A Comprehensive Introduction to Healthcare Data Analytics

Maria Rosa Calvino de Gomez
National University of Tucuman, T4000 San Miguel de Tucumán, Tucumán, Argentina.
gomezrosa@hotmail.com

Correspondence should be addressed to Maria Rosa Calvino de Gomez : gomezrosa@hotmail.com.

Article Info
Journal of Biomedical and Sustainable Healthcare Applications (http://anapub.co.ke/journals/jbsha/jbsha.html)
Doi: https://doi.org/10.53759/0088/JBSHA20240405
Received 16 September 2022; Revised from 18 April 2023; Accepted 30 June 2023.
Available online 05 January 2024.

Published by AnaPub Publications

Abstract – Healthcare data can be collected from various sources, including sensors, and conventional electronic records, photographs, data from clinical notes/biological literature, among others. The variation in data representation and gathering gives rise to issues in both data interpretation and processing. The methodologies required to analyze these diverse sources of data exhibit considerable variation. The presence of heterogeneity within the data gives rise to a distinct set of challenges when it comes to the processes of integration and analysis. This article presents a detailed review of healthcare data analytics and the respective data sources. Secondly, it discusses advanced data analytics for the healthcare sector, and its practical systems as well as applications of healthcare data analytics.

Keywords – Big Data, Healthcare Data Analytics, Data Gathering, Data Representation, Data Processing, Data Interpretation.

I. INTRODUCTION

Big Data refers to information assets characterized by their substantial volume, velocity, and diversity, which necessitate the utilization of specialized technology and methodologies to extract meaningful value from them. Big Data encompasses a vast amount of information characterized by its substantial volume, dynamic nature, and diverse variety, which in turn requires innovative methods of processing. In order to effectively handle the volume, velocity, and diversity of Big Data, the utilization of new technologies is imperative. This is due to the limitations of conventional data-processing systems and software tools, which are unable to adequately manage the aforementioned aspects of Big Data. The distinctiveness of Big Data is undeniable when compared to conventional data warehouses and other archival systems. Organizations necessitate a novel approach in addressing unstructured data. Initially, it is imperative for businesses to embrace the notion of streaming analytics, necessitating a shift in perspective to perceive data as dynamic streams rather than static entities. The aforementioned attributes require the utilization of state-of-the-art information technology (IT) systems that effectively leverage emerging data. The emergence of the "Big Data" concept is closely linked to the rapid expansion of available data, thereby facilitating valuable analyses, drawing conclusions, and enabling more informed decision-making for various organizations and individuals.

The data set is contributed to by patients, institutions, as well as a diverse array of monitoring devices, sensors, and equipment. The healthcare sector employs a combination of paper-based and electronic systems for data storage. The nature and specific characteristics of big data analytics pose technical and organizational challenges for businesses. The healthcare sector has consistently generated substantial amounts of data due to various factors, including the necessity of maintaining comprehensive records of individuals' medical histories. However, the challenge associated with Big Data in the healthcare industry is not solely the immense amount of data, but also the unparalleled diversity of data types and formats, as well as the need for timely evaluation in order to provide vital information on an ongoing basis. One of the primary challenges associated with Big Data pertains to the ability to make well-informed judgments across numerous domains, given the vast quantities of available data. The adaptation of extensive data storage, analysis, presentation of analysis findings, and drawing

inferences from them for application in a therapeutic setting poses a notable challenge within the realm of health-care data. The primary objective of healthcare analytics systems is to facilitate the comprehension of complex data through accurate description, integration, and presentation (see **Fig 1**). The utilization of this would enhance the processes of acquiring, storing, analyzing, and visualizing large volumes of healthcare data.

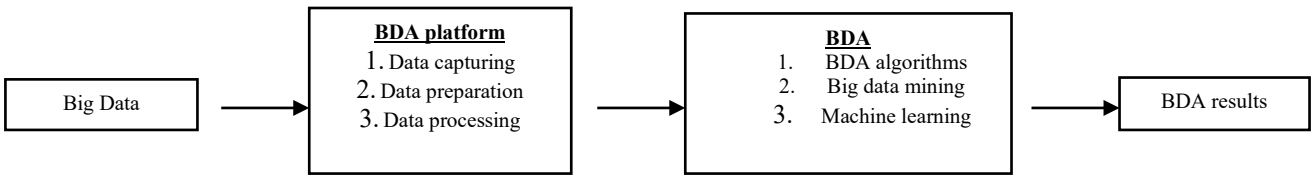

Fig 1. Big data analytics process

The development of sophisticated analytical approaches, which efficiently convert information into actionable and meaningful insights is crucial for comprehending and advancing knowledge derived from healthcare data. Advancements in general computer technology have already commenced to significantly transform the accessibility of medical treatment for patients. Data analytics plays a particularly significant role in these computer systems. When implemented in the context of healthcare data, analytic solutions exhibit significant potential in transforming healthcare delivery from a reactive approach to a proactive one. In the forthcoming years, the healthcare industry is poised to experience a heightened impact from analytics. Analysis is typically employed to reveal the inherent patterns present in health data. Furthermore, this will assist healthcare professionals in developing a distinct profile for every patient, thereby enabling more accurate assessments of an individual's prospective likelihood of encountering a medical condition.

Various types of healthcare data sources include biomedical imagery, clinical text, electronic health records, biomedical signals, genetic data, social media data and sensing data. The utilization of genomic data analysis may enhance the comprehension of the interrelationships among mutations, disease states, and genetic markers. There exist several significant challenges that must be addressed in order to effectively implement genomic discoveries in the realm of personalized therapy. The utilization of clinical text mining can facilitate the conversion of unstructured clinical notes into actionable intelligence. Natural language processing (NLP) and Information retrieval techniques can effectively extract valuable insights from extensive clinical writing. In order to analyze and forecast global health patterns, such as the occurrence of infectious epidemics, scholars are progressively utilizing social network analysis (SNA). This method relies on data derived from various social media platforms, including but not limited to Twitter, web logs, Facebook, and search engines.

Table 1. Disease diagnosis scheme within a system

Diseases	Diagnostic approach	Healthcare approach using IoT
Respiration index	Pattern-matching, frequency matching	Respiration sensors
Stress index	Patter-matching based, frequency based	Emotiv EPOC sensors, other sensors for stress measurement
Water borne diseases or infections	Scale based, frequency based	Camera pill, temperature sensors, ECG
Heart diseases	Pattern matching, frequency matching	ECG patterns
Obesity	Scale based	Blood pressure, body weight
Hypertension	Scale based, frequency based	Blood pressure

Table 2. Individual health data datafication layers

Layers	Descriptions
1	Web searches, smartphones, purchases, EHR, medical service, IoTs sensors, fitness and medical devices
2	Medical and other information related to healthcare in identified types that are gathered, stored, and disseminated to a third party
3	Majorly private entities gather raw data from the first layer, second layer, and other public and private sources. The results are disseminated either in identified or de-identifiale form
4	Internation public or private companies re-process, re-disseminate, or re-sells the data for different aims
5	International treaties and agreements controlling the safeguarding of privacy for individual data about health

The evaluation of the severity of disorders is contingent upon the utilization of appropriate diagnostic procedures. **Table 1** presents a diagnostic approach as demonstrated by Duman and Tolan [1]. Hussain, Siersbæk, and Østerdal [2] present a comprehensive breakdown of the five distinct categories comprising an individual's health data in **Table 2**. Despite the fact that regulations have not kept pace with technological advancements, it remains crucial to safeguard the fundamental rights of individuals whose data is being processed and their right to privacy.

This paper presents a critical review of healthcare data analytics. The following sections have been arranged as follows: Section II focusses on the healthcare basis analytics and data sources. In Section III, a dicussion of advanced data analytics

for healthcare has been provided. Section IV presents a review of the practical systems and applications of healthcare data analytics. Lastly, Section V draws final remarks to the paper.

II. HEALTHCARE BASIC ANALYTICS AND DATA SOURCES

This section will examine the different sources of data and their effects on analytical approaches. The utilization of a broad array of diverse sources in medical data mining necessitates the implementation of a diverse set of data analytics methods.

Electronic Health Records

Electronic health records (EHRs) are utilized to document the medical history of a patient. The comprehensive medical record encompasses various components such as the patient's demographic data, presenting concerns, physician's observations, prescribed medications, past medical history, vital signs, laboratory findings, radiographic assessments, progress updates, and pertinent financial details. Certain EHRs offer data that extends beyond an individual's medical or therapeutic background. The utilization of EHRs offers several benefits due to their ability to facilitate streamlined and proficient communication among healthcare professionals and organizations. EHRs are specifically designed to facilitate immediate access and modification by authorized personnel within this particular context. This exhibits substantial potential for practical application.

In certain cases, a secondary care facility or professional could necessitate accessibility to clinical records of patients' care physician. **Table 3** presents an overview of the diverse contexts in which EHRs can be utilized. The utilization of an Electronic Health Record (EHR) enhances operational effectiveness as it enables personnel to promptly retrieve the most current patient data. EHRs have the potential to offer a comprehensive documentation of a patient's clinical interactions, thereby assisting in various care-based activities like quality monitoring, evidence-based decision support, and outcome reporting. EHRs enable the efficient storage and retrieval of health-related information. According to [3], the utilization of this technology contributes to the improvement of diagnostic accuracy, the promptness of health outcomes, the facilitation of care coordination, and the enhancement of patient satisfaction.

Table 3. Overview of the diverse contexts in which EHRs can be utilized

Administrative Applications	In order to operate effectively, all EHRs necessitate an administrative backend. The administrative application of the Electronic Health Records (EHR) includes patient registration. During the registration process, various details pertaining to the patient's identity are gathered, including but not limited to their name, gender, age, contact number, place of residence, email address, insurance coverage, employment details, and the specific nature of their complaint.
Computerized Physician Order Entry	Computerized physician order entry (CPOE) is an essential software component within EHRs. CPOE is a software application utilized by medical practitioners to submit requests for various medical procedures, such as laboratory examinations, medication prescriptions, and diagnostic imaging. The utilization of CPOE enables healthcare professionals to electronically submit test requests, as opposed to the traditional method of manually transcribing them onto paper. This technological advancement offers numerous advantages for medical practitioners.
Laboratory Systems	Laboratory information systems (LIS) are extensively utilized in the healthcare industry at present, frequently integrated with the EHR to streamline the exchange of patient information and laboratory test outcomes. An interface is utilized to connect the majority of laboratory testing and analysis equipment to the Laboratory Information System (LIS).
Radiology Systems	The Electronic Health Record (EHR) facilitates communication with information systems of various departments, including the Radiology Information System (RIS). Radiology information systems, akin to laboratory information systems, serve as repositories for patient data encompassing radiology orders, test outcomes, appointment notifications, and image surveillance. The utilization of Picture Archiving and Communications Systems (PACS) is commonly observed in conjunction with radiology information systems.
Pharmacy Systems	Contemporary hospital pharmacies employ robots for prescription writing and utilize computerized medication carts. Another type of technology that aligns with EHR is a standalone pharmacy management system. Hospital pharmacies utilize bar codes to ensure accurate medication administration to patients in a timely manner. The integration of pharmacy systems with the EHR is imperative due to the crucial significance of monitoring medication interactions and drug allergies.
Clinical Documentation	EHRs primarily consist of clinical documentation, as healthcare professionals such as doctors, nurses, and other medical staff extensively record substantial quantities of data pertaiing to each patient they attend to. Electronic health records encompass a comprehensive range of medical information pertaining to patients, encompassing their prescription history, vital signs, and laboratory findings.

Biomedical Image Analysis

Medical imaging technology is instrumental in facilitating the production of high-quality photographs depicting various human anatomical features, thereby assuming a crucial role in modern healthcare. The analysis of such images can be of great value to medical professionals and researchers, as it aids in the monitoring of illnesses, planning of treatments, and determination of prognosis. Ultrasound (U/S), computed tomography (CT), Magnetic resonance imaging (MRI), and positron emission tomography (PET) are widely employed imaging modalities for capturing biological images. The capacity to non-invasively observe internal human organs holds significant implications for the field of medicine. These technological advancements facilitate physicians in acquiring a deeper understanding of a patient's health status without the necessity of invasive procedures.

Nevertheless, the mere visual observation of these organs merely marks the process' initial stage. The primary objective of biomedical image evaluation is to produce quantitative data and derive meaningful conclusions from images, thereby providing enhanced understanding of medical conditions. Given the inherent significance of this particular field of inquiry in facilitating a comprehensive understanding of biological systems and the development of efficacious remedies for health-related concerns, it is evident that its ramifications extend deeply into the social fabric. This poses several challenges, primarily stemming from the varied and intricate nature of the images, as well as the existence of irregular shapes and noisy data points. The examination of photos presents several research challenges, including but not limited to object identification, image registration, feature extraction, and picture segmentation Once these issues are solved, healthcare data analytics will possess the necessary tools to deliver significant analytical evaluations.

Sensor Data Analysis

Sensor data is extensively used in the clinical domain for both retrospective research and real-time monitoring. Sensors play a crucial role in various clinical data collection tools, such as electroencephalogram (EEG) and electrocardiogram (ECG). These data collection tools are commonly employed in real-time applications, although they can also be utilized for retrospective analysis. The utilization of critical care units (CCUs) and remote monitoring systems for patients with specific medical conditions presents a favorable prospect for conducting real-time analysis. All of the aforementioned scenarios entail the processing of potentially substantial volumes of data. In an intensive care unit (ICU), a multitude of data streams have the potential to be inputted into the sensor, resulting in the necessity for immediate activation of alarms.

In order to effectively implement such applications, it is necessary to utilize big data frameworks and employ specific hardware platforms. The study of long-term evaluation and real-time occurrences concerning different treatment options and patterns holds significant importance in the field of remote monitoring applications. The utilization of sensor data in healthcare holds significant promise, yet it also introduces a novel apprehension regarding the potential inundation of data. Therefore, it is imperative to develop state-of-the-art data analytic tools that possess the ability to transform these extensive datasets into practical and useful insights. Furthermore, the utilization of these analytical methodologies not only enhances the capacity to monitor patients' physiological metrics and boost situational awareness at the bedside of patients, but also provides valuable insight into inadequacies within the healthcare model, which potentially contribute to the escalating costs.

Biomedical Signal Analysis

The signals analyzed in biomedical signal analysis originate from physiological processes. Electroneurograms, electrocardiograms, electromyograms, electrogastrograms, electroencephalograms, phonocardiograms, and other similar signals exemplify the aforementioned types of physiological recordings. The comprehension of these signals is of utmost importance in order to accurately diagnose and determine the most optimal treatment plan. The quantification or relative assessment of human health can be achieved through the measurement of physiological signals. The acquisition of these readings can be achieved through either invasive or non-invasive means, utilizing a diverse range of sensors and transducers. The nature of the treatment or the level of severity of a medical condition can determine whether these signals exhibit a discrete or continuous nature.

The challenges in processing and interpreting physiological information arise from a combination of factors, including a low SNR (signal-to-noise ratio) and the interdependence of physiological models. In specific instances, a substantial amount of preprocessing may be required to effectively cleanse the signal data obtained from the associated medical devices. The advancement of various signal processing methods has significantly enhanced our comprehension of physical processes. Filtering, noise cancellation, and compact methodologies are among the numerous applications of the aforementioned. The literature has extensively examined advanced methods of analysis. Several techniques can be employed to minimize data dimensionality, such as singular value decomposition (SVD) [4], principal component analysis (PCA) [5], and wavelet transformation [6].

Genomic Data Analysis

Numerous diseases exhibit a hereditary element; however, the precise association between specific genetic markers and diseases remains uncertain. Although it is widely acknowledged that there is a familial component to diabetes, the complete repertoire of genetic markers that confer susceptibility to this condition remains elusive to scientists. The identification of genes associated with specific types of hereditary blindness, such as Stargardt disease, has been accomplished by researchers. However, it is important to note that not all potential mutations responsible for these conditions have been isolated at present.

It is evident that a more comprehensive comprehension of the interrelationships among mutations, disease states, and genetic markers will significantly contribute to the development of efficacious gene therapies. Individuals with a vested interest in data-driven research often seek to ascertain the specific health issues that can potentially be elucidated through the computational analysis of genetic data. There exist numerous challenges that must be addressed prior to the practical implementation of genetic discoveries in personalized medical interventions. The genomic landscapes of diseases such as cancer, which have an impact on multiple organs, are widely recognized for their intricate nature and demonstrate substantial levels of individual variability. The resolution of these problems would constitute a significant advancement in the realization of personalized medicine.

Recent advancements in biotechnologies have led to a significant increase in biological and medical data, as well as notable progress in genomic research. Furthermore, this development has expanded the realm of possibilities and heightened the anticipations surrounding the examination of intricate phenomena in the life science domain, particularly at a genomic level. Recent developments in the genomic technology have facilitated the comprehensive examination of the genetic composition of individuals without any apparent health issues, particularly in relation to complex disorders [7]. Several research avenues have yielded promising findings, indicating their potential to contribute to novel insights into human disease biology and the capacity to forecast individual responses to therapy. In addition, genetic information is commonly depicted in the form of sequences or networks. This implies that individuals working in the field must possess a high level of proficiency in sequence and network mining techniques. The identification of disease biomarkers and therapeutic targets, along with the ability to predict clinical outcomes, are currently prominent areas of focus in medical research. Numerous data analytics-driven solutions are currently being developed to address these pressing concerns.

Clinical Text Mining

Clinical notes, also known as patients' records, serve as the primary method for encoding information pertaining to individuals. The predominant source of healthcare information relies on these records, frequently stored in an unorganized data structure. This section encompasses clinical data obtained through dictation transcription, direct input from healthcare providers, or the utilization of voice recognition software. These resources exhibit significant potential as untapped sources of information. The process of manually encoding a free-text form containing diverse clinical information, even when limited to secondary and primary treatments and diagnoses for the purpose of billing, is evidently cost-prohibitive and time-consuming. The analysis of free-text clinical language is a challenging task due to the complexity involved in converting it into a structured format. As a result, the mechanical analysis of such notes is widely recognized for its difficulty.

The complexity arises due to the unique circumstances of each patient and doctor, as well as the unstructured, heterogeneous, diversified, and multifaceted nature of the data. The timely mechanical encoding of clinical information is significantly facilitated by the utilization of natural language processing (NLP) [8] and entity extraction [9]. In these situations, data preparation approaches hold greater significance compared to actual mining techniques. The utilization of concise and fragmented sentences, dictation methods, condensed lexicons like acronyms and abbreviations, and frequently misspelled technical vocabulary pose challenges in the interpretation of clinical content through natural language processing methods, rendering it more complex compared to the analysis of other types of texts. Due to the aforementioned challenges, the task of clinical text processing is considerably arduous, as these issues significantly impact various fundamental natural language processing (NLP) tasks, including but not limited to shallow or complete parsing, sentence segmentation, and text classification.

III. ADVANCED DATA ANALYTICS FOR HEALTHCARE

This article will examine various advanced methodologies for data analytics in the healthcare sector. Within the realm of medical applications, there exist different machine learning and data mining models, which necessitate customization to suit the specific requirements of this domain.

Clinical Prediction Models

Clinical prediction models have become increasingly important in modern clinical care as they aim to enhance health outcomes and facilitate shared medical decision-making. These models educate healthcare providers, patients, and their families about the risks associated with different outcomes. The objective of diagnostic prediction models is to ascertain the probability of the existence of a disease in a patient at present, while prognostic prediction models aim to determine the likelihood of specific health conditions occurring in the future. The utilization of clinical prediction plays a crucial role within the contemporary medical system. Considerable research efforts have been dedicated to the development and implementation of various prediction models, which have proven to be highly effective in the realm of clinical practice. These types of models have had a significant impact on the identification and treatment of illnesses.

There exist three distinct categories of supervised learning approaches that have demonstrated efficacy in clinical prediction tasks. The first category encompasses conventional statistical approaches such as logistic regression and linear regression. The second category encompasses more sophisticated techniques derived from data mining and machine learning, such as artificial neural networks and decision trees. Lastly, the third category encompasses survival models that purposes to forecast survival results. The main focus of these methods is on covariate variables, also identified as features and traits, including dependent outcome variables. The identification of anticipated outcomes plays a crucial role in determining the most suitable model to be employed in a given healthcare scenario. To address the extensive array of

potential outcomes, various prediction models have been proposed. The binary and continuous formats are commonly observed in the presentation of results. Categorical and ordinal outcomes represent two additional, less commonly observed types. Moreover, a variety of models are available for addressing survival outcomes in situations where the objective is to predict the timing of a specific event.

Temporal Data Mining

The consideration of the temporal component is essential when analyzing or extracting healthcare data, as it is a pervasive feature across the majority of healthcare datasets. In the healthcare sector, two primary sources contribute to the generation of temporal data. One source of information is obtained through the use of sensors, whereas another source is derived from EHRs. The potential for acquiring a more philosophical comprehension of progression, manifestation, and treatment response of illnesses through the analysis of the temporal aspect of electronic health record (EHR) data is considerable. The conventional methodologies prove inadequate in handling electronic health record (EHR) data due to its inherent characteristics of heterogeneity, sparsity, high dimensionality, and unpredictable temporal intervals. In contrast to electronic health record data, sensor data is commonly depicted as numerical time series that are regularly recorded at consistent intervals.

Physiological data obtained through regular patient monitoring serves as an illustrative instance of this category of information, alongside recordings of the patient's electrical activity, including electrocardiograms (ECGs), electroencephalograms (EEGs), and so forth. In contrast to longitudinal electronic health record (EHR) data, which is commonly collected throughout a patient's lifespan, sensor data is captured over a comparatively shorter duration, typically ranging from a few minutes to several days. Distinct approaches are required for temporal data mining of EHR data and sensor data due to their unique characteristics. One of the commonly employed methodologies for extracting valuable insights from EHR data involves the utilization of temporal pattern mining techniques. These techniques involve the representation of data cases, like patient records, as arrangements of discrete occurrences, like diagnostic procedures and codes. The primary objective is to identify and enumerate statistically significant patterns within the data. Time-series analysis and signal processing approaches, such as independent component analysis and wavelet transform, are frequently employed in order to interpret and extract meaningful information from sensor data.

Visual Analytics

Enhanced understanding of diseases and the capacity to determine disruptive clinical workflow patterns necessitate the capability to assess and determine pertinent patterns in multimodal medical data. The integration of data analytics, interactive interfaces, and human cognition collectively renders visual analytics a potent instrument for effectively navigating vast and intricate information. The primary objective of visual analytics is to develop solutions that facilitate individuals in comprehending and interpreting complex data by employing interactive visual interfaces and analytical techniques. Visual analytics has proven to be effective in various domains of healthcare data analysis, primarily due to its ability to provide a wide range of valuable insights. The increasing abundance of health-related data necessitates the development of efficient approaches for data analysis that leverage human-computer interaction and graphical user interfaces. Presenting complex healthcare data in a manner that is easily understandable to individuals is beneficial for facilitating the emergence of novel perspectives.

Numerous diseases are assessed by employing datasets that encompass a multitude of clinical attributes, often numbering in the hundreds or even thousands. The synthesis of information and extraction of insights from medical data have significant effects for users due to the complex nature of the data, which is characterized by its multimodal, noisy, diverse, and temporal properties. The considerable magnitude of data being produced by healthcare institutions offers promising prospects for the advancement of interactive interfaces that facilitate the exploration of extensive databases, the verification of clinical data and coding techniques, and the promotion of internal transparency within departments, facilities, and organizations. Numerous visual methodologies have been derived from the data mining literature [10], alongside additional techniques specifically devised for the healthcare sector.

Clinico–Genomic Data Integration

The identification and understanding of human illnesses pose significant challenges due to the complex interactions among various factors encompassing genetics, clinical manifestations, behaviors, and environmental influences. The capture of various impacts resulting from multiple parameters is achieved through the supplementation of clinico-pathological and genomic databases. To effectively integrate the essential information contained in both genetic and clinical data, it is imperative to develop comprehensive models that consider both factors simultaneously. These types of models have the potential to contribute to the advancement of improved diagnoses, therapies, and medications, thereby facilitating the progression towards personalized medicine.

The integration of clinico-genomic data is an emerging discipline that leverages the potential afforded by the amalgamation of clinical and genetic information to develop integrated predictive models. Genomic data pertains to the comprehensive set of information regarding an individual's genome, encompassing various aspects such as single nucleotide polymorphisms (SNPs), protein profiles, metabolite profiles, and gene expression patterns. On the other hand, clinical data encompasses a broad spectrum of information pertaining to the patient's pathology, behavior, demographics, familial background, environmental factors, and medication history. The ultimate objective of an integrative study is to identify the

genomic and clinical factors linked with a particular illness phenotype, like the presence or absence of cancer, the distinction between tumor and normal tissue samples, or the duration of survival following a certain treatment.

Information Retrieval

Over the past few decades, there has been a vital increment in research and interest surrounding the utilization of information retrieval (IR) methods within the domain of health and biomedicine. This growth can be attributed to similar factors that have contributed to the rapid advancement of the broader field of IR in recent times. The Internet and related technologies have significantly enhanced the accessibility of health resources, resulting in substantial benefits for both health professionals and patients alike. The utilization of applications that offer expedient access to health information that is verified and regularly updated has the capacity to impact the caliber of care dispensed by healthcare practitioners. Consequently, these applications hold significant importance in the dissemination of knowledge. The advent of the World Wide Web has significantly transformed the dynamics between patients, their loved ones, friends, and medical professionals, granting them enhanced accessibility to health-related information.

According to [11], a robust healthcare system should facilitate the discovery of pertinent information, provide personalized health support services, and enable individuals to engage in mutual teaching and learning with others facing similar circumstances. PubMed, an information retrieval (IR) application created by the American NLM (National Library of Medicine), is a longstanding and extensively utilized platform that grants users access to a vast collection of medical research literature from around the globe. Consumers have access to a diverse range of services through which they can obtain health information, with the caveat that the accuracy and reliability of the data provided by these services may vary. Numerous governments worldwide have initiated initiatives aimed at enhancing consumer health by expanding consumer access to health information and streamlining the exchange of patient data. Prominent entities such as Google and Microsoft, which hold substantial influence within the information retrieval (IR) sector, have initiated investments in this domain.

Google Health and Microsoft HealthVault were introduced by Google and Microsoft, respectively, in May 2008 and October 2007, correspondingly. A program called Revolution Health has been recently launched by the firm owned by Steve Case, the creator of AOL. The overarching concept underlying the emergence of "consumer-driven healthcare" is to empower individuals with increased autonomy in making decisions pertaining to their health. Enhancing patients' health and well-being can be achieved through various means, such as facilitating the retrieval of their health records across multiple devices, streamlining access to pertinent data, and utilizing digital resources to offer personalized recommendations. A significant proportion of individuals presently employ the Internet as a means to investigate matters pertaining to health. According to a study conducted in 2006, approximately 80% of web users in America use the web as a major source of clinical data.

Additionally, it was discovered that 53% of individuals seeking health-related information indicated that their most recent health information session had influenced their self-care practices or their caregiving for others. Furthermore, it was observed that health searches are equally prevalent as activities such as reading blogs or utilizing the Internet to retrieve contact details. The [12] anticipate several implications for the future of healthcare resulting from the emergence of the information society. One of their theses posits that patients and their families will possess knowledge regarding the information resources accessible through the Internet and will actively utilize them. In [13] anticipate that electronic media will provide current and extensive information on diseases, which will be easily accessible to individuals and their families. This accessibility will extend to patients and their families, commonly referred to as "consumers."

Additionally, [14] predict the emergence of novel services in this domain, with an estimated increase in access to health websites exceeding 30%. Regardless of its quality, this information will be accessible to users without any cost. This implies that the content produced by individuals or organizations has the potential to obtain certification, such as from industry associations, adhering to a universally recognized set of criteria. The implementation of knowledge assistance within clinical procedures is being strategically planned. It is anticipated that by the year 2020, a significant proportion of clinical practice recommendations, approximately 80%, as well as medical information obtained through surveys, will be readily available through online platforms.

Privacy-Preserving Data Publishing

The concept of privacy in the healthcare sector is commonly defined as the entitlement and inclination of an individual to manage the dissemination of their personal health information. Health records are regarded as highly confidential due to the inclusion of personal information pertaining to patients. Various types of data, such as information related to diseases and genome sequences, may require specific safeguards due to a multitude of factors. The ability of medical organizations to share their data with statistical professionals plays a pivotal role in facilitating medical research. The potential for significant financial savings can be observed through the collective pooling of individuals' medical records. This gives rise to valid concerns regarding the infringement of personal privacy. The challenge of data privacy poses a significant barrier to the advancement of healthcare data analytics. The majority of privacy protection methods hinder the capacity to faithfully depict data, thereby increasing the challenge of concealing personal information.

This can be achieved by introducing random perturbations into the sensitive attribute, or by introducing random perturbations into the other attributes that serve as identifying factors, or by employing both methods simultaneously. It is evident that during the course of this procedure, a certain degree of precision in the representation of data needs to be compromised. Consequently, the preservation of individuals' privacy often necessitates sacrificing the utility of certain

information. Hence, privacy-preserving techniques endeavor to achieve a harmonious equilibrium between the contrasting principles of convenience and confidentiality. This ensures that, for any given level of privacy, the minimum amount of utility is compromised. The primary steps in privacy-preserving data publication algorithms involve selecting an effective privacy level and metric based on the data characteristics and access setting, applying a single or multiple privacy-preserving approaches to effectively accomplish the required primary level, and assessing the employment of processed dataset. The processes are repeated until the required levels of convenience and anonymity are reached.

IV. PRACTICAL SYSTEMS AND APPLICATIONS OF HEALTHCARE

The last section of this article will center on practical implementations and frameworks within the healthcare domain that heavily rely on the utilization of data analytics. These issues have undergone significant development in recent years and continue to garner considerable attention and generate extensive discussion. While not directly related to medical diagnostics, techniques such as fraud detection have notable implications within this field.

Data Analytics for Pervasive Health

Pervasive health is systematic monitoring of an individual's health in real-time and the provision of continuous medical treatment through the utilization of advanced technologies such as wearable sensors. The efficacy of various treatment modalities can be assessed over an extended period by employing wearable monitoring technology. Challenges associated with these methodologies encompass the ability to perform real-time processing and effectively extract valuable insights from vast datasets. In recent years, the practicality of such systems has been facilitated by advancements in technology and software, specifically in the field of data analytics. The aforementioned advancements have facilitated the integration of affordable intelligent health systems into household and residential environments. Intelligent healthcare systems have the potential to utilize a diverse array of sensor modalities, including wearable and ambient sensors.

Wearable sensors encompass electronic devices that can be conveniently worn on the human body or seamlessly integrated into various types of fabric. Distributed three-axis accelerometers have the capability to offer comprehensive information regarding an individual's posture and movement at any given moment. Concurrently with these developments in sensing modalities, there has been an increase in the utilization of analytical techniques on the data generated by these instruments. The integration of analytical solutions is being widely adopted in various healthcare systems in practical settings. Several examples of systems in the field of cognitive health monitoring include activity-recognition-based systems, persuasion-based systems aimed at encouraging individuals to effectively transform their health practices, and fault-detection-oriented models.

Healthcare Fraud Detection

Healthcare fraud, an issue of significant financial magnitude in the United States, incurs annual costs amounting to several billion dollars. The escalating cost of medical care has resulted in a concomitant increase in instances of healthcare fraud. The identification of healthcare fraud has emerged as a primary focus in efforts to reduce healthcare expenses, given the ongoing scrutiny of inefficiencies within the healthcare model in America. Various methodologies could be employed to examine healthcare claims data in order to identify potential occurrences of fraudulent activities. The detection of healthcare fraud presents distinct challenges in comparison to other forms of fraud detection, such as credit card or auto insurance fraud. This is primarily due to the intricate nature of the healthcare sector, which encompasses various stakeholders including beneficiaries (patients), insurance companies, and healthcare providers.

User profiles are constructed based on historical data, and methodologies in these respective domains frequently seek instances where a user's behaviors diverge from their established profile. Nevertheless, these approaches are often unsuitable for addressing healthcare fraud due to the fact that the individuals in the healthcare setting who are affected by the fraud are the beneficiaries, rather than the perpetrators. This implies that the healthcare sector necessitates more sophisticated analytical methods in order to identify instances of fraudulent activity. Numerous data-analytics-driven methodologies have been investigated in [15] as prospective solutions for addressing healthcare fraud. The major advantages of data-based fraud identification are derived from the automated identification of fraud patterns and the prioritization of problematic cases. The predominant unit of analysis in healthcare research is commonly referred to as the "episode of care," denoting a collection of interconnected medical services provided to address a singular health issue.

Data Analytics for Pharmaceutical Discoveries

The process of bringing a new pharmaceutical product to the market, which is developed using groundbreaking chemical techniques, typically requires a time frame of nearly ten years and incurs expenses amounting to hundreds of millions of dollars. The period during which medications frequently fail in clinical trials is commonly referred to as the "valley of death." Clinical trials for novel compounds frequently result in failure due to the occurrence of adverse side effects or the compound's lack of efficacy. The field of drug discovery and development is experiencing significant advancements due to the utilization of interdisciplinary computational techniques that encompass computer science, statistics, medicine, biology, and chemoinformatics. Data analytics possesses the capability to streamline the process of pharmaceutical discoveries by providing guidance to experts regarding the appropriate course of action for hypothesis testing and other modes of inquiry. The utilization of data analytics can serve multiple objectives during the process of drug research and development.

The field of study encompasses distinct data analysis methods that can be broadly categorized into two phases: pre-marketing and post-marketing, each pertaining to the drug in question. The pre-marketing stage of data analytics involves the identification of associations between targets, biomarkers and pharmaceuticals, genes and illnesses, drugs and drugs, proteins and diseases. These discovery activities aim to uncover signals that demonstrate these relationships. One significant application of data and analytics during the post-marketing phase involves the detection and assessment of potential adverse effects associated with medications that have already received approval. These strategies offer a comprehensive repository of potential associations between pharmaceutical substances and negative outcomes, which can be utilized in subsequent investigations.

Clinical Decision Support Systems

Clinical Decision Support Systems (CDSSs) [16] are applications launched to aid medical practitioners in creating critical decisions regarding patient care, encompassing aspects such as diagnosis and therapy. Given their demonstrated ability to enhance both costs of care and patients' outcomes, CDSS have become an essential component in the assessment and enhancement of patient therapy. The diagnostic processes employed have demonstrated the capacity to yield more precise diagnoses, thereby potentially mitigating analytical errors by promptly notifying medical practitioners of potentially hazardous medication interactions. The benefits of CDSS include the ability to estimate both the clinical and economic outcomes of alternative treatment techniques, as well as predict therapy outcomes under specific circumstances [18].

The success of Clinical Decision Support Systems (CDSS) can be attributed to their electronic nature, their ability to seamlessly integrate with clinical processes, and their capacity to provide timely and location-specific decision assistance. The domains of pharmacy and billing within the healthcare sector have witnessed substantial advancements as a result of CDSS. Pharmacists have the ability to employ CDSS in order to conduct screenings for detrimental medication interactions, subsequently notifying the prescribing physician or medical facility of any potential issues. The application of CDSS in billing departments has facilitated the development of treatment plans that effectively optimize both patient care and financial resources.

Computer-Aided Diagnosis

In order to enhance the precision of mammography as a method for breast cancer screening, a computer-assisted detection (CAD) model was developed during the 1960s. Currently, this subject holds significant prominence within the realm of medical image processing and radiomics. Prominent areas of research in the domain of CAD presently encompass computer-aided diagnosis (CADx) and computer-aided detection (CADe). The identical conventions are employed for the abbreviation of these nouns. Computer-aided detection (CADe) is a method that examines the data produced by computer systems in order to precisely identify the specific site of any anomalous lesions on the patient's anatomy. The report generated by CADx, however, delineates the categorization of the patient's lesions into multiple categories. **Fig 2** illustrates the four-stage workflow of a conventional CAD system within the domain of medical image analysis.

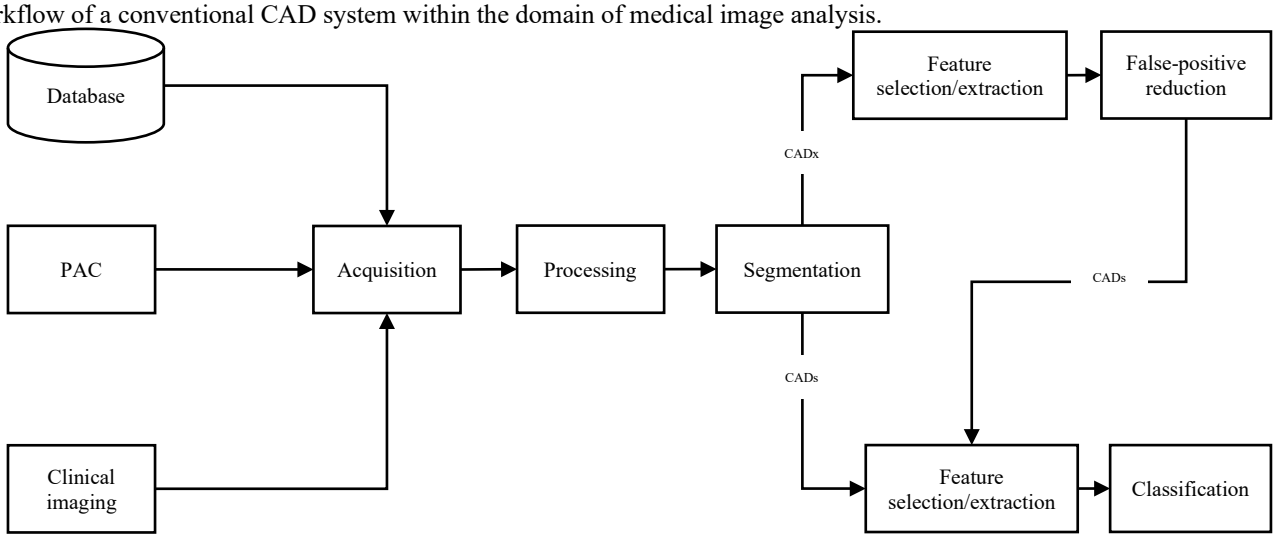

Fig 2. The categorization of medical images using image processing techniques

The methodology encompasses several steps, namely image cleaning, removal of extraneous elements, feature selection and extraction, and lesion classification. Computer-aided design (CAD) systems [17] are extensively utilized in the field of medical image analysis to facilitate the identification and assessment of various conditions and diseases like lung cancer, prostate cancer, and breast cancer, which include malignant growths. Moreover, computer-aided design (CAD) software is employed in the development of diagnostic imaging instruments such as CT and MRI scanners. The utilization of CAD tools has the potential to decrease diagnostic imprecision, decrease the time required to complete tasks, and alleviate stress

experienced by radiologists. It is feasible to categorize an image into distinct classifications, such as "healthy" and "pathological."

V. CONCLUSIONS

Advancements in hardware and software have facilitated the process of collecting healthcare data, resulting in significant advancements in the field of data analytics. The field has encountered substantial challenges due to the inherent lack of structure in the data and the privacy constraints imposed by data collection and distribution systems. The demand for real-time processing and interpretation of data has increased significantly due to the exponential growth in data volumes. The retrieval and analysis of data may necessitate the utilization of more sophisticated methodologies due to its inherent complexity. The rapid accumulation of extensive data facilitated by emerging data collection techniques, coupled with their potential to enhance the field of analytics, concurrently introduces novel challenges. In this study, we present evidence supporting the efficacy of machine learning techniques in accurately and effectively detecting the presence of coronary artery disease (CAD). Our findings are based on the analysis of a publicly available dataset. By utilizing the dataset, we were able to accomplish this task. Recent advancements in medical imaging techniques have facilitated the possibility of minimizing the duration dedicated to anomaly detection through the utilization of novel analytical approaches. Furthermore, it is imperative that these methods yield outcomes of greater reliability. The healthcare sector utilizes a diverse range of methodologies due to the inherent variation in prevailing types of data.

Data Availability

No data was used to support this study.

Conflicts of Interests

The author(s) declare(s) that they have no conflicts of interest.

Funding

No funding was received to assist with the preparation of this manuscript.

Ethics Approval and Consent to Participate

Not applicable.

Competing Interests

There are no competing interests.

References

[1]. E. Duman and Z. Tolan, “Ensemble the recent architectures of deep convolutional networks for skin diseases diagnosis,” Int. J. Imaging Syst. Technol., vol. 33, no. 4, pp. 1293–1305, 2023.
[2]. M. A. Hussain, N. Siersbæk, and L. P. Østerdal, “Multidimensional welfare comparisons of EU member states before, during, and after the financial crisis: a dominance approach,” Soc. Choice Welfare, vol. 55, no. 4, pp. 645–686, 2020.
[3]. A. Haldorai, A. Ramu, and S. Murugan, “Mobile and Pervasive Computing for Urban Development,” Computing and Communication Systems in Urban Development, pp. 1–26, 2019, doi: 10.1007/978-3-030-26013-2_1.
[4]. Y. Jiang, E. Byrne, J. Glassey, and X. Chen, “Reduced-order modeling of solid-liquid mixing in a stirred tank using data-driven singular value decomposition,” Chem. Eng. Res. Des., vol. 196, pp. 40–51, 2023.
[5]. S. Yang and N. Ling, “Robust projected principal component analysis for large-dimensional semiparametric factor modeling,” J. Multivar. Anal., vol. 195, no. 105155, p. 105155, 2023.
[6]. A. Levina and S. Taranov, “Creation of codes based on wavelet transformation and its application in ADV612 chips,” Int. J. Wavelets Multiresolut. Inf. Process., vol. 15, no. 02, p. 1750014, 2017.
[7]. G. Dixon, D. Livingstone, L. Copping, and M. Hollis, “Genomics: new discoveries and commercial developments,” J. Chem. Technol. Biotechnol., vol. 75, no. 10, pp. 867–867, 2000.
[8]. J. Wen and L. Yi, “Natural language processing for corpus linguistics by Jonathan Dunn. Cambridge: Cambridge university press, 2022. ISBN 9781009070447 (PB), ISBN 9781009070447 (OC), vi+88 pages,” Nat. Lang. Eng., vol. 29, no. 3, pp. 842–845, 2023.
[9]. Y. Wang, C. Sun, Y. Wu, L. Li, J. Yan, and H. Zhou, “HIORE: Leveraging high-order interactions for unified entity relation extraction,” arXiv [cs.CL], 2023.
[10]. I. S. Khayal, W. Zhou, and J. Skinner, “Structuring and visualizing healthcare claims data using systems architecture methodology,” World Acad. Sci. Eng. Technol., vol. 11, no. 4, pp. 342–346, 2017.
[11]. C. S. Kruse, M. Mileski, R. Syal, L. MacNeil, E. Chabarria, and C. Basch, “Evaluating the relationship between health information technology and safer-prescribing in the long-term care setting: A systematic review,” Technol. Health Care, vol. 29, no. 1, pp. 1–14, 2021.
[12]. N. Howie, F. Howie, and P. Seville, “Comparison of the scope of practice of physician associates with that of healthcare professions with prescribing responsibility from point of registration,” Future Healthc J, vol. 10, no. 1, pp. 38–45, 2023.
[13]. W. B. Lin and T. Y. Ku, “The influences of service quality of online order and electronic word of mouth on price sensitivity using loyalty as a mediating variable,” Int. J. Electron. Bus., vol. 12, no. 3, p. 215, 2015.
[14]. D. B. George et al., “Technology to advance infectious disease forecasting for outbreak management,” Nat. Commun., vol. 10, no. 1, 2019.
[15]. J. Koreff, M. Weisner, and S. G. Sutton, “Data analytics (ab) use in healthcare fraud audits,” Int. J. Acc. Inf. Syst., vol. 42, no. 100523, p. 100523, 2021.
[16]. S. Ayub, R. Boddu, H. Verma, S. Revathi B, B. K. Saraswat, and A. Haldorai, “Health Index Estimation of Wind Power Plant Using Neurofuzzy Modeling,” Computational and Mathematical Methods in Medicine, vol. 2022, pp. 1–8, May 2022, doi: 10.1155/2022/9535254.
[17]. S. Little and P. Brown, “The functional role of beta oscillations in Parkinson’s disease,” Parkinsonism Relat. Disord., vol. 20, p
[18]. X. Yang, A. Joukova, A. Ayanso, and M. Zihayat, “Social influence-based contrast language analysis framework for clinical decision support systems,” Decis. Support Syst., vol. 159, no. 113813, p. 113813, 2022.