Deep Science Publishing
https://doi.org/10.70593/978-93-49910-80-5_7

DeepScience
Open Access Books

Chapter 7: Big data techniques for analyzing patient data collected from medical devices

7.1. Introduction

Big Data refers to the large amount of complex data that emerges in real time and needs a huge amount of storage, computing infrastructure, software as well as expertise to analyze and infer insight. The advent of the intelligent processing devices embedded in the medical devices and sensors has led to a paradigm shift in the healthcare industry. Nowadays, clinical dataset for a patient is available in various forms such as Electronic Health Record, imaging data, external sensor dataset collected from various devices, motel medical devices and genomics. The integration of the patient data collected from different sources is essential for holistic patient management through early and accurate diagnosis, preventive care as well as personalized treatment (Dubey et al., 2017; Salem, 2021; Rauniyar et al., 2022).

However, the large volume and distributed nature of patient data has made this task challenging. The need for new advanced emerging technologies namely wearable monitoring devices and imaging sensors has led to the evolution of the Digital Medicine Era. The main advantage of Digital Medicine is the availability of patient data trials over a period of time, which can be used to create the Individual Patient Digital Twin. Intelligent Data Analytics techniques from the field of Data Science termed as Big Data Techniques are now being heavily utilized in Digital Medicine to drive important decisions and treatment pathways. Machine Learning and Deep Learning are the most popularly used Big Data Techniques to drive insight and predictions from the patient and the disease related data. Where the traditional ML Algorithms most popular among healthcare practitioners include Numerical Prediction Models such as Logistic Regression, Support Vector Machines, k-Nearest Neighbors, Decision Trees, Random Forests. These algorithms have been popular due to their small implementation

overhead, small footprint in terms of memory, CPU cycle and can be run using Open Source platforms easily and efficiently (Tuli et al., 2019; Shaik et al., 2023).

Fig 7.1: Big Data Techniques for Analyzing Patient Data Collected from Medical Devices

7.2. Overview of Big Data in Healthcare

Big data in healthcare refers to vast amounts of diverse, complex, and rapidly growing data, which demand advanced methods and techniques to enable the capture, storage, management, analysis, and visualization of the information. The convergence of the relentless and exponential growth of computing systems operating in clinical, biomedical, and operational domains of the healthcare ecosystem; increasing computing and storage capabilities creating ever-decreasing costs of hardware, software, storage, and network; and the growing adoption of enabling technologies is contributing to the data deluge on the healthcare ecosystem, becoming the perfect storm where technology can advance in otherwise unattainable new areas. The fields of genomics, proteomics, metabolomics, and other omics are heavily contributing to this data explosion.

Multiple types of big data generated by the healthcare ecosystem stakeholders inundate the internet and enterprise data stores. Clinical data such as Electronic Medical Records and Electronic Health Records, where the biometric, clinical, demographic, and historical health information of patients are stored in a digital format; life sciences data such as genomics, transcriptomics, proteomics, epigenomics, metabolomics,

microbiomics, and other omics; Internet of Medical Things-based data generated by Medical Devices and Internet of Things-enabled biometric sensors; imaging data; patient-generated health data, where patients provide and store their biometrics on mobile devices; and public health data such as vital records, diseases registries, syndromic surveillance, and census data, have been contributing to the digital disruption of the healthcare ecosystem, adopting computational technologies. Artificial intelligence, machine learning, big data analytics, deep learning, and statistical modeling have already started or are on the cusp of transforming healthcare systems globally, driven by driving factors.

7.3. Types of Medical Devices

Medical devices are classified based on the time frame of their operation. They are classified into three major types: wearable, implantable, and remote monitoring devices. Wearable devices are used for monitoring a patient for a short to mid-term time duration. They are noninvasive and do not require surgical procedure for wearing them. Normally, they detect and monitor a patient's health condition or disease in an unattended mode. Implantable devices are those which are inserted in a patient's body for a long-term basis (usually more than 30 days). They are implanted in a patient's body through surgical procedures and stay inside the body for days, months, or even years. Implantable medical devices provide continuous monitoring of critical events or health conditions or diseases of patients from within the body. Remote monitoring devices are designed to assess a patient's health condition or disease either locally or from a remote location. However, both the patient and the healthcare professional have to be in contact with each other during monitoring in real-time. The next subsection discusses different types of medical devices in detail.

Wearable devices are battery-powered or energy scavenging systems equipped with sensors. They are designed to be worn on the body for measuring vital signs or other activities. They are available in the formats such as patches, belts, glasses, watches, clothing, and shoes, and can operate for a few days to weeks with frequent replacement of batteries. Wearable devices can be classified as body-worn electronic devices, smart textiles, peripherals, etc. Body-worn electronic devices refer to unobtrusive patches, belts, or smartwatches that are strapped or worn on the skin. Smart textiles are fabric and threaded-based articles that are embedded with electronic textiles. Sensor-based electronic peripherals are shoes or glasses that have sensors for capturing or monitoring a patient's data.

7.3.1. Wearable Devices

Wearable medical devices or simply wearable devices are the type of medical equipment that patients wear and monitor important biomedical parameters. The wearable devices distinctively have two features. First, it can provide timely critical health-related data for analysis. Second, it should not interfere with the normal daily routine of a patient. The wearable devices can be extremely helpful since they can track patient health data continuously. This rapid real-time surveillance is especially advantageous during periods of post-surgery recovery or times of self-isolation. In recent years, smart technology has entered the field of medical monitoring. The development of intuitive smart technology accelerates the integration of a wide variety of wearable devices with ambient telemetry technology at an accelerated pace.

Wearable devices have increasingly become an integral part of the implementation of telehealth services today. This is mainly due to the sudden outbreak of the pandemic and the growth of the global technological market. The demand for low-cost wearable devices for health monitoring and Remote Patient Monitoring applications during the pandemic increased several-fold because the pandemic highlighted the fact that patients with chronic diseases are especially vulnerable to an infection. This led to an increasing demand for advanced and innovative low-cost wearable sensors equipped with technologies that enable continuous assessment and monitoring of physiological signals from patients to facilitate timely intervention. The field of wearable biosensors has continued to grow in recent years. Wearable devices are widely used for vital sign monitoring, respiration monitoring, and wearable ECG monitors. Wearable devices can be worn anywhere and are flexible to wear, they have also assisted in the development of wrist- and clothing-based wearable sensors that are comfortable and fit for all-day use.

7.3.2. Implantable Devices

Implantable medical devices are therapeutic devices that are placed inside the body using surgical techniques. Pacemakers, defibrillators, drug delivery pumps, orthopedic implants, and prosthetic devices are examples of commonly implanted medical devices. The patient data collected from these devices is exercise performance data such as stroke volume, cardiac output, tidal volume, and respiratory rate; activity detector sensor data for motion classification; remote device configuration data available as event logs; biometric data available from embedded sensors; patient-centric data available from questionnaires; data incorporated from external monitoring devices; device-conditioning-assisted and clinical-coded medical records; device-assisted electrogram signal data; and related information resources. The rapid advancement of microelectronic technologies has led to the development of miniature implants that utilize implantable sensors for in vitro and in vivo biological parameter monitoring.

These tiny devices are capable of sensing, recording, and transmitting physiological information via wireless communication protocols.

7.3.3. Remote Monitoring Devices

Remote medical monitoring is becoming part of the standard mode of patient care. Remote monitoring devices, which provide insight into how patients are progressing post intervention, are becoming a regular part of standard operating procedures in hospitals, clinics, and at-home patient care. These devices include things such as implanted and external continuous glucose monitors for diabetic patients, implanted cardiac monitors for arrhythmia detection, implanted hemodynamic monitors following heart failure patients postdischarge, wearable activity monitors, prosthesis-integrated load cells, and medication adherence monitoring devices. These devices are primarily focused on health maintenance. Devices designed for internal surveillance following patient intervention, such as surgical techniques, may also be included in the list of remote medical monitors.

The global trend in healthcare is towards personalized preventive medicine focusing on early disease detection and prevention of morbidity and mortality utilizing risk factors and disease predictors. Remote medical monitoring systems also include algorithms and clinical decision support systems to predict disease processes, prevent patient readmission, avoid patient morbidity, and provide real-time data for healthcare providers to proactively manage the health status of monitored patients. These types of devices usually use internal or external telemetry to transfer observed data from the patient back to monitoring systems managed by healthcare providers for post-intervention morbidity prediction. Telemetry also enables communication from the healthcare provider back to the monitored patient for data sharing and health maintenance reporting.

7.4. Data Collection Methods

Data collection methods have grown considerably over recent years with the increase in the accessibility of open datasets, APIs, and sensor networks. These novel data extraction and analysis techniques enable researchers and industry experts to leverage enormous sensor datasets to extract valuable insights and discover hidden patterns from them. Recently, there has been a lot of effort towards optimizing various methods and frameworks for processing and extracting knowledge from sensor collected data. On the front of biomedical research, researchers have recently turned their attention to developing novel techniques for monitoring, processing, and analyzing patient data that has been streamed or collected from various biomedical sensor devices. It is expected that in the near future, various physiological signals that were previously collected in a

controlled environment would be continuously monitored and analyzed using advanced decision making algorithms.

As discussed earlier, two common methods for data management are typically utilized by researchers. The first method is based on real-time data flow from relevant data sensors or devices continuously. This means that as soon as the data is generated from the sensor devices, it is processed either on-device or in the cloud and ultimately transferred to a main database which consolidates all such data from various connected devices of interested patients. This model resembles what is currently offered by many medical device manufacturers who continually collect health data from their patients for clinical management of patients. The second method is the batch data processing model, whereby several patients are monitored over a period, and their health data is collected periodically and compiled to create a large dataset which is then stored in the database for validation and research purposes. Both methods have their advantages and disadvantages based on various circumstances and available resources.

7.4.1. Real-time Data Streaming

Chapter Summary Big data analytics for medical devices demand capturing a diversity of devices and sources. Current data collection methods impose network architecture, information carrier mechanisms, and transferring protocols restrictions on the kind of data available for further analyzes. It may be an asset for certain small big data classes. Typical practical purposes are the streaming real-time data from the access level to the aggregation and transferring level and the batch transferring from the aggregation to the access level of the IoT network architecture. In this chapter, we describe the practical issues related to the data collection of standard commercial IoT architectures for the big data field. Though out of the typical architecture, we focus on the wireless body area networks example.

Real-time Data Streaming In practice, every standard IoT topology proposes a variety of purposes for both the access sensor level and the transferring connectivity level. The characteristics imposed by sensors and their local contact with the measured subjects and actuators concerning proximity, energy, foreign interference, access, and burst of parameters restrict the range of applicable medical devices, wearable or implantable, to a quality class. Medical devices must operate for long periods of time and make reliable measurements on the wearers. The limited energy resources open big challenges in clustering the sampled information, data fusion techniques, priority data-based transmitting, and the wearable processing capability. These limitations impose the use of proprietary sensors and custom-made firmware that restrict the variety in the measured parameters. Also, the lack of diverse surplus information captured can

compromise the expected quality for medical big data analytics, introducing biases related to the missing not at random mechanism.

Fig 7.2: Real-time Data Streaming

Appropriately chosen and combined, these real-time and reduced data kinds might help monitor particular populations and detect critical states and events, using established wearable or implantable devices. In that direction, aperiodical fast data collection might also exploit the characteristics of periodical parameter variations to jump aperiodical. A compromise between the capturing device quality and the big data purpose must drive the interface design engineering, especially for neural recording devices. These devices face the biggest challenge of the energy resources, needing complicated power transfer fair rules to enable periodic performed protocols.

7.4.2. Batch Data Processing

Very commonly, an initial biomedical analysis is unnecessary, and thus the data remains in the devices for an extended period of time (up to several months) until a specific medical condition requires the physician to extract the contents. In such cases, the analysis is not performed in real time: these are batch analysis cases. Also, when heavy analysis is required, it is very common to execute all of these processing steps at a later time, fully isolated from the patient. For this reason, batch processing is extremely common in the medical area. The batch processing interactions occur between the

biomedical device that collects the information in files and a remote server that handles a multitude of operations on sets of data contiguous in time from different patients.

Within the biomedical field, several batch data processing systems at public level already exist, which allow for certain analyses to be performed on patient samples in a general way. Some of them charge a fee for the use of their data analysis infrastructure and some others are free. The paid systems commonly contain powerful servers connected to extensive net storage and are generally focused on a specific area of application. The free systems are more numerous and diverse, covering multiple application fields, but they require that the researchers develop their own sequence of programs to retrieve and when necessary reformat the input data, calling the specific applications and then collecting and processing the output data on patient scale. In exchange for the effort of building these orchestrated sequences of functions, the researchers of these free systems have made available high-throughput specific applications operating on powerful servers.

7.5. Data Storage Solutions

As shown in the fourth phase in our roadmap is data storage. After the processing steps, the data is ready for analytic efforts during the exploration and analytic phases, during which the data will be queried multiple times. Typically a data storage system will be utilized during the entire exploration and analytic phases. Many factors factor into the decision of what data storage solution to utilize including project timeline, technical capabilities of the team, size and complexity of data, required performance level, and cost. These considerations will guide the decision toward options being potentially stored in cloud services, utilizing on-premises servers, or being implemented with hybrid options.

There are many cloud storage services popular to use for big data projects. One provider offers a collection of services to aid in storage, processing, and frameworks specifically designed to work with big data. Another offers a relatively new player on the market that focuses on being an analytical database and utilizes the cloud for storage and compute operations. A third provides similar storage capabilities and functions for handling big data through its cloud services. These cloud adoption options provide flexibility in the amount of resources provisioned and are ideal for cases where the project team is unprepared to invest in physical infrastructure like rack servers with scalable storage. One key challenge with the cloud option is related to keeping costs low, especially in cases where long-term permanent storage is needed and retrieval of previous data values is not frequent and sudden spikes of traffic cannot be forecasted. However, many companies utilize these services for their lower barrier of entry cost based on limited project sizes and budgets.

7.5.1. Cloud Storage

In our information-driven age, healthcare organizations are continually challenged to find ways to maintain operational efficiencies while keeping patient care costs in check. One viable solution to achieving this operational health is the adoption of cloud storage solutions – shared resource pools of data storage made available by third party providers. Moving data to the cloud offers organizations viability in both flexibility and cost, as resources can be expanded on-demand and there are no hefty upfront hardware costs. Because user sessions as well as storage demand can be intermittent throughout the office hours, the pay-as-you-go pricing feature associated with cloud virtual storage is a financially appealing option. Organizations are also able to obviate the burden of dedicated, on-site IT resources who must implement and perform routine maintenance on an in-house storage solution. Despite these advantages, the move to cloud data storage is not without risks. Healthcare patient data is now increasingly regulated through compliance measures which monitor sensitive data access and use within the healthcare service paradigm.

7.5.2. On-Premises Solutions

The term "on-premises" refers to an enterprise-class IT solution that is hosted and managed within the physical confines of the enterprise infrastructure. The company relies on its own resources to install the software in its own data center or server room. Although many IT resources today have taken on the subscription-based, utility pricing model of the public cloud, a significant number of software applications still run on-premises, including many legacy enterprise resource planning systems. For decades, the majority of organizations operated under an on-premises IT structure. Thousands of businesses, from small shops in suburban office complexes to Fortune 500 giants with global operations, built out data center racks filled with servers, routers and switches. A band of technicians was devoted to keeping the systems up and running. With the rapid rise of cloud services, this model has largely faded into the background.

Meanwhile, organizations also enhance on-premises IT with other types of technology. They often utilize colocation services to reduce the cost and hassle of operating their own data centers but still prefer to manage their own servers. Many organizations utilize small cloud infrastructures that connect to a larger external cloud. These are often put into place to provide backup or additional processing capacity for the larger cloud infrastructure. On-premises IT is expected to continue as the dominant model for certain enterprise applications and specialized workloads. Security and privacy regulations are the most common reasons cited for requiring on-premises IT. Organizations that manage sensitive or personal information, such as banking, health care and law firms often prefer to keep everything in-house rather than place that information into a cloud environment

that might be vulnerable to security breaches. Other organizations might have custom-built applications that are deployed on-premises and support crucial business operations. Because these applications are highly customized, moving them to the public cloud would be too costly and time-consuming.

7.5.3. Hybrid Approaches

Hybrid solutions, involving a combination of on-premises and cloud-based components, are becoming increasingly popular. Such models optimize price while meeting security and compliance constraints. DICOMWeb is a DICOM image storage standard that allows images to be transferred via a RESTful interface, enabling access to cloud storage vendors that support simple HTTP calls to upload and fetch images. Various PACS vendors leverage this API to transfer and archive studies into cloud storage to reduce costs of storing medical images. Hybrid cloud solutions allow radiologists to receive studies in a familiar interface while shifting storage costs to the cloud vendor.

Iguana is a radiology workflow product that runs in a hospital's network but allows users to call the DICOMWeb API to archive studies to the cloud. Also, several medical device vendors use cloud storage to enable on-site radiologists to review the data via a web-based viewer. These storage solutions leverage medical imaging ecosystem products to enhance their image transfer and display performance. Hybrid cloud solutions that remain within a vendor's cloud ecosystem can minimize security concerns and allow for a reduced infrastructure support load. Such support load reductions enable clinical staff to focus on customer service.

7.6. Data Preprocessing Techniques

In the previous sections, we discussed how to pre-process patient data from wearable sensors. Typically, this patient data initially undergoes feature-based processing techniques that segment the raw sensor data into labeled segments. This includes feature selection, data cleaning, data normalization, and data transformation. Then, using this labeled output, a classifier can be trained either to classify patients or to identify important segments of given patients in the data. The quality of the output of the feature-based processing technique is important because a useful classifier can be generated only if high-quality labeled input data is provided for training. This training can be supervised with known segment labels or unsupervised with assumed segment labels. Suppose there are errors in the labels of the training data. In that case, either the resulting classifier will be inaccurate or the trained classifier will not provide appropriate segments corresponding to various states of the patient. Feature extraction is critical for the effectiveness of machine learning. Manual feature extraction typically relies on domain

expertise to develop application-specific features. Machine learning can eliminate the design of many manual feature extractions using raw data for training. In the end, the optimal features to extract from the data depend on the task. Given that our goal is to preprocess the data to develop input for supervised learning strategies, we concentrate on feature-based data preprocessing methods. The goal of data cleaning, normalizing, and transforming the data is to ensure data of high quality as input to the classifiers. The data cleansed should be free of duplicates and outliers. The data features should also be in the same range as others to avoid introducing biases to the classifier. Popular normalization procedures include z-score normalization, min-max normalization, and robust normalization. The raw data need to be transformed into more meaningful values or formats to help the classifier performance better. For example, a 2D input array of data could be concatenated to form a 1D vector for classifier training.

7.6.1. Data Cleaning

Data cleaning is one of the crucial steps in big data preprocessing techniques. Every dataset has some noise and errors in it; these may arise from different sources and can lead to the unreliability of data and also to misinterpretations. Removing such noise is also essential to get accurate relationships and correlations in the data.

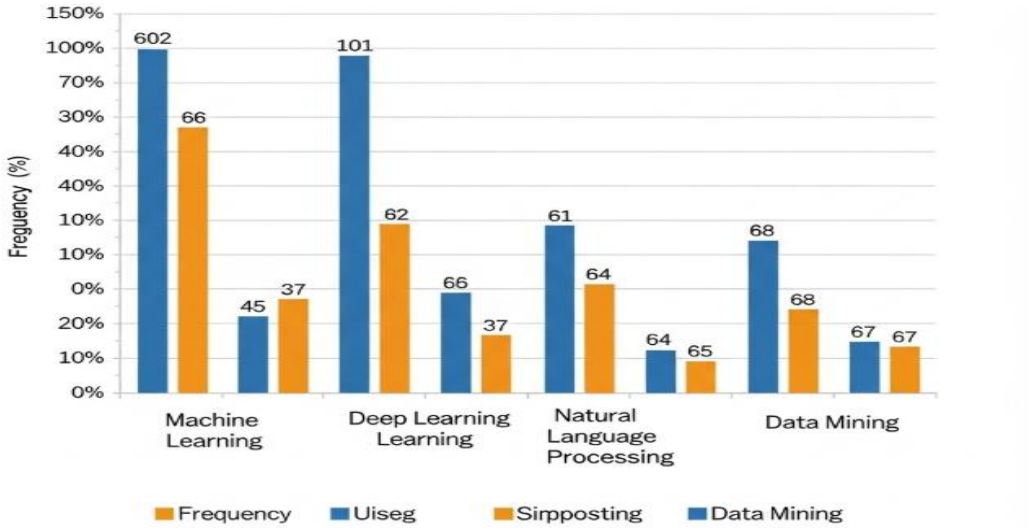

Fig 7.3: Analyzing Patient Data Collected from Medical Devices

It has been observed that 30% to 90% of data cleaning efforts occupy most of the time in the data preprocessing phase of big data. Data cleaning can be an expensive or time-consuming or hard task to execute, especially for extremely large scale and disparate datasets which gives augmentation to most commercial data warehouses and big data

tools. Although some amount of cleaning may take place before data is copied through ETL programs, some cleaning will definitely need to be undertaken after the data is copied to the data warehouse or Big Data for analysis purposes.

Cleaning of big data normally requires automated cleaning processes for data cleansing. One of the most significant processes used for removing noise from big data is De-duplication which can dynamically and aggressively detect duplicate data items using the programming model and is implemented together with certain database system products. This approach uniquely identifies and maps each item to a bucket that contains its specific version; any duplicates are simply excluded at the time of data insertion. DBMSs that use such a logic will include certain database systems. Notably, one system relies on customers for the deduplication step by requiring them to develop a patient and noticed process.

7.6.2. Data Normalization

Data normalization is a process that can improve the performance of machine learning algorithms. It is a standardization technique that works to achieve the same level of automation in medical devices without carrying out needless feature engineering. Essentially, data normalization works to rescale the data to be in a specific range to avoid exploding gradients and to have each feature contribute equally to the loss function of the neural network model. In addition to avoiding exploding gradients, normalizing the medical device data also grants a comparative analysis of the features collected from the devices. This comparative analysis can also benefit the model learning the data and improve accuracy.

Data normalization techniques also can work on time series data. One such method is called z-score normalization which is implemented by simply subtracting each data point by the mean of the entire data sequence and then dividing the resulting number by the standard deviation. This method adds no constraints to the temporal data and can be used when there are outliers present in the sequences. Other normalization techniques that do add constraints to the data are min-max normalization which forces the data to be within a range of −1 to 1 and scaling which restricts the data points of a feature to have a unit norm with unit variance. Other normalization methods include log normalization, mean normalization, and decimal scaling.

Modeling medical device data has its challenges. Fortunately, different normalization techniques can alleviate these challenges. This will aid in a better performance of models used in further analysis of the patient data collected from medical devices without needing elaborate feature engineering strategies. Overall, the importance of

normalization techniques for microcontroller-based circulatory monitoring devices cannot be stressed enough in both the technical area and the healthcare domain.

7.6.3. Data Transformation

There are two data transformation techniques available, PCA and ICA. PCA is a linear method of analyzing multidimensional data and it is mainly used for dimensionality reduction. It is a common choice for visualization, allowing further analysis and interpretation of large-dimensional biomedical datasets. Recently, some new PCA-based techniques have been proposed. However, the basic PCA technique is a linear inductive approach, incorporating the concept of finding a linear transformation that represents the data in the space with maximum variance. It is similar to other dimensionality reduction techniques, like Fisher linear discriminant analysis that classes the multidimensional data into two or more sample classes. The implementation of PCA is easy and the discovery results are easy to understand.

Any multidimensional datasets frequently include unique and interesting features that are hidden in the multidimensional space. Furthermore, ICA is a statistical and computational method to perform blind source separation of a set of signal sources from the signaled sources. ICA uses a dimension-reducing method to find a representation of a group of random variables into a smaller set of non redundant variables. Many ICA algorithms are open source and available. ICA can analyze any mixed multidimensional datasets. Applying ICA for dimension reduction corresponding to the temporal and functional characteristics distinguishes the biomedical signals from the rest of the undesirable signals.

7.7. Data Analysis Techniques

Data analysis is the etymological descendent of data through Latin and Old French. The term to analyze (analysis) comes from the Greek, meaning “to loosen” or “to open”. These facts, coupled with an intelligence definition — “Intelligence means educating and training a person to be able to think and analyze as much as possible” — help us to realize that it is not enough to just apply techniques to analyze data. We should educate people to think as intelligence training enables people to extract real knowledge. Data collection techniques are used to analyze data based on what, how, where, and when to collect data. The type of data collection depends mainly on the objectives of the analysis. The selection of a particular technique is not simple and sometimes its success can be obtained through testing many techniques. Data analysis techniques can be manually executed — in smaller problems — or can be automatically executed. Data analysis consists of the application of different techniques to data with the objective of drawing

conclusions and/or estimating parameters. In the case of hidden data, inference is performed.

Our goal with the following analysis is to present data analysis techniques that can be applied to patient data, with a focus on data from medical devices. We first review statistical analysis methods that can be applied to sensor data. Next, we present machine learning algorithms, both supervised and unsupervised. Finally, we review natural language processing techniques, which are important in the case of text data generated by sensors.

7.7.1. Statistical Analysis

Regular analysis of the biomedical signals recorded from a patient is very useful since they may signify a change in physiological state. The continuous tracking of these signals can be exploited to gather several insightful messages to assist healthcare professionals in decision making and improve the quality of life of patients. The biomedical signal processing steps typically involve filtering, sketching, annotating, and symbolization of the raw data before it may be analyzed for a decision. We first present an advanced and automated approach by utilizing prediction intervals of a fitted statistical model to explore critical patterns in the recurrent measures of phospholipids from patients. In another approach, we utilized Bayesian Regression and Forecasting Model to explore the repetitive biochemical measurements for Thyroid, Alkaline Phosphatase and Hemo-Gram for patients with Type 2 Diabetes Mellitus.

The use of this approach was motivated by the practical difficulties typically faced using time series models to accommodate for missing/overlapping data. This enabled us to analyze across a group of patients and select some patients for whom the predictions are extremely high or actually very low and later validate with claim records. One of the key features of this methodology is that it uses standard multiple linear regression techniques available in most statistical packages and doesn't require specialized expertise in time series models to implement. In addition, the method is flexible enough to accommodate a wide variety of experimental designs with missing data and allows regression modeling of experimental conditions. This work highlights the use of statistical modeling techniques as valuable research tools in exploring relevant patterns in such recurrent and often unbalanced clinical data.

7.7.2. Machine Learning Algorithms

Advances in the field of Artificial Intelligence (AI) and Machine Learning (ML) have influenced almost every area of science and technology. Among the various ML

algorithms that have gained popularity in recent years are Support Vector Machines (SVMs), Artificial Neural Networks (ANNs), Recommender Systems (RS), and Reinforcement Learning (RL). Support Vector Machines (SVMs) are well known for their simplicity, strong theoretical foundation, and good generalization performance. The advantage of the SVM algorithm is that it is able to construct high dimensional borders up to a few support vectors in the data set. Finding the optimal hyperplane for a two-class classification problem in Support Vector Machines in either the primal or dual space is guaranteed to reach the global optimum when the kernel function is convex and positive semi-definite. A well-known disadvantage of the SVM algorithm is the long training time, especially when there are a large number of classes in the data set.

Artificial Neural Network (ANN) is an ML algorithm that has been applied to various fields, including finance, marketing, medical diagnosis, and tourism. The Neural Network approach is a system of computing that is inspired by the biological structure of the brain. The strength of the Neural Network model comes from the fact that it can model nonlinear relationships and large datasets. A disadvantage of the ANN method is that it requires a large number of samples to build an accurate model. With concerns about privacy issues and missing or incomplete data, collecting large amounts of data is not always feasible.

In Recommender Systems (RS), the central idea is to use preferences in order to help end users make informed decisions when choosing information, products, or services. One of the primary drivers of the popularity of recommender systems is the social and economic impact of their use, especially in the areas of e-commerce, online entertainment, social networks, and personalized content delivery. Reinforcement Learning (RL) is an area of machine learning that is generally concerned with how agents ought to take actions in an environment in order to maximize rewards. Many problems in robotics, control theory, operations research, economics, and game theory involve learning how to take actions so as to maximize a numerical reward signal.

7.7.3. Natural Language Processing

Natural Language Processing (NLP) provides techniques for the analysis and representation of natural language text, and it is a branch of Artificial Intelligence (AI). The term representation refers both to the mental faculties and to the issues of logical representation; it deals with the different methods to encipher and decode your memories and perception of the world, your maps in the brain, which can be created and managed by a computer working with AI. As such, NLP provides mechanisms for the analysis of free text with the goal of uncovering patterns in the data. These patterns then represent hidden knowledge that will be of interest to any investigator and will improve the understanding of some phenomena that deal with free textual communications by people.

NLP has great advantages in biomedical research. In the last decade, NLP has been focused on improving its microbial techniques for the understanding of articles in the biomedical field. Modern biomedical NLP must deal with the issue of the generalization of the rules needed to segment a specific field of application. In fact, an expert has to code these field rules to improve the precision in a specific domain. Despite these technical problems, it is necessary to note that NLP can help any area of medicine in the task of transforming natural language to hexadecimal encoding. In addition, NLP can ease the process of defining and distinguishing between specific medical terms that need special attention and also information and arguments and what can be ignored.

7.8. Data Visualization Tools

Data visualization tools are essential components for visualizing big data collected from medical devices by providing colorful and interactive views of the underlying data. Extensive research has been done to create new algorithms and techniques for static and dynamic visualization of medical data. Static visualization of medical images and shapes has been made possible using 3D visualization devices, taking advantage of the latest computer graphics techniques. Further, virtual reality environments have been explored allowing immersed interaction with the data and more natural displays present the data in 3D. Interactive visualization tools such as 2D and 3D; adding dimension tools, and 3D slice processing tools make it easy to visualize big data from medical devices. Scientific visualizations such as Ellipse-formed voxelization for fast full-view 3D printing visualization of 3D computed tomography scans have also been reported.

Dashboards

Medical data dashboards are user-friendly analysis tools that allow end users to quickly and easily explore huge quantities of multi-dimensional patient data. Dashboards are commonly used in many different situations. For example, the rate and severity of outages can be readily summarized in a dashboard containing just 10 numbers or less; however, analyzing the factors that give rise to these outages often requires a much more sophisticated set of data visualizations. Dashboards that provide analytical summaries are known as Reporting Dashboards, while those that enable ad hoc analysis users are known as Exploratory Dashboards. While reporting dashboards contain static visualizations such as ideal for monitoring sales performance and exploring sales relationships, exploratory dashboards provide user-selectable filters and dynamic visualizations, allowing the user to drill down, slice, and dice the data using different selection criteria, such as defining customer segments, and product groups. A dashboard's interactive visualizations may be driven by calls or statements that share common data model definitions. User interaction typically allows data stored in servers

to drive ad hoc analysis but can also be directly stored in the dashboarding hardware or software.

7.8.1. Dashboards

Dashboards are very popular tools for business users, and are becoming increasingly common in healthcare, for patients, doctors, and other healthcare stakeholders to have a visual representation of what is happening with various components. Dashboards in healthcare could show the number of patients currently being treated with Covid-19 and if they are required to be admitted to a hospital. It could show the number of deaths due to Covid-19. It could also show how many patients are being treated from other diseases such as diabetes, cardiovascular diseases, and types of cancer. When informed with this rate, healthcare authorities, governments, and hospitals can make informed decisions. Are the number of admissions to the hospital from Covid-19 increasing? Do we need to revise our guidelines and policies to ensure that the number of people not following Covid-19 protocols is reduced? Such dashboards made for healthcare stakeholders only provide a view of the situation in healthcare, are not interactive, do not allow any drill-down into data available, or do not have any storytelling capabilities.

In summary, dashboards made for non-technical and non-experienced users such as the general public, patients, doctors, insurance companies, and even administrative staff in hospitals provide the most important data usually asked by them which is on display, or visualized. This visualized data gives insights to these users about healthcare that helps in fast decision-making. Dashboards are made by using specialized tools and display visuals that are not interactive. Dashboards are made on some common business intelligence tools that either extract from databases or from underlying data visualized in some other data visualization tools, and run with functionalities to customize the look and feel of dashboards.

7.8.2. Interactive Visualizations

Rather than visualizations changing as the user interactively explores the data, the interactive visualization framework described in this section works by letting the user modify which variables are visualized such that the visualizations update to explore the new variables. This approach is better than having visualizations changing while the user explores the data because instantaneously changing visualizations would not allow the users to see the dependencies between variables or the different properties of the data, nor would it let them view and search for relationships that are nearly identical. Since the majority of knowledge discovery tasks can be reduced to finding relationships between a small set of variables, the interactive visualization approach is appropriate for

visualizing any kind of data. More importantly, when generating a large number of visualizations, there is a need to keep the number of visualizations small. Several of the existing approaches visualize the distributions for several groups in a single visualization in the same frame, which requires heavy luminosity modeling or luminance correction. Red color blindness affects a large group of users, and to support this user group, heat maps and other methods of encoding complex data in the luminance dimension should only be used when it is not necessary to provide an effective visualization for these users. Allowing users to select which groups multiplexed in a single visualization are visualized is helpful, but does not completely address the need to keep the number of visualizations small. For example, the groups correspond to the values of the grouping attribute, and the methods let users select which attributes are used to create the visualization. A simple method of creating such attribute visualizations is to create small multiples. These launch several visualizations at once.

7.9. Challenges in Analyzing Patient Data

There are several challenges in the method outlined in this chapter in the context of analyzing patient data. Machine learning has made great strides in various domains based on visual, speech, and textual analysis due to large amounts of easily accessible labeled data. However, such approaches for analyzing patient data could have different levels of success compared to conventional medical approaches due to lesser amounts of real-world labeled patient data. Although generating synthetic data based on generative models has made some progress in dealing with such reduced data challenges, this is still an area of preliminary research. There are some specialized data augmentation techniques for specific modalities which can partially help mitigate this problem. In addition to reduced amounts of labeled patient data, a second key challenge is the impact of data privacy-related regulations on the quality of patient data which can be made publicly available, thereby again limiting the data size.

There have also been calls to come up with techniques that incorporate prior medical knowledge into the analysis of patient data; either related to specific modalities or conditions being treated in order to assist in classification or prediction. Due to the necessity of analyzing and interpreting data obtained from heterogeneous data collection sources and levels of data available about individual patients due to various data privacy-related regulations limiting the usage of such data, ideally multi-modal, multi-source and multi-level data, some have called for a multi-parametric patient data-centric approach that aims to use complementary data in a multi-parametric approach in order to enhance the quality of prediction for individual patients. At the same time, we note that such data-centric methods will also be impacted by the concerns due to data privacy regulations discussed previously due to these approaches generating exposure to more significant

amounts of details related to the health history of individual patients during the prediction or classification task.

7.9.1. Data Privacy Concerns

One primary challenge in analyzing patient data collected from medical devices is concerned with the privacy of patient information. Studies involving medical device data must adhere to the strict rules enforced by institutions. While traditional security and data protection techniques may be effective at anonymizing patient data, the volume and variety of data produced by devices allows for a re-identification risk, mandating the need for a different strategy to ensure protection of patient identity. Quasi identifiers, regularly found in sensitive data, can often help facilitate the identification and re-identification of patients.

Quasi identifiers include age, gender, date of birth, location, and biometrics. Because data from a patient's wearable device can help track their movements and daily activities, the data is highly sensitive and can jeopardize the safety of patients. For this reason, it would be wise to investigate the data privacy challenge posed by the patient data being collected. With movement data readily available in a communication packet, utilizing techniques developed in the field of data inoculation, data masking, and perturbation to provide differential privacy guarantees to the data before transmitting or sharing would help reduce the risk of patient identity exposure. However, while such techniques are valuable at providing identity protection, patients must also be aware of the risks involved in opting into wearable device projects to help further research in medical device informatics.

7.9.2. Data Integration Issues

Integrating and managing the types of complex and heterogeneous data from various medical devices, sources and formats can also be very challenging. Clinical data may include, but is not limited to, medical device data, physician notes, pathology results, laboratory results, pharmacy records, other clinical test results, and medical histories, all of which are usually stored in different formats. Data from these diverse sources, however, require different cleaning, transformation, and integration pipelines in order to be integrated and used for any downstream predictive analytics tasks.

Integrating or joining each of the heterogeneous modalities may result in considerable information loss, especially in the early life stages of patients or during relapses when relatively few events have been recorded within that specific modality or sub-modality. Integrating these modalities remains overly too demanding since appropriate labels are

not available. We may only be able to label whether or not a patient has any events or transitions within that time period without knowing which of the events occurred during that time. We could use a multi-instance learning paradigm, but that incurs a relatively high and unwanted computational cost. Most of the existing multi-disciplinary works focus on either single-level data fusion or data fusion at the feature level.

7.9.3. Scalability Challenges

Scalability is an important challenge when it comes to processing complex data in the context of big data for healthcare. The terabytes of patient data generated by low-cost medical devices, while particularly helpful in studying the time-variant length and nature of certain health conditions of patients over a massive amount of time, are not easy to store and process for analytical tasks. The fact that devices like magnetocardiography can continuously monitor patients for days together and produce hundreds of millions of matrices representing magnetic field measurements is ample proof of the scalability issues relating to efficient and timely storage and processing of such vast amounts of health data being generated by low-cost health devices. Moreover, running analytics at scale also requires building scalable algorithms that can be parallelized and run on multi-core and distributed environments. Most of the state-of-the-art algorithms for analytical tasks on longitudinal patient data have been developed in the context of clinical or experimental data which while possibly large and verbose, are not Big Data by any means. A majority of the scalable algorithms that attempt to address the volume aspect of health Big Data either rely on simplified heuristics lacking in refinement or focus on specific tasks within one or a combination of tasks without any thorough consideration.

Most of the scalability issues of the state-of-the-art algorithms for analyzing the massive amount of longitudinal patient data generated by low-cost medical devices stem from the fact that the algorithms primarily deal with clinical experimental data, which while verbose, are not Big Data by any stretch of the imagination. Moreover, with medical devices expected to keep evolving in terms of complexity and volume of data being generated, there is a dire need for developing data-driven and domain-aware yet computationally efficient analytical algorithms, to make intelligent use of the serviceable medical device signals before they suffer from the effects of information overload.

7.10. Conclusion

Numerous challenges exist for healthcare stakeholders when working with patient data collected from medical devices but these challenges also represent numerous opportunities. Numerous techniques have been proposed in the Big Data community for properly managing such data, while others have been proposed for deriving insight from

such data and yet other methods utilize such data for training machine learning algorithms. With the proliferation of patient-generated data and the ever-increasing connectivity of biomedical devices, the possibilities seem to be limitless. Moreover, the value of such data is great since these data represent vast real-world patient populations rarely represented in traditional clinical trials. Personalized healthcare seems like an achievable goal thanks in part to these advancements in patient data science. Furthermore, traditional research in biomedical fields exploring treatment efficacy and post-market surveillance after approval might start looking for guidance in such large amounts of patient data, which would enable quicker results.

Utilizing Big Data techniques for patient data collected from medical devices and improving upon them is an exciting research direction. Various repositories of biomedical device data including registry data can help in the validation of Big Data methods and techniques. By properly stressing the techniques proposed for other domains and building from them, we believe that they can be successfully employed in the domain of patient device data science. In particular, we see avenues for progress and development in the domain of novel statistical techniques and visualization techniques. Various novel statistical techniques might help overcome challenges in missing data, unbalanced data, or generating generalized results from cohort studies of small patient populations. Visualization techniques might also provide unique opportunities when faced with overlaid data where overlapping points occur enormously. Furthermore, we see opportunities for progress in the various lessons previously described, especially in how other fields manage, process, and glean insight from data on a huge scale.

7.10.1. Future Trends

The growing amount of patient data collected in health and wellness issues provides physicians and healthcare administrators with useful information about patient conditions to improve diagnosis, care, and treatment procedures. Research data are also more and more available, with the aim of designing and assessing novel solutions for detecting and addressing patient problems. For specific ailments/conditions, collaborative research projects exploit patient-generated health data such as those acquired by mobile health sensors, including wearables and implanted devices. In particular, these devices gather facilitated access to vast amounts of continuously witnessed information about patients' physiological parameters. However, analyzing this data to obtain actionable, personalized knowledge is not simple. To contribute to the healthcare knowledge domain, this work provides a compilation of techniques for effectively analyzing medical sensor data.

References

Shaik, T., Tao, X., Higgins, N., Li, L., Gururajan, R., Zhou, X., & Acharya, U. R. (2023). Remote patient monitoring using artificial intelligence: Current state, applications, and challenges. arXiv preprint arXiv:2301.10009. arXiv

Tuli, S., Basumatary, N., Gill, S. S., Kahani, M., Arya, R. C., Wander, G. S., & Buyya, R. (2019). HealthFog: An ensemble deep learning based smart healthcare system for automatic diagnosis of heart diseases in integrated IoT and fog computing environments. arXiv preprint arXiv:1911.06633. arXiv

Dubey, H., Monteiro, A., Constant, N., Abtahi, M., Borthakur, D., Mahler, L., ... & Mankodiya, K. (2017). Fog computing in medical Internet-of-Things: Architecture, implementation, and applications. arXiv preprint arXiv:1706.08012. arXiv

Rauniyar, A., Hagos, D. H., Jha, D., Håkegård, J. E., Bagci, U., Rawat, D. B., & Vlassov, V. (2022). Federated learning for medical applications: A taxonomy, current trends, challenges, and future research directions. arXiv preprint arXiv:2208.03392. arXiv

Salem, A.-B. (2021). Innovative Smart Healthcare and Bio-Medical Systems: AI, Intelligent Computing and Connected Technologies. CRC Press. Routledge