient visits. Model training on synthetic datasets has increased robustness against variants. Patient monitoring models have been developed to support long-term follow-up of heart disease progression under treatment. Collectively, these technologies can be combined to obtain a patient diagnostic and monitoring technique that is cost-effective and easy to apply. An Intelligent Synthesis Approach (ISA) offers the opportunity to create a novel heart disease detection and progressive monitoring technique. The thesis proposes a heart disease detection and progressive monitoring model based on a traditional data collection toolbox-based organized dashboard that records heart structure motions, outcomes, and synopses in textual logs. These data are processed with a domain-specific natural language processing model to extract and clean features for subsequent simulated and data-driven modeling. The former is fulfilled with a particle-filter-based Kalman modeling that simulates the extracted and augmented heart structure motion, while the latter uses a two-stage classifier ensemble consisting of missing value imputation, model selection, and output ensembling to detect heart disease from non-imaged numerical data. Collectively, the proposed design and a smartphone application serve as a toolkit to help doctors and clinics remotely diagnose patients and monitor their disease progression.

7.3. Cancer Screening Innovations

As of 2022, cancer remains a leading cause of morbidity and mortality worldwide, with approximately 20 million new cases and nearly 10 million deaths in 2020. Cancer often progresses insignificantly, which results in disease complication and late diagnosis due to non-specific symptoms. Therefore, early cancer detection is key to reducing morbidity and mortality . Currently, the most commonly used approaches for early detection include patient screening or 'triaging' through risk scores and medical history or imaging investigations of symptoms. Patient screening is used in asymptomatic individuals to track high-risk populations who have not received a diagnosis of cancer. Early screening can be performed by laboratory tests or imaging. For patients at-risk populations, rapid imaging investigation of symptoms can detect specific tumours faster than the traditional pathway.

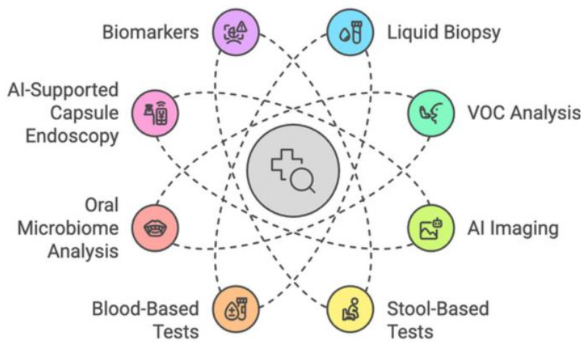

Figure 4: Advances in Non-Invasive Screening Methods for Gastrointestinal Cancers

Algorithms are trained with masses of past data to replace the healthcare professional's expertise. Algorithms for early cancer diagnosis can assist doctors fittingly, and thus the diagnosis is less dependent on physicians' diligent patterns. Historically, for ill-disciplined diseases such as cancers, there is a multitude of information shared publicly to foster knowledge dissemination, but it could also impede the patient when considering what is important and what is not. It is not only cost-effective but also self-love to make intelligent disease predictions models that can mitigate fear while helping patient monitoring. This elaboration demonstrates the robustness of the algorithm with voxel size preference, wrapper analysis of other radiomics features. The proposed Radiogenomic models can accurately predict the gene expressions from radiomics extracted from images across 15 cancer types that can help oncologists determine the best chemotherapeutics/intensive follow-ups prior to treatment initiation.

Computer-assisted image segmentation capabilities are demonstrated and to be broadly exploitable in other cancers, as the proposed image-first then pixel-next segmentation pipelines directly enumerate the advantage of incorporating a 3D model-direct voxel-based selection strategy. The incorporation of transformer encoders and the stacking of multi-header self-attention remarkable improvements in the tumour localisation task on breast, lung, and prostate cancers. This foundational model is extendable with extra modern computer vision components to serve the promise of open-source segmentation systems.Furthermore, the modular design of the architecture allows for seamless adaptation to different imaging modalities, including MRI, CT, and PET scans. This adaptability enhances the generalizability of the model across various clinical settings, supporting more accurate and efficient diagnostic workflows.

8. Future Trends in Machine Learning and Healthcare

Machine learning in healthcare has currently gained traction and visibility. Various forms of artificial intelligence are being utilized in healthcare and biotechnology to harness the growth of big data in these fields. The recent improvements and discoveries in machine learning, particularly in big data analysis and generation, have generated a great opportunity and solution to assist the biomedical field in the understanding and advancement of knowledge at an excellent resolution, thus improving the quality, accuracy, and prediction of care [15]. Physicians and data analysts alike have utilized current advancements in machine learning to aid in the understanding and prediction of large amounts of complex multi-parameter data, in the improvement of knowledge and curability of diseases, and even in the automation of repetitive yet time-consuming tasks. Currently, research and trading applications have been focused on deep learning, an aspect of machine learning more focused on the artificial neural network aspect of machine learning. Far superior outputs of adjectives versus inputted variables have been demonstrated with deep learning when compared to a broader machine learning technique that often contained bias due to the underlying statistical methods being analyzed.

Regardless of its potential benefits, machine learning has currently served only in the supportive or second-hand role of large physician or analyst organizations in healthcare. Anti-viral agents, anti-inflammatory compounds, and vaccines against the virus were highlighted as examples of machine learning implementation in the batched research and development stage. Current machine learning implementation in major medical organizations includes a type of supervised machine learning-based approach called the Cost Function Approach, which attempts to provide a solution to the common complications of large medical institutions, especially concerning the digitizing and organizing of their patients' electronic health records. As demonstrated in the aforementioned case example of COVID-19, research and attempts have been made to gather information regarding the virus's origin, genetic sequence, and prevalence in humans. Various approaches were utilized, including static and dynamic scrutiny of nucleotide and protein bond conservation across kin-bacteria cybernetics, investigation of high-frequency bacteria, and X-ray crystal structure analysis of protein.

8.1. Personalized Medicine

Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine Techniques and methods inspired by biologically based learning, also known as machine learning (ML), are receiving considerable attention in the medical and personal health fields. Social media data, IoT data, electronic patient records (EPR), and genomic screening data are now extensively gathering and sharing, resulting in big health data. In spite of the huge potential of this big health data, the rate of successful adoption of AI and ML in real world healthcare systems continues to lag behind others; hence, the global burden of diseases has only been reduced in a slow pace over past decades. There are several barriers that prevent systems from taking a big leap into the realm of population health tell it's cultural aspect, resources aspect, computation/IT aspect, and Modeling aspect. The Cassiopea project simplifies these barriers with a few mutual philosophy and techniques, the architecture, platform development, application & testing strategies and even internal closures with a capability to keep updating, enhancing, and adapting after it has been established.

The classical necessity of medical AI and ML is recognition and prediction of disease ailments/symptoms. It is necessary in achieving good patient compliance to minimize human errors in personal health monitoring activities; it can also give a reliable decision support for professionals. The ML endeavor in this domain is massive, with neck-breaking performance and model-space complexity. But the applicability of these state-of-the-art ML technologies are severely limited as there lacks mutual and secure platform for heterogeneous complexity, unknown features and resultant mutual comparison across domains. AI and ML technologies, and models in this domain are hence incompatible across various studies.

The technologies and methods of AI and ML in this domain tend to raise multiple barriers either in pre-processing, modeling, or post-processing, result interpretation and performance evaluation. Hence it is a mutual hope that a whole background system of adaptable, usable and transferable mutual/compatible architectures be developed for modeling in this domain with only very minimal modifications. Since such a development and establishment attempt has not been conducted by possible popular framework providers or private AI companies yet, this Cassiopea project would play a pioneering role in promoting the development towards better healthcare and precision medicine.

8.2 AI-Driven Decision Support Systems

These systems are computer software designed to aid clinicians in making evidence-based diagnoses, therapeutic, and operational decisions. They integrate patient information, clinical guidelines, and the latest medical research to generate recommendations. These systems leverage advanced computing and ML techniques to analyze diverse types of complex medical data, including electronic health records, images, and text, to support clinical decision-making. This systematic review provides a comprehensive overview of the state and challenges of AI-CDSS in dermatology, radiology, cardiology, and overall healthcare delivery. Moreover, both implementation and regulatory challenges, with unique attention to the European context, are discussed, along with potential solutions.

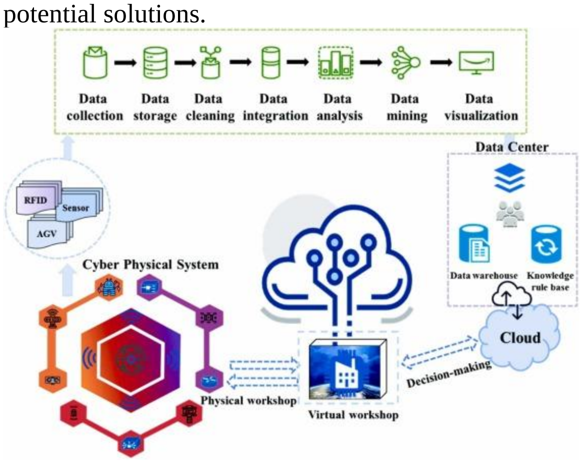

Figure 5: AI-Driven Decision Support Systems

AI-driven decision support systems are means of augmenting clinical reasoning and actions by taking into account patient data. In the last decades, AI-supported clinical decision support (AI-CDSS) systems have gained traction in healthcare. Such systems use AI to process a wide variety of patient data, including patient histories, laboratory test results, or physiological parameters, as well as clinical guidelines or protocols and scientific evidence. These systems support clinicians' decisions, offering early alerts, possible diagnoses, raising suspicion for diseases, alternative therapies, or justifying treatment choices. In addition, they can undertake actions, such as requesting additional tests or conducting medical imaging. The decisions and actions of existing AI-CDSS systems may hold professional, legal, and financial responsibility. Thus, clinicians remain in control, assisted by advice from the systems.

Health data is complex, heterogeneous, and often incomplete. This poses great challenges for mitigating the biases that underlie AI predictions, particularly in low-resource settings where enhanced clinician focus is needed the most and overworked health systems exacerbate biases. Achieving fair AI-CDSS solutions will require augmented and pooled data repositories. Specialized standards and practices must also be developed for certain applications, such as dermatology, where patient data is more easily annotable and tractable. Nevertheless, increased and diversified data may introduce a new level of bias, especially for less frequent diseases, warranting exploration of federated learning systems for algorithm training.

9. Conclusion

Machine learning (ML) has been employed in numerous healthcare applications to exploit the abundance of healthcare-related information. Wearable devices can offer various types of vital signs (e.g., body temperature, blood pressure), and their advancement enables real-time patient's vital sign measurements. These vital signs can be correlated with several diseases that affect vital signs behavior. In addition, a vital sign increase or decrease could indicate an acute situation for the patient. Patient monitoring is crucial in healthcare applications as it can impact doughiness during a health state distance estimation by the monitoring specialists. As each patient's vital sign information can have several measurements, i.e., frequency and/or unit; it can be challenging for the healthcare specialists to continuously monitor several patients. Also, the healthcare specialists perform various measurements for each patient and for a specific healthcare situation in order to optimize the captured information. These challenges motivate the development of a big data analytics platform for real health data exploration and to assist healthcare specialists in real-time patient monitoring and monitoring state estimation in timely intervals.

The proposed platform should work on a Hadoop ecosystem and have the capability of real-time patient monitoring and vital signs supervision for the detecting of their abrupt changes with the sampling of each patient's vital sign every 15 minutes at minimum. The maintenance of the monitoring and Hz estimation of each patient's vital sign can be conditioned of a healthcare situation that can control the maximum time intervals between the monitoring motions based on the timestamp of the last assessment request sending. Thus, the proposed platform architecture consists of four layers: real-time patient monitoring, real-time decision, and data storage, patient classification, and disease diagnosis & data retrieval and visualization.

References

[1] J. Fritzsch, J. Bogner, S. Wagner, and A. Zimmermann, "Microservices Migration in Industry: Intentions, Strategies, and Challenges," 2019.
[2] S. Talja, "Service-Oriented Process Models in Telecommunication Business," 2013.
[3] Routhu, K., Bodepudi, V., Jha, K. M., & Chinta, P. C. R. (2020). A Deep Learning Architectures for Enhancing Cyber Security Protocols in Big Data Integrated ERP Systems. Available at SSRN 5102662.
[4] K. Maroukian and S. R. Gulliver, "Leading DevOps Practice and Principle Adoption," 2020.
[5] L. Silva, M. Unterkalmsteiner, and K. Wnuk, "Monitoring and Maintenance of Telecommunication Systems: Challenges and Research Perspectives," 2020.