A Review of Medical Data Sources, and Advanced Data Analytics in the Medical Sector

June Huh Eddie
Department of Biomedical Engineering, Bauman Moscow State Technical University, Moscow, Russia, 105005.
eddiehuh@bmstu.ru

Correspondence should be addressed to Mia Ella Grace: miagrace@outlook.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/JBSHA202303011
Received 18 March 2022; Revised from 10 September 2022; Accepted 10 November 2022.
Available online 05 July 2023.

Published by AnaPub Publications

Abstract – This article provides a thorough examination of healthcare data analytics and identifies various unresolved questions that necessitate further investigation. The implementation of healthcare analytics has the potential to generate two supplementary advantages for healthcare providers, specifically heightened patient contentment and ameliorated health consequences. The field of data analytics has been propelled forward by the facilitation of healthcare data acquisition through technological and software advancements. The lack of a predetermined framework in the data, along with the constraints imposed by privacy considerations in data gathering and dissemination methods, have presented significant obstacles to the progression of the field. The expeditious handling and comprehension of data requires prompt decision-making in the presence of substantial information quantities. In specific situations, the retrieval and analysis of data may necessitate the utilisation of more advanced techniques owing to the intricacy of the data. The emergence of data collection technologies that facilitate analytics has presented new obstacles, despite their effectiveness in accumulating vast quantities of data. The healthcare industry employs a wide array of methodologies, which can be attributed to the inherent distinctions in the fundamental data types.

Keywords – Healthcare Data Analytics, Electronic Health Records, Clinical Text Mining, Clinic-Genomic Data Integration, Temporal Data Mining.

I. INTRODUCTION

The field of medicine is currently experiencing a significant shift due to technological progress in domains such as electronic health records, portable devices, and intelligent wearables. The proliferation of large-scale healthcare data, coupled with advancements in computational techniques, has enabled researchers and medical practitioners to extract and visualize medical big data to an unparalleled extent. The utilization of scientific programming has the potential to facilitate the dissemination of vast quantities of healthcare data, enable the construction of composite systems models, and facilitate the derivation of insights from healthcare data and simulations. These capabilities are crucial for the identification of solutions to present and future healthcare challenges.

Visualization is a potent tool that enables the use of diagrams, graphics, or animations to communicate healthcare information and enhance comprehension. Programming tools such as Apache Hadoop [1], InformaticaPowerCenter [2], and Tableau [3] are among the options that can be utilized to proficiently scrutinize data and present the acquired insights derived from big data. Further investigation is necessary from a scientific programming perspective to effectively incorporate big data into the healthcare industry. The healthcare industry stands to gain significant advantages from research endeavors that concentrate on the technical and computational obstacles associated with the management of extensive volumes of data, commonly referred to as big data. A diverse range of techniques has been employed by both scholars and professionals.

The progress in general computer technology has significantly impacted the provision of medical treatment for patients. Data analytics is a crucial aspect of these computer systems. Healthcare big data encompasses a variety of sources, including MRI, X-ray, and Electronic Health Records (EHRs), CT scan images, biomedical signals such as EEG and ECG, handwritten prescriptions, and datasets from medical and wearables devices. The dynamic nature of health data poses a significant challenge to conventional methods and tools, impeding their ability to effectively manage and assess such information. The demand for efficient techniques of data analytics is imperative to analyze diverse datasets and enhance the quality of decision-making processes. **Fig 1** shows the different forms of big data analytics utilized in the medical industry.

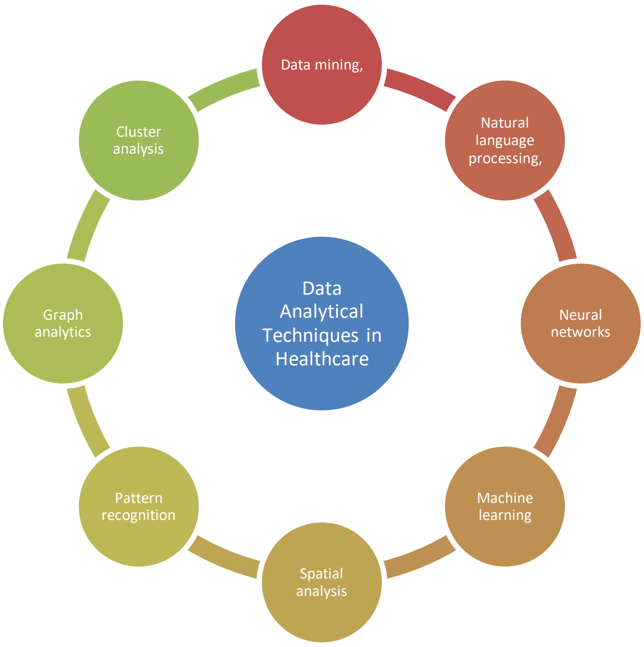

Fig 1. Healthcare data analytics methodologies

Analytical solutions possess significant possibility to transform medical delivery from reactive to proactive approach when applied to healthcare data. In the forthcoming years, the healthcare industry will be increasingly impacted by analytics. The latent patterns inherent in the data can typically be comprehended through the analysis of health data. Moreover, this technology will empower medical practitioners to establish distinctive patient profiles and accurately assess the likelihood of a particular patient encountering a medical condition in the immediate future. The healthcare industry is characterized by a wealth of data that originates from diverse sources such as sensors, images, textual data from biological literature and medical nodes, and traditional electronic records. The variability in the process of data representation and gathering poses a significant hindrance to the interpretation and processing of the data. The diverse nature of the data necessitates a range of strategies for its examination.

Furthermore, the diverse nature of the data inherently gives rise to several challenges pertaining to the analysis and integration of data. The employment of different forms of data can often facilitate the identification of valuable insights that would be unattainable through the examination of a solitary data source. Recently, the significant potential of integrated data analysis methods has been revealed. The interdisciplinary nature of healthcare poses a significant challenge for both researchers and practitioners. The healthcare industry has often reaped the rewards of advancements stemming from distinct domains such as data mining, medical research, databases, information retrieval, and medical practitioners. The presence of multiple disciplines within the field contributes to its richness, yet simultaneously poses challenges to achieving significant advancements.

The mathematical and statistical foundation required for data analytics is often lacking among medical practitioners and researchers, whereas computer scientists typically lack training in domain-specific medical concepts. Although it is apparent that a significant portion of the available data could potentially be enhanced through the utilization of sophisticated analytical methodologies, this has posed a difficulty in consolidating a cohesive body of literature on this topic. The diversity often leads to distinctive career trajectories stemming from a wide range of perspectives. Professionals working in the domain of data analytics are liable to disengaging from practical domain-oriented challenges, typically suggesting problem formulations that are technically impressive but lack practical application.

This article makes an effort to bridge these disparate groups by thoroughly examining the most significant contributions from every field. Only by uniting these many communities can the enormous scope of data analysis techniques be fully realized.The rest of the article has been organized as follows: Section II provides a discussion of basic analytics and medical data sources. In this section, various concepts are discussed: electronic health records, biomedical image analysis, sensor data analysis, biomedical signal analysis, genomic data analysis, clinical text mining, mining biomedical literature, and social media analysis. Section III focuses on advanced data analytics for the medical sector. In this section concepts such as clinical prediction models, temporal data mining, visual analytics, clinic-genomic data integration, information retrieval, and privacy-preserving data publishing, have been discussed. Lastly, Section IV provides final remarks to the paper.

II. BASIC ANALYTICS AND MEDICAL DATA SOURCES

This section will address the various sources of data and their effects of analytical techniques. The diverse nature of medical data sources necessitates the utilization of a broad spectrum of methodologies drawn from various fields of data analytics for effective mining.

Electronic Health Records

Electronic Health Records (EHRs) [4] are utilized to document the medical history of patients. The comprehensive medical record encompasses various aspects of the patient's medical information, such as their personal characteristics, medical conditions, prescribed medications, physician's observations on their recovery, vital signs, as well as their past medical history, laboratory findings, radiographic images, and financial data. EHRs have the potential to encompass not only a patient's medical or treatment history, but also broader perspectives on their care. The utilization of EHRs is beneficial due to their ability to facilitate streamlined and productive communication among healthcare professionals and organizations.

Electronic health records (EHRs) are designed to enable authorized users to access and edit them in real-time. This exhibits significant potential for practical application. A specialist or secondary care facility may require access to the medical records of a patient's primary care physician. The utilization of EHRs enhances operational productivity as personnel can promptly retrieve the latest patient data. The utilization of a comprehensive clinical encounter record for a patient has the potential to facilitate various care-related responsibilities such as quality monitoring, outcomes reporting, and evidence-based decision support. EHRs streamline the process of storing and accessing medical records. This development has resulted in several favorable outcomes, including improved care coordination, enhanced diagnostic precision, better health outcomes, and increased patient engagement.

Biomedical Image Analysis

Medical imaging is gaining increasing prominence as a means of visualizing and analyzing data in the fields of medicine and biology. In the last decade, there has been a significant surge in the complexity of instruments employed for the acquisition, retention, transfer, examination, and demonstration of visual information. These instruments are significantly enhancing the capacity of biologists, healthcare scientists, and biochemists to observe their subjects of research and gather quantitative data to support scientific hypotheses and medical diagnoses. Medical photographs of low quality are widely known to be difficult to comprehend due to various factors such as artifacts, inadequate contrast, and noise. The efficacy of traditional or modern algorithms for detection and analysis may be limited due to the presence of distinct forms of interference, which vary between mammograms and ultrasound images. The biomedical imaging scenario poses a significant challenge due to the presence of non-Gaussian, nonlinear and nonstationaryevents such as bursts, ruptures, and transients as well as complex interactions between various components. Consequently, the majority of present research endeavors are concentrated on improving the inadequate caliber of extant biomedical imaging data.

The precise examination of medical images could be of great significance to medical practitioners and researchers in aiding with disease surveillance, devising treatment strategies, and predicting patient outcomes. Computed tomography, and magnetic resonance imaging, ultrasound, and positron emission tomography are widely employed imaging modalities for the acquisition of biological images. The capacity to non-invasively inspect internal human organs holds significant implications for the field of medicine. These technological advancements enable medical practitioners to gain further insights into a patient's medical state without necessitating invasive procedures. Observing images of such anatomical structures can be advantageous; however, it merely marks the initial phase of the procedure. The ultimate aim of biomedical image analysis is to derive numerical information and make inferences regarding a patient's health status based on their visual medical records.

The analysis holds significant societal importance as it serves as a crucial tool in comprehending biological systems and addressing health-related issues. Nonetheless, the task presents numerous difficulties due to the diverse and intricate nature of the images, which may feature irregular shapes and noisy values. Several broad research problem categories that emerge in the analysis of images include image segmentation, object detection, feature extraction, and image registration. The resolution of these challenges will facilitate the production of significant analytical computations, which could operate as an input for other fields of medical data analytics.

Sensor Data Analysis

According to the assertions made by Rodrigues, da Righi, da Costa, and Antunes [5], the foremost concerns regarding IoT sensor networks are centered on the precision and capacity for expansion of the sensor data. The utilization of techniques for mining and analyzing sensor data, including data management, data collection, data mining, and knowledge discovery has been employed to tackle these issues. The acquisition of knowledge and the ability to make informed decisions are achieved within this domain by utilizing deep learning and machine learning models.

Machine learning models

Iannacci [6]has discussed the necessity of a simplified approach to extract valuable insights from the data obtained through Internet of Things (IoT) sensors. In the context of analyzing data obtained from IoT sensors, it is imperative that the models of machine learning are implemented on embedded processors that are integrated within the sensor. In order to accomplish this task, it is necessary to employ specialized system programming and implement an efficient data system that can effectively handle the unique characteristics of sensor data generated by the IoT in real-time. The utilization of a Gaussian Mixture Model (GMM) has been proposed by the authors as a means of addressing the heterogeneity of characteristics present in sensor data. Real-time hardware and software collaboration, along with continuous algorithms of machine learning, were employed to attain data categorization and training for IoT-oriented applications.

Deep Learning Models

The utilisation of deep learning for feature learning from Internet of Things (IoT) sensor data was proposed by Chaudhary, Goyal, Benslimane, Awasthi, Alwadain, and Singh [7]. To enhance the precision of the indistinct attributes of IoT sensor data, it is imperative to employ real-time deep learning-oriented classification. Nevertheless, the execution of deep learning frameworks on resource-limited sensor boards incurs high processing costs. Consequently, Alam, Shuvo, Ali, Ahmed, Chakma, and Jang [8] recommended a pre-processing approach in the spectral field, followed by the utilisation of deep learning frameworks.

Neural Networks for IoTs Sensor Data Analysis/Processing

Artificial neural networks (ANNs) are a suitable choice for addressing pattern recognition and function estimation tasks through the utilization of supervised learning methodologies. The neural network's architecture comprises the hidden layer, output layer, and the input layer. The Convolutional neural networks (CNNs) are comprised of layers within the network that performs convolution operations on one-dimensional or two-dimensional sensory data, as stated by Parcham and Fateh [9]. The convolution operator is a fundamental tool utilized in machine learning to acquire knowledge of the local patterns present within the data. The appropriate convolutional neural network variant is selected based on the available data. CNN is primarily well-suited for processing image data, particularly in the context of analyzing image sensor data. The majority of Internet of Things (IoT) devices, including but not limited to unmanned aerial vehicles (UAVs), intelligent automobiles, and mobile phones, are outfitted with optical sensors.

Numerous Internet of Things (IoT) applications, such as predicting floods, landslides, and forest fires through drone imagery, as well as managing traffic, rely on the utilization of cameras installed on vehicles. Convolutional Neural Networks (CNNs) are capable of extracting high-level features from 2D or 1D images, speech, or signals by utilizing a sequence of hidden layers. The neural network architecture comprises of concealed layers, which consist of a completely interlinked pooling layer and a convolutional layer positioned at the terminal end. The Convolutional Neural Network (CNN) utilizes a collection of filters, known as learnable parameters, within its convolution layer. These filters serve as the fundamental component of the CNN, as they are responsible for filtering high-dimensional dataset to a low dimension. This process aids in the extraction of the most appropriate feature from input images.

Sensor data is extensively utilized in the clinical field for both retrospective and real-time analysis. The Electrocardiogram (ECG) and Electroencephalogram (EEG) are medical instruments that are utilized for data acquisition. These tools function as sensors, which gather impulses from different parts of the human body. Occasionally, and with greater frequency, these instruments are utilized for the analysis of acquired data in real-time. Real-time analysis has been identified as having significant potential in the remote monitoring and intensive care unit contexts of patients with categorical clinical conditions. In each of these contexts, the data volume that necessitates processing can be substantial. In the context of an intensive care unit (ICU), it is commonplace for sensors to obtain input from different sources of data, often numbering in the hundreds, with the imperative that alerts be promptly activated. The utilization of big-data models and dedicated hardware systems is imperative for the successful implementation of these applications.

The study of remote monitoring applications is concerned with both the long-term analysis and real-time events of diversified trends as well as treatment alternatives. The rapid proliferation of sensor data holds significant potential for transforming the healthcare sector, yet it also presents the challenge of managing data overload. Hence, the development of novel data analysis tools is of paramount importance to effectively process copious amounts of acquired data into comprehensible and valuable insights. The implementation of analytical techniques has the potential to enhance situational awareness in clinical settings, facilitate more accurate monitoring of patients' physiological signals, and potentially reveal inefficiencies within the healthcare system that contribute to escalating costs.

Biomedical Signal Analysis

The signals that are analyzed in biomedical signal analysis originate from physiological processes. Electroneurograms, electrocardiograms, electromyograms, electrogastrograms, electroencephalograms, and phonocardiograms are among the various types of signals that can be observed. Comprehending these signals holds paramount importance in accurately diagnosing the condition and determining the optimal treatment plan. The quantification or comparison of physiological signals can offer an assessment of an individual's state of well-being. Data is obtained through the utilization of invasive and non-invasive sensors and transducers. The nature of the medical intervention or the gravity of the ailment may determine whether the indications are discrete or continuous.

The intricate interrelation among physiological systems and the limited ratio of useful signal to extraneous noise pose challenges in analysis and processing of physiological data. Extensive preprocessing of signal data collected from medical instruments may be required owing to the presence of high levels of background noise. The development of various signal processing algorithms has significantly contributed to our understanding of physiological phenomena. Various applications of [36] include filtering, noise cancellation, and compact methods. The literature has extensively examined various approaches to dimensionality reduction, including but not limited to singular value decomposition (SVD) [10], wavelet modification [11], and principal component analysis (PCA) [12].

Biomedical signal analysis confers significant advantages to medical diagnosis, monitoring, treatment, and outcome evaluation. The evaluation phase entails assessing the capacity to fulfill functional requirements, obtaining evidence of performance, and conducting quality control measures. The process of information gathering involves the measurement of phenomena to facilitate the interpretation of a system. Diagnosis, on the other hand, entails the detection of any malfunction or abnormality within the system. Monitoring is a continuous process of gathering data on systems, while control and therapy involve modifying the behaviors of system-oriented on the results to meet the requirements. Lastly, the assessment is conducted to assess the system's ability to meet the specified requirements. The utilization of biological signal analysis in conjunction with computer-assisted diagnosis and treatment is illustrated in **Fig 2**. Biomedical instrumentation is utilized for the purpose of gathering patient information, including but not limited to electrocardiographic (ECG) data. Hence, it is crucial to design all biomedical equipment and signal analysis systems while prioritizing the patient's well-being and comfort. The process of acquiring signal data involves several components, including transducers, signal conditioning hardware, and analog-to-digital conversion.

Transducers can manifest in various forms, including but not limited to sensors and ECG patch electrodes. Amplifiers and filters are signal-conditioning mechanisms that are utilised to either increase or decrease a signal, respectively. The transducer does not effectively intensify the power. The acquisition of data is succeeded by the processing of signals. The presence of numerous signals in the vicinity often hinders the detection of biological signals, thereby necessitating the mitigation of interference, artefacts, and noise for accurate signal analysis. At this juncture, the selection of a filter type is determined, which may include a time-domain filter, a frequency-domain filter, or an adaptive filter.

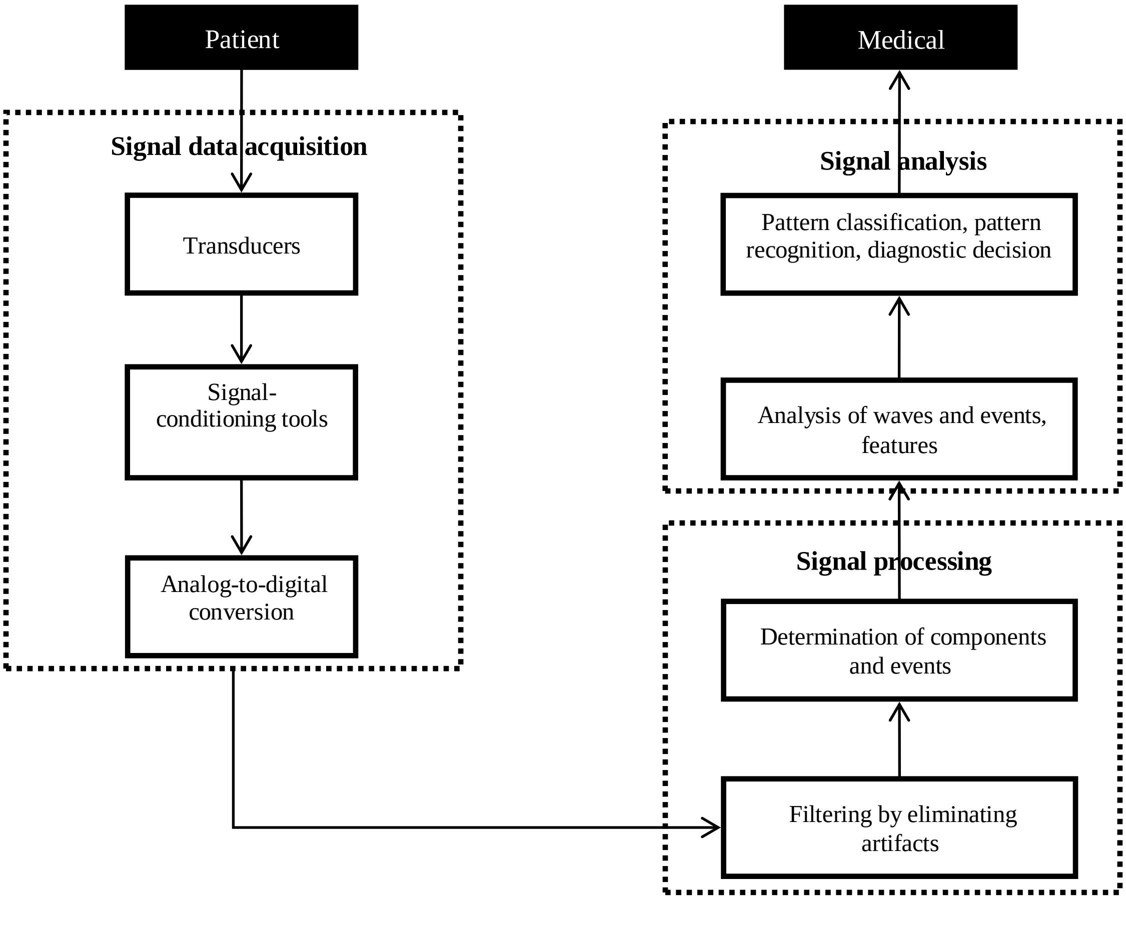
Patient
Medical
Signal data acquisition
Transducers
Signal-conditioning tools
Analog-to-digital conversion
Signal analysis
Pattern classification, pattern recognition, diagnostic decision
Analysis of waves and events, features
Signal processing
Determination of components and events
Filtering by eliminating artifacts

Fig 2. Biomedical signals through the use of computer-aided methods for the purposes of diagnosis and therapy

The process of detecting events and waves is essential for the purpose of isolating the segment of a signal that is relevant to particular events of interest. The identification of QRS, T, and P, waves is a common practise in electrocardiography. Upon identification of the events, it is imperative to conduct signal analysis. Fixed segmentation and adaptive segmentation are two techniques employed for the analysis of non-stationary information. The process of categorising a signal into multiple potential categories necessitates the utilisation of pattern recognition techniques that rely on signal analysis. This, in turn, facilitates the diagnostic procedure. The primary objective of analysing biological signals is to arrive at a diagnosis by evaluating the symptoms presented by the patient.

Genomic Data Analysis

Numerous diseases exhibit a heritable aspect, however, the precise correlation between specific genetic indicators and disease remains ambiguous. Although it is widely acknowledged that diabetes has a hereditary component, the complete array of genetic markers that confer susceptibility to the disease remains unknown to researchers. In certain circumstances, such as cases of blindness caused by Stargardt disease, the crucial genes have been ascertained, although not all conceivable mutations have been delineated. It is evident that an enhanced comprehension of the associations among genetic markers, mutations, and disease will significantly aid in the advancement of varied gene therapies for the treatment of these disorders. Specifically, it is imperative to ascertain the types of health-related inquiries that can be addressed through conventional data-driven investigation utilising in-silico scrutiny of genetic data. Several challenges need to be addressed before genetic discoveries can be effectively implemented in customised medical interventions. The genomic profiles observed in diseases such as cancer exhibit a significant degree of complexity and manifest considerable heterogeneity among individuals. By addressing these issues, we can significantly advance the development of personalised medicine.

Advancements in biotechnology have enabled the prompt production of extensive amounts of medical and biologicaldata, thereby enabling more comprehensive investigations in the realm of genomics. The aforementioned development has facilitated novel opportunities for investigating recalcitrant concerns in the field of biological sciences pertaining to the entirety of the genome. The genomic technology advancements have enabled the comprehensive exploration of complex disorders by studying the complete genetic makeup of healthy individuals. Encouraging results have been identified in various research avenues, indicating that they will facilitate novel perceptions into the science of human ailments and enhance the capacity to anticipate individual treatment response. Furthermore, genetic data is frequently represented through the utilisation of sequence models and network models. Professionals operating within the sector are required to possess a proficient understanding of sequence and network mining techniques. The identification of biomarkers for diseases and targets for treatment, along with the anticipation of clinical outcomes, are currently among the most critical issues in the field of medical research.

Clinical Text Mining

The vast majority of patient data is logged in the medical notes format. The aforementioned records serve as the fundamental framework of healthcare data and are frequently maintained in an unorganised data structure. The aforementioned comprise of health information pertaining to patients that is acquired through methods such as dictation transcription, direct input from physicians, or employment of voice recognition software. These sources represent a valuable reservoir of knowledge that has yet to be fully explored. The process of manually encoding clinical information from a free-text form, even if limited to secondary and primary treatments and diagnoses for billing, is evidently a costly and time-consuming endeavour. The process of converting unstructured clinical data into a structured format is challenging, resulting in the complexity of mechanically analysing such notes. The complexity of the matter arises due to the uniqueness of each patient and physician's circumstances, coupled with the unstructured, heterogeneous, diversified, and multifaceted nature of the data. The utilisation of natural language processing (NLP) and entity extraction is crucial in the timely and automated encoding of clinical information. These techniques enable the inference of valuable knowledge from extensive amounts of clinical material.

Data preparation methodologies hold greater significance than actual mining techniques. The utilisation of brief and concise sentences, dictation, lexicons such as acronyms and abbreviation, and occasionally erroneous clinical terminology all contribute to the increased complexity of processing clinical material through natural language processing methods compared to other forms of text. The challenges posed by the aforementioned issues render clinical text processing a complex task, as they impact various fundamental natural language processing (NLP) tasks like complete or shallow parsing, text classification, and sentence segmentation. **Fig. 3** illustrates the fundamental framework of TM's clinical implementation. The categorization of cancer, such as breast cancer, will limit the potential application of TM to a specific area. The determination will be made regarding the specific categories of textual data, such as mammography reports, that will be utilised.

Professionals, whether they be human or artificial intelligence, require specialised knowledge in a particular field in order to accurately analyse data. Various sources encode pertinent information in machine-readable formats, such as comprehensive dedicated databases (such as BCGD), knowledge representation frameworks (such as UMLS), and standardised reporting protocols (such as BI-RADS). The information that is extracted, including the stage of cancer, is a result of the combination of data and knowledge, and is obtained through the utilisation of specialised software systems for

Text Mining. The extracted data can be utilised for various purposes, including clinical research such as epidemiology, and decision-making processes such as selecting an appropriate therapy. It is imperative to verify the reliability of information obtained through automated means prior to its application in a therapeutic setting. TM systems are evaluated using a range of metrics, such as negative and positive predictive values, to regulate the accuracy of the extracted data. This process is similar to the validation of diagnostic tests, which undergo a series of tests to ensure their reliability.

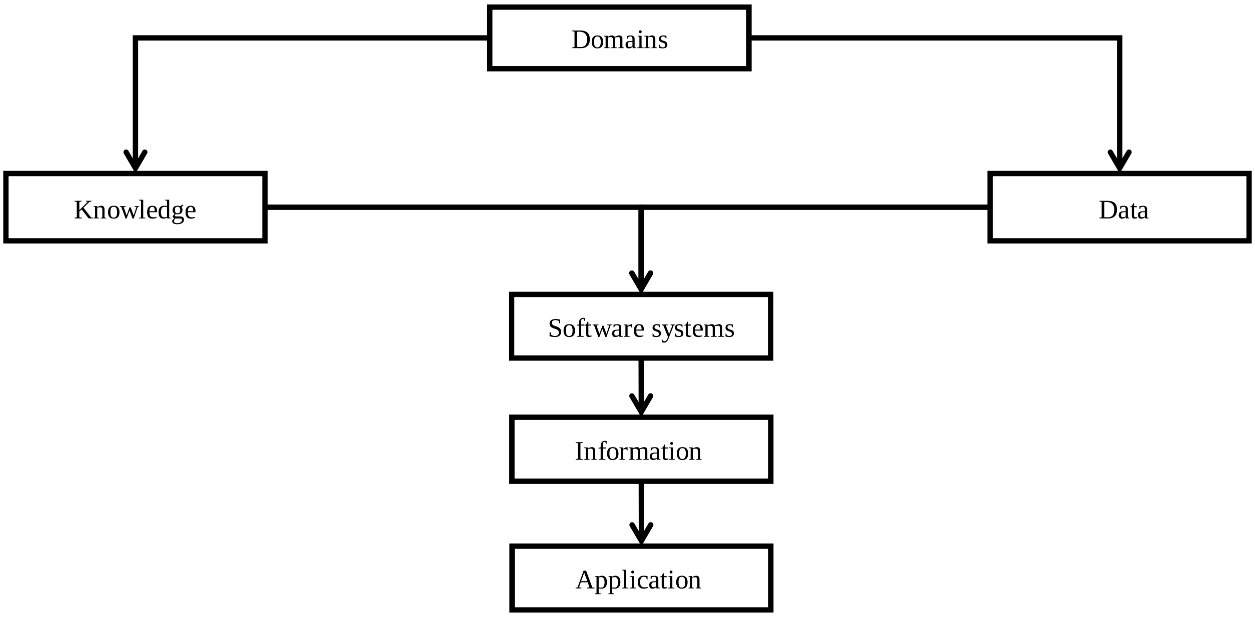

Fig 3. Text mining model for medical applications

Mining Biomedical Literature

The utilisation of biomedical literature evidence is prevalent across diverse contexts. The latter phenomenon is characterised by a significant abundance and notable expansion over an extended period. Text mining technologies play a crucial role in ensuring the longevity, accessibility, and usefulness of resources in biomedical applications that draw data from scientific publications. In the field of biomedicine, the utilisation of text mining tools and techniques presents novel prospects for the adoption of innovative methods of knowledge exploration. These technologies facilitate the process of knowledge discovery and creation by enabling efficient search, extraction, integration, analysis, and summarization of textual content. The interdisciplinary nature of text mining in the biomedical field poses a considerable challenge. Biologists utilise proprietary nomenclature to designate chemical compounds, whereas scientists typically opt for IUPAC-compliant language or unequivocal descriptors likeglobal biological identifiers.

Text mining algorithms are needed to manage and their relationships that are not well-defined in the literature, while cheminformatics tools can be used to address the former. In order to achieve this objective, it is imperative to employ methodologies for extracting entities and events from unstructured databases. The utilisation of text mining techniques presents innovative opportunities for the efficient population, updating, and integration of datasets that would otherwise be excessively costly to sustain. Text mining offers several advantages to biomedical research, such as the capacity to formulate hypotheses, the minimization of costs associated with expert knowledge validation, and the integration of textual data with biological pathways. This approach has the potential to be widely employed for enhancing the arrangement of biological data and revealing latent associations.

The formulation of hypotheses grounded on biomedical literature has demonstrated promise; however, there exist impediments that must be surmounted. Firstly, the limited complexity of certain methodologies, such as ABC co-occurrence-oriented technique, poses a challenge in effectively communicating intricate biological processes. Presently, there exist two primary challenges associated with LBD. Numerous contemporary LBD methodologies and systems have been developed primarily for research objectives, making it challenging to augment these technologies with comprehensive biological context. The specialised nature of biomedical articles may result in a biassed representation of their content. Therefore, it is crucial to implement these findings in practical settings, such as pharmaceutical study, medical care, and basis scientific research, to effectively utilise these systems and evaluate their potential benefits. Divergences could exist among the outcomes presented in different scholarly articles. Obtaining reliable theories in this particular scenario poses a challenge.

Social Media Analysis

The propagation of social medical instruments, such as blogs/microblogs, social networking platforms, online communities, and question-answering forums, has resulted to diversification of data on public opinions concerning different aspects of medical care. The utilisation of data obtained from social media platforms has the potential to provide valuable insights into population health and facilitate the monitoring of public health. The insights and recommendations of individuals on social media platforms may provide valuable knowledge to professionals in the field of public health. The aggregation of social media posts and messages, despite their limited individual informative value, can yield valuable insights when analysed on a large scale. The duration required for the collection of such complex data could be significantly reduced by half if these voluminous sets of information are appropriately analysed. The examination of social media analytics in the healthcare sector has been previously explored with the aim of enhancing intervention capabilities for health-related endeavours and gathering collective health trends, including instances of infectious disease outbreaks and accounts of adverse medication interactions.

The evaluation of the advancements in social media content has the potential to shed light on the emergence of disease outbreaks, rendering it a valuable resource for timely identification. Topic models are frequently employed to conduct a comprehensive analysis of health-related information. Online forums for doctors and patients offer an additional information resource that can be accessed through social media platforms. Owing to the pervasive prevalence of numerous medical conditions, online support communities have emerged as a valuable repository of knowledge. The reliability of data obtained from social media platforms is widely recognised as questionable, underscoring the necessity of approaching any resulting conclusions with caution.

III. ADVANCED DATA ANALYTICS FOR MEDICAL SECTOR

This section will provide an evaluation of different advanced data analytics techniques that are currently being utilised in the healthcare industry. Several machine learning and data mining models need to be tailored for application in the medical domain.

Clinical Prediction Models

The ability to make clinical predictions is a fundamental component of contemporary healthcare practises. Considerable research has been conducted, and numerous predictive models have been efficaciously employed in clinical settings. The impact of these models on the identification and treatment of illnesses has been significant. The clinical prediction tasks have been aided by three primary categories of supervised learning techniques, namely: (i) conventional statistical approaches such as logistic regression and linear regression; (ii) complex data mining and machine learningalgorithms such as ANNs and decision trees; and (iii) survival frameworks, which focus on the prediction of survival results. The main focus of these methods is on covariate variables, also identified as features and traits, and dependent results variables.

The foremost determinant in selecting an appropriate healthcare model for a given scenario is the anticipation of the expected outcomes. The scholarly literature presents a diverse array of predictive models to address the vast spectrum of potential outcomes. Two common types of outcomes are binary and continuous representations. Categorical and ordinal outcomes represent two additional, comparatively infrequent categories.

Numerous models have been devised to address survival outcomes when attempting to forecast the timings of certain events of interest. Survival frameworks are extensively investigated in the realm of medical data evaluation for their capacity to predict a patient's survival duration. The effectiveness of these forecasting models can be assessed and appraised through diverse methodologies.

Temporal Data Mining

Temporal data mining may be illustrated as the procedure of discovering knowledge within temporal datasets by identifying systems, such as temporal models or patterns, within temporal information and data. Any approach, which classifies temporal patterns or fit frameworks to temporal dataset, can be considered a temporal data mining technique. The purpose of temporal data mining is to employ a fusion of statistics, machine learning, and database algorithms to discover hitherto undiscovered associations within extensive data sequences, including temporal sequences and time series that comprise nominal symbols from the alphabet. The fundamental components of temporal data mining encompass the depiction of temporal data, the formulation of likeness metrics, and the mining operations per se.

Temporal analysis and inference are crucial in healthcare data due to the omnipresence of time-related information. Within the healthcare sector, there exist two main categories of temporally marked information. The initial type of data pertains to information obtained via sensors, whereas the subsequent type of data pertains to information obtained from electronic health records (EHRs). The likelihood of discovering temporal patterns within electronic health record (EHR) data that can enhance our understanding of the onset, development, and treatment outcomes of illnesses is considerable. The heterogeneity, sparsity, high dimensionality, and unpredictable time intervals of EHR data render traditional approaches inadequate. In contrast to electronic health record data, sensor data is typically presented as numerical time series that are regularly recorded at fixed intervals. The recordings encompass a range of electrical activity records, such as electrocardiograms (ECGs), electroencephalograms (EEGs), and other routine patient monitoring data.

In contrast to longitudinal electronic health record (EHR) data, which is frequently collected throughout a patient's lifespan, sensor data is captured over a significantly shorter duration, typically ranging from a few minutes to several days. The variances between EHR data and sensor data necessitate distinct methodologies for extracting temporal data. The utilisation of temporal pattern mining techniques is frequently observed in the extraction of EHR data. The employed methodologies involve the representation of data instances, such as patient results, as discrete events sequences, such as diagnostic procedures and codes. Subsequently, the identification and enumeration of statistically significant patterns contained in the data are pursued. Signal processing techniques and time-series analysis methodologies, such as the wavelet transform and independent component analysis, are frequently employed to extract meaningful insights from sensor data.

Visual Analytics

The vast majority of healthcare components are dependent upon the aggregation, arrangement, and evaluation of data derived from diverse origins, a significant portion of which is progressively increasing in terms of its magnitude and diversity. The aforementioned trend is evidenced by an increasing demand for methodologies and software tools capable of scrutinising and presenting said data to facilitate informed decision-making. The proposition has been put forth that the utilisation of data visualisation and visual analytics could potentially address this requirement.Data visualisation is a technique that involves generating graphical or pictorial depictions of data for the purpose of facilitating the recognition and understanding of patterns and trends. Visual analytics encompasses a range of disciplines, including machine learning, data mining, interactive visualization-oriented user interfaces, and computational evaluation that collectively support human cognition and analytical reasoning.

Health informatics has been a subject of focus for several years, with an emphasis on the effective utilisation of data for decision-making based on evidence. The transition from traditional paper-based medical charts to a comprehensive medical information technology (IT) system has led to a significant surge in the utilisation of data. According to Baksh et al. [13], Electronic Health Records (EHRs) have been extensively adopted by medical professionals such as doctors, certified nurse-midwives, nurse practitioners, and physician assistants at rates of 95% or higher. Additionally, 96% of hospitals in the country have implemented EHRs.Moreover, a considerable and increasing demographic in the United States and other regions employs individualised devices to gather self-generated health information with the intention of promoting personal well-being. According to a survey conducted by the Pew Research Centre in 2015, just over 50% of individuals who owned smartphones indicated that they utilised an application related to health. Individuals who possess an iPhone operating on iOS 10 or a more recent version may have a pre-installed health application that covertly monitors their physical activity.

In recent times, there has been a notable increase in scholarly inquiry and innovation directed towards comprehending and leveraging the substantial quantities of data produced through the extensive implementation of technological interventions in the healthcare industry within the United States. In this domain, there exist various activities such as health data exchanges, machine learning algorithms for data-oriented risks prediction and evaluation, and the incorporation of patient-based healthcare data into EHRs. The proliferation and expansion of health-related activities have led to an increased demand for visual analytics services. Consequently, novel visual analytics technologies and products have been developed and integrated into health IT systems to meet this demand. The implementation of data visualisation and visual analytics in healthcare settings has a broad spectrum of practical uses. Medical records are utilised at the point-of-care to determine the course of treatment for individual patients. Conversely, community health practises and health policy are informed by information derived from longitudinal cohort studies and public health data at the population level.

The utilisation of data visualisation and visual analytics tools holds promise in facilitating the examination and dissemination of intricate data and information, albeit their evaluation may present difficulties. In comparison to traditional interfaces, which are evaluated based on complex metrics such as error rates and tasks completion time, visualisation and visual analytics instruments are regularlypremeditated to enable the identification of insights derived from data. Various solutions have been proposed by researchers in the field of visualisation to address this issue. These solutions include quantifying insights, utilising structural approaches, which model visualization at distinct granularity levels, and adapting approaches from fields, such as visual analytics and heuristic analysis. Randomised controlled trials may not always be the optimal approach for evaluating medical therapies. Nevertheless, they can be useful in assessing design decisions and presenting proof of effectiveness.

The utilisation of datasets containing numerous clinical variables is a prevalent practise in the evaluation of diverse medical conditions. The synthesis and interpretation of clinical data can be challenging due to its diverse, noisy, multimodal, and temporal characteristics. The vast amount of data generated by healthcare organisations presents the possibility of creating novel interactive interfaces for the purpose of exploring extensive databases, verifying clinical data and coding methods, and enhancing transparency across various departments, hospitals, and organisations. Such opportunities hold significant potential for advancing healthcare practises. The healthcare industry has devised various distinctive visual techniques, whereas some of them can be conveniently modified from the data mining literature.

Clinico–Genomic Data Integration

The pathogenesis and advancement of human illnesses are influenced by a multitude of factors, including genomic, clinical, behavioural, and environmental variables, thereby introducing an additional layer of intricacy. The diverse effects of numerous parameters can be comprehended by supplementing clinico-pathological and genomic databases. To effectively integrate the essential information contained in genetic and clinical data, it is imperative to develop integrative frameworks that consider both genetic and medical features instantaneously. Models of this nature have the potential to facilitate the advancement of improved diagnoses, therapies, and medications, thereby moving us closer to the realisation of personalised healthcare. The emergence of clinico-genomic data integration has led to the development of promising prediction models that incorporate both clinical and genetic information.

The term "genetic data" pertains to the genetic information of a patient, encompassing metabolite profiles, gene expression, SNPs, and protein. On the other hand, "clinical data" pertains to a wider spectrum of information regarding the patients' surroundings, including pathology, behaviour, demographics, family, environment, and medications. The identification of clinical and genomic factors associated with a specific disease phenotype, commonly referred to as biomarker discovery, and represents a prevalent motive for conducting integrative studies. The outcome of interest may manifest as a binary event, such as the presence or absence of cancer, or as a continuous measure, such as the duration of survival following a particular therapeutic intervention.

Information Retrieval

The field of medical data analytics frequently encompasses the utilisation of data mining and analysis techniques to examine patient-related data. However, supplementary information such as scientific research and literature may also prove beneficial in this context. Frequently, information retrieval (IR) techniques are employed to obtain access to said data. The primary focus of Information Retrieval (IR) is knowledge-based information, which refers to data that is generated and organised through observational or experimental study. In contemporary times, infrared (IR) systems have gained extensive usage [14]. A significant proportion of internet users in the United States, exceeding 80%, have engaged in online searches for personal health information. Similarly, nearly all medical practitioners have leveraged the internet in their professional activities. The challenges pertaining to text mining in clinical and biological domains are closely linked to the models of information retrieval. The principal aim of information retrieval is to identify and retrieve the relevant material based on the user's requirements. The customary procedure involves initiating an inquiry aimed at the information retrieval system. Metadata serves as a means for a search engine to identify and retrieve pertinent content.

Information Retrieval (IR) comprises two fundamental components: indexing, which involves assigning information to the material, and retrieval, which involves the user inputting a query and receiving pertinent content. The inverted index is a widely recognised data structure utilised for efficient information retrieval, wherein each document is assigned a distinct identifier. Subsequently, every individual letter is associated with a directory comprising of file names. This representation is optimal for conducting a keyword search. The implementation of ranking systems is crucial to manage the potentially voluminous outcome of search results. Over time, numerous assessments have been carried out with a user-centric approach, specifically targeting individuals seeking biological information and evaluating the efficacy of searches within clinical settings.

Privacy-Preserving Data Publishing

There is a common belief that specific types of data, such as geographical units with high precision, cannot be adequately sanitised for release. It is also believed that the sanitization process may render the data unusable for certain types of research. However, it should be noted that numerous statistical organisations distribute sanitised public-use data sets. In response to this requirement, the governments of the United States, Canada, and Germany have established secure research data centres to facilitate access to more accurate and comprehensive data for researchers from other nations. The objective is to create a secure environment wherein proficient researchers can carry out investigations utilising computer systems that have limited connectivity to external networks and are subject to surveillance by census personnel. Approximately twelve analogous locations can be identified within the United States. Prior to commencing work on a study, it is imperative that a researcher's credentials undergo verification through a background check and a sworn declaration. Patel, Moore, Craver, and Feldman [15] conducted a thorough evaluation of the outcomes and information to ensure strict adherence to disclosure protocols prior to their removal from the facility. Several countries offer remote access servers, which are occasionally referred to as "virtual" secure research data centres.This instance illustrates the intricacy of the challenge pertaining to safeguarding privacy, notwithstanding the fact that this study does not centre on secure facilities and data centres. Legal statutes, conventional security protocols, and analogous tools comprise a larger repertoire that encompasses technological competencies for safeguarding privacy in data dissemination.

The focal point of process mining is gaining an understanding of the underlying processes themselves. Information systems frequently capture and provide access to such data. The confidential data that they encompass necessitates careful scrutiny. Consequently, an increase in discourse regarding privacy issues associated with process mining has been observed. Whilst privacy-protection techniques are anticipated to modify the initial data, they must execute this task in a manner that maintains the data's usability. The implementation of data alterations aimed at safeguarding privacy could potentially result in the introduction of bias into the research. Consequently, it is imperative to construct novel infrastructures that facilitate the dissemination of privacy-conscious event data. The primary objective of such infrastructures is to furnish metadata that delineates the privacy-related alterations made to event data, while refraining from divulging the particulars of privacy preservation tactics or the safeguarded information.

In the healthcare industry, privacy is commonly defined as an individual's entitlement and inclination to manage the dissemination of their personal health data. Patient health information is considered highly confidential due to the potential inclusion of sensitive details pertaining to individual patients. Various reasons may necessitate the implementation of specialised measures to safeguard diverse types of data, such as those related to illnesses and genetic codes. The ability of medical organisations to share their data with statistical professionals is of paramount importance in promoting medical research. The potential cost savings resulting from the aggregation of medical records by individuals could be significant. This gives rise to valid concerns regarding the infringement of personal privacy. The challenge of safeguarding data privacy poses a significant barrier to the advancement of healthcare data analytics. The majority of privacy protection methods involve the manipulation of data or the concealment of information that has the potential to reveal an individual's identity. To effectively accomplish this, it is fundamental to modify either the sensitive property itself or other properties that serve as means of identification. It is evident that during the course of this process, a certain degree of accuracy in the depicted information must be relinquished.

Thus, the preservation of individuals' anonymity frequently entails relinquishing certain valuable data in the course of action. Hence, the aim of privacy preservation methodologies is to attain an optimal equilibrium between ease of use and secrecy. This ensures that minimal utility is compromised while maintaining a specific level of privacy. In broad terms, privacy-preserving data publication algorithms encompass three primary stages: firstly, identifying a suitable privacy metric and level that aligns with the accessibility settings and data attributes; secondly, implementing one or more privacy-oriented algorithms to attain the intended privacy dimensions; and thirdly, performing a post-processing evaluation of the data's utility. By iterating through these three procedures, it can be ensured that both the utility and privacy requirements are met.

IV. CONCLUSION

Healthcare analytics employs both historical and real-time data to improve decision-making and optimise patient outcomes by generating valuable insights. The utilisation of health care analytics can yield two additional benefits for healthcare providers, namely enhanced patient satisfaction and improved health outcomes. Advancements in technology and software have facilitated the acquisition of healthcare data for analysis, thereby driving progress in the field of data analytics. The absence of a pre-defined structure in the data, coupled with the limitations imposed by privacy concerns in data collection and distribution mechanisms, have posed substantial challenges to the advancement of the discipline. The processing and interpretation of information necessitates instantaneous action in the face of extreme data volumes. In certain circumstances, retrieval and analysis of data may require more sophisticated methods due to the complexity of the data. The advent of data collection technologies that enable analytics has introduced fresh challenges despite their efficacy in amassing copious amounts of information. The diverse range of methodologies utilized in the healthcare sector can be attributed to the inherent distinctions in the fundamental data types. The present article has provided a comprehensive overview of healthcare data analytics and highlights several unresolved inquiries that require further exploration.

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. Azhir, M. Hosseinzadeh, F. Khan, and A. Mosavi, “Performance evaluation of query plan recommendation with Apache Hadoop and Apache Spark,” arXiv [cs.DB], 2022.

[2]. “The official InformaticaPowerCenter download resource,” Informatica.com. [Online]. Available: https://www.informatica.com/download.html. [Accessed: 09-Jun-2023].

[3]. C. Wernhard, “Range-restricted interpolation through clausal tableaux,” arXiv [cs.LO], 2023.

[4]. M. Kowalski-Mcgraw et al., “Electronic health records (EHRs) and occupational data: A call for promoting interoperability,” J. Occup. Environ. Med., 2023.

[5]. V. F. Rodrigues, R. da R. Righi, C. A. da Costa, and R. S. Antunes, “Smart hospitals and IoT sensors: Why is QoS essential here?,” J. Sens. Actuator Netw., vol. 11, no. 3, p. 33, 2022.

[6]. J. Iannacci, “Study of the Radio Frequency (RF) performance of a Wafer-Level Package (WLP) with Through Silicon Vias (TSVs) for the integration of RF-MEMS and micromachined waveguides in the context of 5G and Internet of Things (IoT) applications. Part 2: parameterised 3D model and optimisation,” Microsyst. Technol., vol. 27, no. 1, pp. 223–234, 2021.

[7]. M. Chaudhary, N. Goyal, A. Benslimane, L. K. Awasthi, A. Alwadain, and A. Singh, “Underwater wireless sensor networks: Enabling technologies for node deployment and data collection challenges,” IEEE Internet Things J., vol. 10, no. 4, pp. 3500–3524, 2023.

[8]. S. S. Alam, S. B. Shuvo, S. N. Ali, F. Ahmed, A. Chakma, and Y. M. Jang, “Benchmarking deep learning frameworks for automated diagnosis of ocular Toxoplasmosis: A comprehensive approach to classification and segmentation,” arXiv [eess.IV], 2023.

[9]. E. Parcham and M. Fateh, “HybridBranchNet: A novel structure for branch hybrid convolutional neural networks architecture,” Neural Netw., vol. 165, pp. 77–93, 2023.

[10]. C. Reinsch and M. Richter, “Singular value decomposition in extended double precision arithmetic,” Numer. Algorithms, vol. 93, no. 3, pp. 1137–1155, 2023.

[11]. W. P. Richardson, M. L. Reba, and B. R. K. Runkle, “Modification of a wavelet-based method for detecting ebullitive methane fluxes in eddy-covariance observations: Application at two rice fields,” Boundary Layer Meteorol., vol. 184, no. 1, pp. 71–111, 2022.

[12]. Y. Tang et al., “Characterization of Calculus bovis by principal component analysis assisted qHNMR profiling to distinguish nefarious frauds,” J. Pharm. Biomed. Anal., vol. 228, no. 115320, p. 115320, 2023.

[13]. R. A. Baksh et al., “Multiple morbidity across the lifespan in people with Down syndrome or intellectual disabilities: a population-based cohort study using electronic health records,” Lancet Public Health, vol. 8, no. 6, pp. e453–e462, 2023.

[14]. P. Rajagopal, T. Aghris, F.-E. Fettah, and S. D. Ravana, “Clustering of relevant documents based on findability effort in information retrieval,” Int. J. Inf. Retr. Res., vol. 12, no. 1, pp. 1–18, 2023.

[15]. N. U. Patel, B. A. Moore, R. F. Craver, and S. R. Feldman, “Ethical considerations in adherence research,” Patient Prefer. Adherence, vol. 10, pp. 2429–2435, 2016.