Neural Computing and Applications (2020) 32:4417–4451
https://doi.org/10.1007/s00521-019-04095-y

COGNITIVE COMPUTING FOR INTELLIGENT APPLICATION AND SERVICE

Big data analytics for preventive medicine

Muhammad Imran Razzak[1] · Muhammad Imran[2] · Guandong Xu[1]

Received: 1 October 2018 / Accepted: 12 February 2019 / Published online: 16 March 2019

Abstract
Medical data is one of the most rewarding and yet most complicated data to analyze. How can healthcare providers use modern data analytics tools and technologies to analyze and create value from complex data? Data analytics, with its promise to efficiently discover valuable pattern by analyzing large amount of unstructured, heterogeneous, non-standard and incomplete healthcare data. It does not only forecast but also helps in decision making and is increasingly noticed as breakthrough in ongoing advancement with the goal is to improve the quality of patient care and reduces the healthcare cost. The aim of this study is to provide a comprehensive and structured overview of extensive research on the advancement of data analytics methods for disease prevention. This review first introduces disease prevention and its challenges followed by traditional prevention methodologies. We summarize state-of-the-art data analytics algorithms used for classification of disease, clustering (unusually high incidence of a particular disease), anomalies detection (detection of disease) and association as well as their respective advantages, drawbacks and guidelines for selection of specific model followed by discussion on recent development and successful application of disease prevention methods. The article concludes with open research challenges and recommendations.

Keywords Disease prevention · Data analytics · Healthcare · Knowledge discovery · Prevention methodologies

1 Introduction

Due to the rise of healthcare expenditures, early disease prevention has never been important as it is today. This is particularly due to the increased threats of new disease variants, bio-terrorism as well as recent improvement development in data collection and computing technology. Increase amount of healthcare data increases the demand to develop an efficient, sensitive and cost-effective solution for disease prevention. Traditional preventive measures mainly focus on promotion of healthcare benefits and have lack of methods to process huge amount of data. Using IT to promote healthcare quality can serve to improve health promotion and disease prevention. It is true inter-disciplinary challenge that requires number of types of expertise in different research areas and really big data. It raises some fundamental questions.

- How do we reduce the increasing number of patients through effective disease prevention?
- How do we cure or slow down the disease progression.
- How do we reduce the healthcare cost by providing quality care?
- How do we maximize the role of IT in identifying and curing the risk at early stage?

Clear answer to these question is the use of intelligent data analytics methods to find information from glut of healthcare data. Data analytics researchers are poised to come up with huge beneficial advancement in patient care. There is vast potential for data analytics applications in healthcare sector. Currently, data analytics, machine learning and data mining made it possible for early disease identification and treatment. Early monitoring and detection of disease being in practice in many countries, i.e., BioSense (USA), CDPAC (Canada), SAMSS, AIHW (Australia), SentiWeb (France) etc.

✉ Muhammad Imran
dr.m.imran@ieee.org

Guandong Xu
guandong.xu@uts.edu.au

[1] Advanced Analytics Institute, University of Technology, Sydney, Australia

[2] College of Applied Computer Science, King Saud University, Riyadh, Saudi Arabia

This paper discusses the IT-based methods for disease prevention. We chose to focus on data-mining-based prevention methodologies because recent development in data mining approaches led the researcher to develop number of prevention systems. Tremendous progress has been made for early disease identification and its complication management.

1.1 What is data mining and data analytics

Exponential time increase in data made tough to get useful information form that data. Traditional methods showed much performance; however, their predictive power is limited as traditional analysis deals only with primary analysis, whereas data analytics deals with secondary analysis. Data mining is the digging or mining of data from many dimensions or perspectives through data analysis tools to find prior unknown pattern and relationship in data that may be used as valid information; moreover, it makes the use of this extracted information to build predictive model. It has been used intensively and extensively by many organizations especially in healthcare sector.

Data mining is not a magic wand but in fact a big giant tool that does not discover solutions without guidance. Data mining is useful for the following purposes:

- Exploratory analysis: Examining the data to summarize its main characteristics.
- Descriptive modeling: Partitioning of the data into subgroups based on its properties.
- Predictive modeling: Forecasting information form existing data.
- Discovering pattern: Discover pattern that occur frequently.
- Retrieval by content: Discovering hidden patterns

Big data and machine learning holds great potential for Healthcare providers to systematically use data and analytics to discover interesting pattern that are previously unknown and uncover the inefficiencies from vast data stores in order to build predictive models for best practices that improve quality of healthcare as well as reduces the cost. EHR system are producing huge amount of data on daily basis which is a rich source of information that can be used by healthcare organization to explore the interesting fact and findings that can help to improve patient care. Figure 1 shows the data analytics generic architecture for healthcare applications.

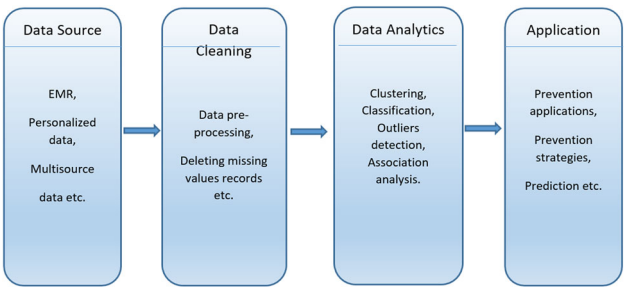

Fig. 1 Architecture of health care data analytics

As health sector data is moving toward really big data, thus better tools and techniques are required as compared to traditional data analytics tools. Traditional analytics tools are user friendly and transparent as compared to big data analytics tools that are complex and programming intensive and required variety of skills. Some famous big data analytics tools are summarized in Table 1.

1.2 What is disease prevention and its challenges

Every year millions of people die of preventable death [1]. In 2012, about 56 million people died worldwide and two-thirds of these deaths were due to non-communicable disease including diabetes, cardiovascular and cancer. Moreover, 5.9 million children died in 2015 before reaching the fifth year of their life and most of these death were due to infection (i.e., diarrhea, malaria, birth asphyxia, pneumonia etc.); however, this number can be reduced to half at least by treating or preventing through the access to simple affordable interventions [2]. Core problem in healthcare sector is to overcome the huge number of causalities as well as reduce the cost. The goal is to reduce the prevalence of disease, help people to live longer and healthier life as well as reduce the cost. One of the main interests in disease prevention is driven by the need to reduce the cost. The lifetime medical expenditures are increased from $ 14k to $ 83k per person and this increase is up to 160K after the age of 65.

Thus, the proportion of average world GDP devoted to healthcare sector is increased from 5.28% in 1995 to 5.99% in 2014 and is expected to increase in future (i.e., from 17.1% in 2014 to 19.9 of US GDP by 2022%) [3, 4]. This increase in medical expenditures is mainly due to the aging and growing populations, the rising prevalence of chronic diseases as well as for infrastructure improvement. Thus, the cost-saving and cost-effective preventive solutions are required to reduce the burden on economy. Traditional preventive measures mainly focus on promotion of healthcare benefits. The cost-effectiveness ratio is said to be unfavorable when intervention incremental cost are larges relative to the healthcare benefits. USA spent 90% of budget on disease treatment and their complication rather than prevention (only 2–3%) whereas many of these diseases can be prevented at first stage [5, 6]. Spending more on health does not guarantee of health system efficiency. The investment on prevention can help to reduce the cost as

Table 1 Big data analytics tools

Platforms and tools	Description
Advanced data visualization	ADV can reduce quality problems which can occur when retrieving medical data for extra analysis
Presto	Distributed SQL query engine used to analyze huge amount of data that collected every single day
The Hadoop Distributed File System (HDFS)	HDFS enables the underlying storage for the Hadoop cluster and enhances healthcare data analytics system by dividing large amount of data into smaller one and distributed it across various servers/nodes
MapReduce	Breaks task into subtasks and gathering its outputs and efficient for large amount of data
Mahout	An apache project, goal is to generate free applications of distributed and scalable ML algorithms that supports healthcare data analytics on Hadoop systems
Jaql	Functional, declarative query language, aim to process large datasets. It facilitates parallel processing by converting high-level queries into low-level ones
PIG and PIG Latin	Configured to assimilate all types of data (structured/unstructured, etc.)
Avro	Facilitates data encoding and serialization that improves data structure by specifying data types, meaning and scheme
Zookeeper	Allows a centralized infrastructure with various services, providing synchronization across a cluster of servers
Hive	Hive is a run-time Hadoop support architecture that permits to develop Hive Query Language (HQL) statements akin to typical SQL statements

well as improve the health quality and efficiency. Health industry is facing considerable challenges in the promotion and protection of health at a time when there is huge pressure due to the considerable budgets constraints and resources in many countries. Early detection and prevention of disease plays a very important role in reducing deaths as well as healthcare cost. Thus, the core question is: How data can help to reduce the patients or disease effect in the population?

1.2.1 Concept and traditional methodology

Disease prevention focuses on prevention strategies to minimize the future hazards to health by early detection and prevention of disease. An effective disease management strategy reduces the risks from disease, slow down its progression and reduces symptoms. It is the most efficient and affordable way to reduce the risk of disease. Preventive measures strategies are divided into different stages, e.g., primary, secondary and tertiary. Disease prevention can be applied at any prevention level along with the disease history, with the goal of preventing its progression further. *Primary* It seeks to reduce the occurrence of new cases, e.g., stress management, exercises, smoking cessation to prevent lung cancer and immunization against communicable diseases. Thus, it is most applicable at the suspected stage of a patient. Strategies of primary prevention include risk factor reduction, general health promotion and other protective measure. This can be done by bringing up the healthier lifestyles and environmental health approaches through health education and promotion program. *Secondary* Purpose of secondary prevention is to either cure the disease, slow down its progression, or reduce its impact and is the most appropriate for those in the stage of early-stage or pre-symptomatic disease. It attempts to reduce the number of cases through early detection of the disease and reducing or halting its progression, e.g., detection of coronary heart patient after their first heart attack, blood tests for lead exposure, eye tests for glaucoma, lifestyle and dietary modification. Common approach to secondary prevention includes procedure to detect and treat preclinical pathological changes early through screening for disease, e.g., mammography for early-stage breast cancer detection. *Tertiary* The key aim of tertiary disease prevention is to enhance life quality of patient. Once the disease is firmly established and has been treated in its acute clinical phase, it seeks to soften the impact of disease on the patient through therapy and rehabilitation, e.g., tight control of type-1 diabetes, assisting a cardiac patient to lose weight and improving the functioning of stroke patient through rehabilitation program.

Effective primary prevention to avert new cases, secondary prevention for early detection and treatment and tertiary prevention for better diseases management are not only to improve the quality of life but also helps to reduce unnecessary healthcare initialization. Extensive medicine knowledge and clinical expertise are required to predict the probability of patient that are contracting disease (Table 2).

1.2.2 Challenges

Un-automated analysis of huge and complex volumes of data is expensive as well as impractical. Data mining provides great benefits for the disease identification and

Table 2 Prevention level

Leavell's levels of prevention		
Stage of disease	Prevention level	Type of response
Pre-disease	Primary prevention	Specific protection and Health promotion
Latent disease	Secondary prevention	Pre-symptomatic diagnosis and treatment
Symptomatic disease	Tertiary prevention	Disability limitation for early symptomatic disease

treatment; however, there are several limitations and challenges involved in adapting DM analysis techniques. Successful prevention depends upon knowledge of disease causation, transmission dynamics, risk factor and group identification, early detection and treatment methods, implementation of these methods and continuous evaluation and development of prevention and treatment methods. Additionally, data accessibility (data integration) and constraints(missing, unstructured, corrupted, non-standardized data and noisy) add more challenges. Due to the huge number of patients, it is impossible to consider all those parameters to develop cost-effective and cost-efficient prevention system. The expansion of medical records databases and increased linkage between physician, patient and health record led the researcher to develop efficient prevention system.

Healthcare applications generate mound of complex data. To transform this data into information for decision making, traditional approaches are lagging behind, and they barely adopt advanced information technologies, such as data mining, data analytics, big data etc. Tremendous advancement in hardware, software and communication technologies opens up opportunities for innovative prevention by provided cost-saving and cost-effective solution by improving the health outcomes, properly analyzing the risk and overcoming the duplicate efforts. Barriers to develop such system include non-standard (interoperability), heterogeneous, unstructured, missing or incomplete, noisy or incorrect data.

Disease prevention mainly depends on the data interchange across different healthcare system thus interoperability plays major role in success of prevention system, whereas healthcare sector is still on the way. ISO/TC 215 includes standards for disease prevention and promotion. Standards are a critical component, whereas it is not yet mature in healthcare sector. Many stakeholders (HL7, ISO and IHTSDO (organization that maintain SNOMED CT) with aim to have common data representation are working to address semantic interoperability. Healthcare data is diverse and have different format. Moreover, with the rapid use of wearable sensors in healthcare results in tremendous increase in the size of heterogeneous data. For effective prevention methods, integration of data is required. For years, documentation of clinical data has trained clinician to record data in most convenient way irrespective, how this data could be aggregated and analyzed. Electronic health record systems attempt to standardize the data collection but clinician are reluctant to adopt for documentation.

Accuracy of data analysis depends significantly on the correctness and completeness of database. It is a big challenge to find problems in data and even harder to correct the data, moreover data is missing. Using incorrect data will defiantly provide incorrect result. Whereas ignoring the incorrect data, or issue of missing data introduce bias into analysis that leads to inaccurate conclusion. For the extraction of useful knowledge from large volume of complex data that consist missing data and incorrect values, we need sophisticated methods for data analysis and association. Moreover, data privacy and liability, capital cost, technical issue are other factors. Data privacy is another major hurdle in development of prevention system. Most of the healthcare organizations have HIPAA certification; however, it does not guarantee the privacy and security of data as HIPAA is considering security and policy rather than implementation [7]. With the increase popularity of wearable devices, mobiles and online availability of healthcare data put it on emerging threat. In addition to that, it may increase racial and ethnic disparate because these may not be equally available due to economic barrier.

Rest of the paper is organized as: Sect. 2 describes the existing prevention methodologies and is categorized into three subsections nutrients, policies and HIT. Section 3 presents the data mining development for disease prevention followed by data analytics-based disease prevention application in Sect. 4. Finally, some openly available medical datasets are discussed in Sect. 5 followed by open issues and research challenges are presented in Sect. 6.

2 Existing disease prevention methodologies

Although chronic diseases are among the most common and costly health problems, however, these are the most preventable. Early identification and prevention is the most effective, affordable way to reduce morbidity and mortality as well as helps to improve the life quality [8]. Not only data mining, several other prevention methods are being to reduce the risk factor.

2.1 Nutrients, foods, and medicine

Diet acts as medical intervention, to maintain, prevent, and treat disease. It is major lifestyle factor that contributes extensively for disease prevention such as diabetes, cancer, cardiovascular disease, metabolic syndrome and obesity etc. Poor diet and inactive lifestyle are lethal combination. Joint WHO/FAO expert consultation on diet, nutrition and the prevention of chronic diseases states that chronic diseases are preventable and developing countries are facing consequences of nutritionally compromised diet [9]. Individual has power to reduce the risk of chronic disease by making positive changes in lifestyle and diet. Use of Tobacco, unhealthy diet, and lack physical activity are associated with many chronic conditions. Evidence shows that healthy diet and physical activity does not only influence present health but also helps to decrease morbidity and mortality. Specific diet and lifestyle changes and their benefits are summarized in Table 3. Food and nutrition interventions can be effective at any prevention stage. In primary stage, food and nutrition therapy could be used to prevent the occurrence of disease such as obesity. What if disease is already identified and how diet can help to reduce the effect of disease? Recent studies shows, potential of food and nutrition interventions as secondary and tertiary prevention is also effective preventive strategy that reduce the risk factor and slow down the progression or mitigate the symptoms and complications. Thus, at secondary, it could be used to reduce the impact of a disease and at tertiary stage, it helps to reduce the complications, i.e., stomach ulcer. Dietitian plays critical role in disease prevention, i.e., change in lifestyle can help to delay or prevent type II diabetes.

Please add the following required packages to your document preamble:

2.2 Policy, systems and environmental change

After many year focus on individual; policies, systems and environmental changes are new way of thinking to improve the quality of healthcare sector. It affects large segments of the world population simultaneously. Disease prevention is much easy if we develop such environment that can help community to adopt health lifestyle, proper nutrition and medications etc. Developed nations are promoting social, environmental, policy, and systems approaches to support healthy life such as low fat diet at restaurants, smoking restricted areas, increase in prices of tobacco items and urban, healthy food restriction for all students, infrastructure design that leads to lifestyle change, i.e., increase in physical activity. United nations economic commission for Europe (UNECE) member states have committed to

Table 3 Convincing and probable relationships between dietary and lifestyle factors and chronic diseases [10]

Factors	CVD	Type-2 diabetes	Cancer	Dental disease	Fracture	Cataract	Birth defect	Obesity	Metabolic syndrome	Depression	Sexual dysfunction
Life style											
Avoid smoking	↓	↓	↓	↓	↓	↓		↑			↓
Physical activity	↓	↓	↓		↓			↓	↓	↓	↓
Avoiding overweight	↓	↓	↓		↑	↓			↓		↓
Diet											
Using healthy fat	↓	↓						↓	↓		
Fruits and veggetable	↓		↓		↓	↓	↓	↓			
Using whole grains	↓	↓						↓	↓		
Reducing sugar	↓	↓		↓				↓	↓		
Reducing calories								↓	↓		
Reduction in sodium	↓										

↓ Decrease risk, ↑ increase risk

implement health policies to ensure the increase in longevity by quality of life [11]. Some good examples are 5 A DAY program to increase healthy nutrition in the UK, WHO: Age-friendly cities, Lifetime homes in the UK, Support program for dementia patients living alone in Germany [12].

Policy change includes new rules and regulation at the legislative or organizational level, i.e., tax on tobacco item and soft drinks, provide time off during office hours for physical activity, changing community park laws to allow fruit trees. System change includes change in systems strategies within organization such as improving school infrastructure, transportation systems. Environmental change includes the changes in physical environment such as incorporating sidewalks and recreation areas in community areas, healthy food in restaurants.

2.3 Information technologies

Health Information Technology (HIT) strategy is to put information technology to work in healthcare sector in order to reduce healthcare cost and increase efficiency. HIT makes it possible to get maximum benefits for patient, healthcare organizations and government through intelligent patient information processing. Happy marriage of healthcare and information technology includes variety of electronic approaches that are used not only to manage information but to improves the quality of clinical and other preventive services, i.e., disease prevention, early disease detection, risk factor reduction and complication management. Healthcare transaction generates huge amount of data. This expansion of medical record databases and increased linkage between physician, patient and health record led the researcher to develop efficient disease prevention systems. Thus, there is a need to transform health data to information through the use of innovative, collaborative and cost-effective informatics and information technology. IT provides strategic value to achieve health impact and health quality by transforming these mounds of data into information. Computer-based disease control and prevention is an ongoing area of interest to the healthcare community. Disparities in access to health information can affect preventive services. Increased access to the technology (i.e., handled devices and fast communication) and availability of online health information reduced the disease risk. It makes the user to be able to access health information and make good decision. Increased use of HIT tools (i.e., reminders, virtual reality applications and decision support system) helps to reduce the risk factor.

Effective disease prevention requires identifying and treating individuals at risk. Several preventive measures are being used to improve therapy adherence. Relevant information, such as blood pressure, cholesterol measurement, fasting plasma glucose etc., is recorded electronically, and it makes automatic risk prediction possible. Moreover, the rapid growth of cellular networks provides opportunity to be in touch with patient [13]. Reminder-based preventive services involves continuous risk assessment of several life-threatening disease (cardiac disease [14–17], diabetes management [18–21] through vital sciences [14, 15, 18–20] or reminder of due for specific preventive services such as vaccination, follow-up appointment, weight loss [22–25] or reminder for medicine [22, 23, 26]. This type of communication could by any electronic medium, i.e., call, SMS or emails. Studies found that automated reminder services are very effective in boosting patient adherence [27, 28]. Healthcare databases contain cardiac disease data, but practical methods are required to identify patient at risk and continuous monitoring of those patient through electronic vita science data, i.e., Blood pressure, cholesterol and glucose etc. Several studies have been performed to reduce risk through automated reminder using collected primary care data [14, 15]. Text message-based intervention for supporting diabetic patient helps to improve self-management behaviors and achieve better control over glycemic [18–20]. Research studies show that reminders (through emails, SMS, calls, social media and wearable devices etc.) based individual data leveraged to reduce risk. Integration of wearable sensors, EHR and expert system with automatic reminder will end up at efficient disease prevention system.

The wish of efficient healthcare is coming within the reach after the advent of smart wearable technology. These sensors are reliable and efficient for real-time data collection and analysis [29–31] thus are really good source of preventative methods for several threatening disease, i.e., cardiovascular [32–34], cardiopulmonary [35, 36], diabetes [37, 38] and neurological function [39]. Several companies (i.e., Medtronic, BottomLine, Medisafe, ImPACT, Actismile. Allianz and AiQ) are providing low cost wearable solution like concussion management, cardiovascular, monitoring, reminders, diabetes, overweight, metabolism etc. Recent development in intelligent data analysis methods nano-electronics, communication and sensor technology made possible the development of small wearable devices for healthcare monitoring. Wearable sensors are able to collect most of the useful medical data required for early identification and prevention, but it is still costly and most of the healthcare system are not yet able to process this data. In near future, wearable devices are getting cheaper (few dollars), smaller (almost invisible, i.e., sensor-clad smart garments, implanted devices), accurate (no approximation of data) and powerful (low battery consumption, high communication and processing power). Moreover, due to the recent development high processing

sensors, research is shifting form simple reasoning to high-level data processing by implementing pattern recognition, data mining methodologies to provide much valuable information, i.e., fall detection, heart attack detection via data processing and sending message to emergency services or family member for immediate action.

3 Algorithms and methods

Biomedical data is more complicated and is getting larger day by day. Efficient analysis of this huge amount of data provides large amount of useful information that could be used for health promotion and prevention. Traditional methods deal with primary data and fail to analyze big data. Thus question arises: how to analyze it efficiently? Answer is: we need much smarter algorithms to analyze this data. The major objective of data analytics is to find hidden pattern and interesting relations in the data. The significance of such approaches is to provide timely identification with less number of clinical attributes [40]. This secondary data is used for important decision making.

Healthcare data analytics is multidisciplinary area of data mining, data analytics, big data, machine learning and pattern recognition [41]. Intelligent disease control and prevention is an ongoing area of interest to the healthcare community. The basic goal of data analytics-based disease prevention is to take real-world patient data and to help to reduce the patient at risk. Recent development in machine learning and data analytics for handling complex data opens new opportunities for cost-effective and efficient prevention methods that can handle really big data. Data mining provides the methodology and technology to transform these mounds of data into useful information for decision making by exploring and modeling big data [42]. This section is divided into four subsection based on data mining functionalities in healthcare, i.e., classification of disease based on symptoms. In this section, our goal does not go much into detail but shows the basic concept and difference between different algorithms.

- Classification is the process of finding set of variables to classify data into various types, i.e., disease identification, or medication etc.
- Clustering: data grouping or data distinguishing used for decision making, i.e., finding pattern in EHR, prediction of readmission.
- Associate analysis: discover interesting, hidden relation between medical data, i.e., frequently occurrence disease.
- Anomaly detection: is the identification of abnormality that does not follow any specific pattern.

3.1 Classification

Success of disease prevention is based on early detection of disease. Accurate disease identification is necessary to help the physicians to deal it with proper medication. The goal of classification is to accurately predict target case dataset from unseen data. Classification is the process of assigning given object to one of the predefined class. There has always been a debate on the selection of best classification algorithm. Several statistical and machine learning-based contributions were made for disease prediction and wide range of classifiers (shown in Fig. 2) has been used, i.e., decision tree [43–46], SVM [47–49], Naive Bayes [43, 50], neural network [51] and *k*-nearest neighbor [43, 52], PCA [53, 54] for disease classification.

3.1.1 Decision tree

Decision tree is used as prediction model that maps observation sequence to classified items. Classification method uses tree-like graph and is based on sorting of feature values. Each node represents the feature in an instance to be classified, whereas each branch in decision tree represents the value that the node can assume. In decision tree, nodes may have two or more child nodes; internal nodes are denoted by rectangle and ladled with input features; leaf nodes are denoted by oval and ladled with class or probability distribution over class; classification rules are represented by path from root to the leaf and internal nodes. The decision tree are of two types: regression tree analysis (predicted outcome is real value number) and classification tree analysis (predicted outcome is class). Several popular decision tree algorithms used for data mining purpose are ID3 (C4.5, SPRINT), AID (CHAID), CART and MARS.

ID3 or Iterative Dichotomiser-3 is used to generate decision tree from dataset and one of the simplest learning method [55]. C4.5 is the extension of ID3 and uses gain ratio as splitting criteria whereas ID3 uses binary splitting [56]. Unlike ID3, it can handle data with missing values,

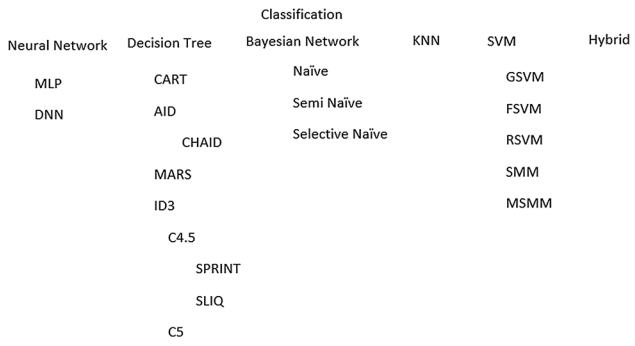

Fig. 2 Classification methodologies

attributes with differing costs and handle both discrete and continuous attributes. It became quite popular and was ranked at No. 1 in the list of top 10 algorithms in data mining. C5.0 is the extension of C4.5 and introduces new features such as misclassification cost (weight on classification error based on its importance), case weight attribute (quantification the importance of each case). Misclassification cost boosting leads to improvement in predictive accuracy [57]. It also supports cross-validation and sampling. Several new data types are included especially not application if data values are missing.

ID3 and its descendant construct flowchart-like tree structure and are based on the recursive and divide-and-conquer algorithm.

- If set S is small or all examples of S belong to same class
- return a new leaf and label it with class C in S.

Otherwise

- Run a test based on single attribute with two or more outcome;
- Make this test as a root of the tree with one branch for each outcome.
- Partition S into subset S1, S2 and apply same procedure recursively.
- Stop if all its instances have the same class.

3.1.2 Support vector machine

It is supervised learning approach that is based on statistical learning theory [58]. It is one of the most simple, robust, accurate and successful classification method that provided high performance in solving complex classification problems especially in healthcare. Due to its discriminative power especially for small data and large set of variables, data driven and modal free property, it is has been widely used for disease identification and classification problems. Unlike logistic regression that depends on pre-determined modal for prediction of binary event through fitting of data onto logistic curve, SVM produces binary classifier that discriminates between two classes by separating the hyperplanes through nonlinear mapping of data into feature space that is high dimensional. Unlike regression-based methods, SVM allows more input features than samples, so it is particularly suitable in classifying high-dimensional data. Moreover, some kernels even allow SVMs to be considered as a dimensionality reduction technique.

SVM hyperplanes are the decision boundaries between two different sets of classes. SVM uses support vectors to find hyperplane with maximum margin form set of possible hyperplane. In ideal case, two classes can be separated by a linear hyperplane; however, in real world, data is not that simple. Due to the limitation of SVM such as binary classification, memory requirement and computational complexity, several other variations of SVM are presented such as GVSM, RSVM, LSVM [59], TWSVMs (twin support vector) VaR-SVM (value-at-risk support vector machine), SMM [60], SSMM [61], RSMM [62, 63], cooperative evolution SVM [64]. Using Kernel functions (polynomial, linear, sigmoid, and radial) to add more dimensions to low dimensional space, two class problem could be separable in high-dimensional space. As compared to other classification methods, SVM training is extremely slow and computational complex. However, it does always rank well among the list of best classifier.

3.1.3 Machine learning

Machine learning have emerged as advanced data analytics tools and recently have been successfully applied in a variety of applications including medical diagnosis assistance to the physician. Why machine learning is important for data analytics even though state-of-the-art data analytics algorithms are available? The essence of the argument is that in some cases where other analytics methods may not produce satisfiable predictive result, machine learning can help to improve the generalization ability of the systems by training on the data annotated by human experts. Mostly, the predictive accuracy obtained using machine learning is slightly higher than other analytics methods or human experts [65] and it is high affordable to the noise data. However, despite these fact, neural networks was not preferable choice for data analytics due to facts: NN is quite slow in training and classification making it impractical for large data and trained neural networks are usually not comprehensible [66]. However, due to the recent advancement in computation power, neural network is increasingly favored for the development of data analytics applications. Important neural network algorithms are backpropagation neural network (BPNN), associative memory, and the recurrent network.

Deep machine learning also known as hierarchical learning or deep learning is a branch of machine learning based on a set of algorithms which have one or more hidden layers that automatically extract high-level and complex abstractions as data representation [67]. Data is passed through multiple layers in hierarchical manner and each layer applies nonlinear transformation on to data. With its amazing empirical result over past couple of years, it is one of the best predictive algorithms for big data. Main advantage of deep learning is automatic analysis and learning of huge amount of unlabeled data typically learning data representations in a greedy layer-wise fashion [67, 68]. Several attempt has been made for disease

prevention using deep learning methods [69, 70]. An important question is whether to use deep learning for medical data analytics or not. As most of the healthcare data is unstructured and unlabeled, due to the automatic high-level data representation property of deep learning, it could be used in effective way for prediction problems, but it needs huge volume of data to be trained.

3.1.4 Bayesian networks

Bayesian network, Bayes network or Bayes model is probabilistic directed acyclic graph that encodes probabilistic relationships among variables of interest. Nodes represent set of random variable and their conditional dependencies (represented by edges) via directed acyclic graph. Unconnected nodes represents conditionally independent variables, for example, probabilistic relationships between diseases and its symptoms. Nodes are associated with probability distribution function that takes set of values as input and output the probability of variable represented by node. Advantages of Bayes networks are model interpretability, efficient for complex data, requiring small training dataset. For example, Bayes classifier to diagnose correct disease based on patient observed symptoms. Important algorithms of Bayesian networks are Naive Bayes, semi Naive Bayes, selective Naive Bayes, one-dependence Bayesian classifiers, unrestricted Bayesian classifiers, and Bayesian multinets, Bayesian network-augmented Naive Bayes and *k*-dependence Bayesian classifiers etc.

Naive Bayes has proved itself a powerful supervised learning algorithm for solving classification problems and has been extensively used for healthcare data analysis specially for disease prevention. It is built upon strong assumption that features are independent with each other (so that classifier could be simple and fast) and it assumes that the effect of variable value on the given class is independent of other variable values. Mostly, Naive Bayes uses maximum likelihood for parameter estimation and classification is done by taking highest posterior of classified values. Based on given set of variables that belong to known class, the aim is to construct rules for classification of unknown data based. Limitation of Naive Bayes is sensitivity to correlated features. Selective Naive Bayes uses only subset of given attributes in making prediction [71]. It reduces the strong bias of Naive independence assumptions owing to variable selection [72]. The objective is to find among all variables, the best classifier, compliant with the Naive Bayes assumption. Several selection methods have been presented so far [71–74]. Selective Naive approach is good for datasets with a reasonable number of variables; however, it does not scale for large datasets with large number variables and instances [72].

3.1.5 *k*-nearest neighbor

The *k*-nearest neighbor or *k*-NN algorithm, a non-parametric classification and regression method is based on nearest neighbor algorithms and is one of the top 10 data mining algorithms [75]. The aim is to find output with a class membership. kNN finds a group of *k* objects in the training dataset that are closest to the test object and decides the assignment of a label on the predominance of a particular class in this neighborhood. *k*-nearest neighbors are identified by computing the distance of the object to labeled object. The calculated distance is used to assigner class. The main advantages of kNN are simple in implementation, robust with regard to search space, and online updation of classifier and few parameters to tune.

There are a lot of different improvements in the traditional KNN algorithm, such as the wavelet based *k*-nearest neighbor partial distance search, equal average nearest neighbor search, equal average equal-norm nearest neighbor code word search, equal-average equal-variance equal-norm nearest neighbor search and several other variations.

Guidelines The complexity in classification of data arises due to the uncertainty and the high dimensionality of medical data. Studies [76–78] show that there is a large amount of redundancy, missing values as well as irrelevant variables in healthcare data that can effect the accuracy of disease prediction. Conventional wisdom is that larger corpora yield better predicting accuracy however data redundancy and irrelevancy introduces bias rather than benefits that distorts learned models. Thus, before applying any prediction model, dataset needs to be processed carefully by removing the redundant data (e.g., age, DOB, redundant lab reports) as well irrelevant data (e.g., gender attribute in case of gestational diabetes). However, little is known about the redundancy that exists in the data as well as what type of redundancy is beneficial as opposed to harmful. Statistical methods (correlation analysis etc) could be used to identify irrelevant variables, other alternative is feature selection based on the importance of specific variable. To deal with missing values, it is recommended to use data analysis methods that are robust to missingness that are good to use when we are confident that mild to moderate violations of the technique's key assumptions will produce little or may be no biasness in the resultant conclusions. As healthcare data is highly sensitive, one drawback of these method for missing or irrelevant information may lead to distortion of important relationship between set of dependent and independent variables. For example, smoking, drinking and helicobacter pylori individually might not affect stomach cancer significantly whereas all together could effect significantly [41]. Thus, efficient approaches are required for data preprocessing to avoid ignorance of that type associated data as deletion of

such depended variable will significantly effect prediction power.

As discussed above, no generic classifier works for all type of data and it varies form data to data and nature of the problem. For example, small and labeled data, classifier with high bias (Naive Bayes) is best choice. However, if the data is quite large, classifier doesn't really matter so much, so the selection of classifier is based on its scalability and run-time efficiency. Classifier performance can be significantly improved through effective features selection. In case of big data, deep learning is good choice as it learns features automatically. For better prediction, different feature sets and a couple of classifiers should be selected and tested. The best classifier that beats all other classifiers could be selected for prediction activities.

3.2 Clustering

Clustering is a descriptive data analysis task that partitions the data into homogeneous groups based on the information found in data that best describes the data and its relationship, i.e., classifying patients into groups. Organization of data into clusters shows the internal structure of the data. The goal is to group the individuals or objects that resemble each other for the purposes of improved understanding. The greater the dissimilarity between groups and the greater the similarity within group provides better clustering, i.e., a new observation is assigned to the cluster with closest match and it is assumed to have similar properties to others in same cluster. It is unsupervised learning that occurs by observing only independent variable. Unlike classification, it does not have training stage and does not use "class".

The heart of clustering analysis is the selection of the clustering algorithm. Selection of proper clustering method is important because different methods tend to find different type of cluster structure and is based on the type of data structure. A number of clustering methods have been proposed and has been widely used in several disciplines, i.e., model fitting, data exploration, data reduction, grouping similar entities, prediction based on groups etc. It is categorized into partitional(unnested), hierarchical(nested), grid-based, density-based, subspace-clustering and some other clustering algorithms fuzzy, conceptual clustering as shown in Table 4.

Partitional clustering requires a preset number of clusters and it decomposes a set of N observations into a set of k disjoint clusters by moving them from one cluster to another, starting from an initial partitioning. It classifies the data into k cluster based on requirements: each observation belongs to exactly one cluster and each cluster contains at least one observation. An observation may belong to more than one cluster in fuzzy partitioning [79]. One of the most important issues in clustering is the selection of number of clusters and it is complicated if no prior information exist. Another issue is the selection of proper parameter and proper clustering algorithm. Inappropriate choice of clustering algorithm or wrong choice of the parameters, clustering may not reflect the desired data partition [79]. Not all methods are applicable to all type of problem, so selection of algorithm is based on the type of problem. Several contributions have been made to address this issue [79–84]. One of the most common partitional algorithms is k-means. It uses an iterative refinement technique and attempts to minimize the dissimilarity between each element and the center of its cluster. k-mean partition the data into k cluster represent by their center. Cluster center is computed as mean of all instance belong to that cluster. k-medoids, k-medians, k-means++, Minkowski weighted k-means, bisecting k-means, spherical k-means etc are some variations of k-mean.

Hierarchical clustering also called hierarchical cluster analysis or HCA is a clustering method that seeks to build a hierarchy of clusters (permit clusters to have sub-clusters) and can be viewed as sequence of partitional clustering. Based on pairwise distance between sets of observations, HCA successively merges (hierarchically group) most similar observations until a termination condition holds. Unlike partition, it does not assume a particular value of k. First step in hierarchical clustering is to look the most similar/closet pair which are then joined to make clustering tree. Division or merging of cluster is performed on the similarity measure that is based on optimal criteria. Strategies for hierarchical clustering generally fall into two types agglomerative and divisive also known as bottom-up and top-down respectively. Divisive clustering starts with single cluster containing all observation and consider all possible way to split into appropriate sub-clusters recursively, whereas agglomerative starts with one point cluster and recursively merges two or more clusters into a new cluster in bottom-up fashion. Main disadvantages of the hierarchical methods are inability to scale and no backtracking capability. Standard hierarchical approaches suffer from high computational complexity. Typical hierarchical clustering includes BIRCH, CURE and ROCK etc. To improve the performance, several approaches have been presented. In *fuzzy based clustering* also called soft clustering, where each data point can belong to more than one cluster. It is continuous interval [0, 1] of belonging label, 0, 1 is being used instead of discrete value in order to describe relationship more reasonably. Most widely used fuzzy clustering algorithms are the fuzzy C-means (FCM), fuzzy C-shells and mountain method (MM) algorithm. Fuzzy C-Means get membership of each data point to every cluster by optimizing the object function. *Distribution-based clustering* is iterative, fast and natural clustering of large

Table 4 Clustering algorithms

Category	Algorithm
Partition	*k*-mean, *k*-medoids, weighted *k*-mean, CLARA CLARANS PAM etc.
Hierarchy	Agglomerative algorithms (CURE, CHAMELEON, ROCK), Divisive algorithms (average link divisive, PDDP etc)
Distribution	DBCLASD, GMM etc.
Density	DBSCAN, OPTICS, mean-shift etc.
Fuzzy	FCS, FCM, MM etc.
Grid	STING, CLIQUE etc.
Graph	CLICK, MST etc.
Fractal	FC etc.
Model	GMM, SOM, ART, genetic algorithms etc.

dataset. It automatically determines the number of clusters and produces complex models for clusters that captures the dependences and correlation of data attributes. Data generated from the same distribution belongs to the same cluster if there exists several distributions in the original data [85]. GMM using expectation maximization and DBCLASD are of the most prominent method. *Density-based clustering* identifies distinctive clusters in the data based on high-density region (contagious region of high-density area separated from other clusters having low density region). Typical methods of density-based clustering includes DBSCAN, OPTICS and Mean-shift. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is the most well-known density-based clustering algorithm and does not require the number of clusters as a parameter. Further detail on clustering algorithms can be found at [85].

Guidelines Clustering is normally an option when there is little or no data available. Moreover, they do not concentrate on all requirements simultaneously which makes the result uncertain. Clustering efficiency could be affected by outliers, missing attributes, skewness of data, high dimensionality, distributed data and selecting of proper clustering method with respect to application context. Missing and outlier can significantly effect the measurements of similarity and the computational complexity. Thus, before going to clustering of data, it is recommended to eliminate the outliers. Grouping the data into cluster provides signification information about the object. Without prior knowledge of number of clusters or any other composition information of clusters, cluster analysis cannot be performed. Clustering (distance-based clustering) performance is also effected by distance function used. High dimensionally is another issue of data clustering. Number of features are very high and may even exceed the number of samples. Fixed number of data points become increasingly sparse as the dimensionality increase. Moreover, the data is so skewed that it cannot be safely modeled by normal distributions. In that case, data standardization and model based clustering or density-based clustering could perform better.

Selection of clustering algorithms for particular problem is quite difficult. Different algorithms may produce different results. In case when no data or little data is available, then hierarchical clustering could be good option as they do not requires predetermination of number of clusters. However, hierarchical clustering methods are computationally expensive and un-scalable thus not good option for large data. When the number of sample are high, algorithms have to be very conscious of scaling. Generally, successful clustering methods scale liner or log-liner. Quadratic and cubic scaling is also fine but with linear behavior. Generally, healthcare data is not totally numerical, thus conversion is required to make it useful. Conversion into numerical value could distort the data that could affect the accuracy, for example there attributes values A, B, C are converted to numerical values 1, 2 and 3. Here, there could be distortion conversion from attribute value to numerical, i.e., if distance of A and B is 2 whereas B and C is 1.

3.3 Association analysis

It is a highly unsupervised approach for discovering of hidden and interesting relations between variables in large datasets (Fig. 3). It is intended to identify strong rules that will predict the occurrence of an item based on occurrence of other item. For example, which drug combinations are associated with adverse events?. Association rule is an implication of the form $\{X\} \rightarrow \{Y\}$ where $X \cap Y = \emptyset$.

Strength of association rule is the measure in term of its support and confidence (rules are required to satisfy user-specified minimum support and confidence at the same time). Support is an indication of how frequently a rule is applied to a give dataset and confidence is an indication of how frequently the rule has been found to be true.

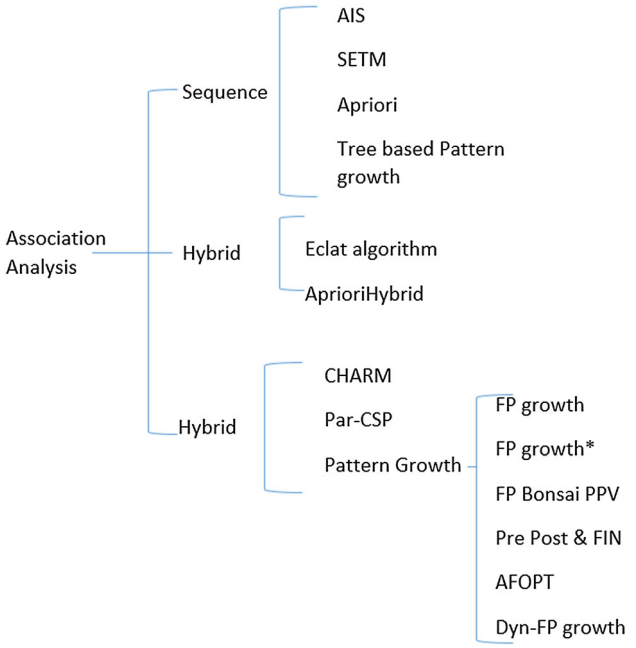

Fig. 3 Association mining methodologies

Uncovered relations could be represented using association rules or set of frequent items. The following rules shows that there is strong relationship between diabetic patient and retinopathy or heart-block can leads to hypertension. Association rules interestingness is measure through support and confidence. Support of 6% means that pregnancy and gestational diabetes occurred together in 6% of all the transactions in the database and Confidence of 90% means that the patients who are pregnant, 90% of the times, they also suffer gestational diabetes.

$$\{\text{Pregnancy}\} \rightarrow \{\text{Gestational diabetes}\}$$
$$[\text{support} = 6\%, \text{confidence} = 90\%]$$
$$\{\text{Pregnancy}\} \rightarrow \{\text{Type-II Diabetes}\}$$
$$[\text{support} = 1\%, \text{confidence} = 8\%]$$

Support and confidence are commonly used to extract association rules through measuring interestingness of rule. Support of 6% means that pregnancy and diabetes occurred together in 6% of all the transactions contained in the database. While confidence of 75% means that, the woman who are pregnant, 75% of the times, suffer type-II diabetes as well. However, it is well known that even the rules with a strong support and confidence could be uninteresting.

In healthcare, this kind of association rules could be used to assist physician to cure patient or to find the relationships between various diseases and drugs or occurrence of other associated disease as it is revealed that occurrence of one disease can lead to several other diseases. Association rules are created by analyzing data for frequent if/then patterns and using the criteria support and confidence to identify the most important relationships. Thus, the problem of association rules is divided into two phases: generation of frequent itemset (find all the itemset: frequent pattern that satisfy the minsup threshold) and rule generation (extract the all high confidence rules form frequent patterns). Computation requirement of frequent-pattern generation are more expensive than rule generation that is straightforward. Finding of all frequent item-sets involves searching all possible item-sets in the database thus is difficult and computational expensive that can be reduced by reducing the number of candidate (a priori method) or reducing the number of comparison (FP-Growth, tree generation etc).

To design efficient algorithms for association rules computation, several efforts were presented over time, i.e., Sequence (Apriori, AprioriDP, tree based partition), Parallel (FP-growth, partition based, i.e., Par-CSP [86] and hybrid (Eclat). The most common algorithm for association rule is the Apriori algorithm. It is an algorithm for frequent item set mining and association rule learning over transactional databases. It uses breadth-first search and hash tree strategy to count the support of itemsets and uses a candidate generation function which exploits the downward closure property of support. AprioriDP is extension of Apriori that utilizes Dynamic Programming in frequent itemset mining. Finding of frequent sets requires several passes thus performance depends upon the size of the data. Processing of high density in primary memory is not feasible. Accesses to secondary memory could be reduce by effective partitioning. Tree based partitioning approach organizes the data into tree structures that can be processed independently [87]. To relieve the sequential bottlenecks, parallel frequent-pattern discovery algorithms exploit parallel and distributed computing resources. FP-growth counts item occurrence and stores into 'header table' followed by construction of FP-tree structure by inserting instances. Core of GP-growth is the use of frequent-pattern tree (FP-tree), a special data structure to retain the itemset association information. FP-tree is a compact structure that store quantitative information about frequent pattern. Further detail on association algorithms can be found at [30, 88].

Guidelines Traditionally, association analysis was considered an unsupervised problem and has been applied in knowledge discovery. Recent studies showed that it has been applied for prediction in classification problem. Generating association rules algorithms must be tailored to the particularities of the prediction to build effective classifiers. Computational cost could be reduced by sampling database, adding extra constraints, parallelization and reducing the number of passes over the database. To expedite multilevel association rules searching as well as

avoiding the excessive computation in the meantime, much progress has been made during last few years. Other issues involved in association mining are selection of non-relevant rules, huge number of rules and selection of proper algorithm. A small number of rules are used by domain expert whereas all rules are used by classification system. For good accuracy, it would be better, if the non-relevant rules are eliminated by domain experts (e.g., gestational diabetes females) and to readily organize raw association rules using a concept hierarchy.

3.4 Anomaly detection

Anomaly detection has been widely researched areas and provides useful and actionable information in a variety of real-world scenarios such as timely detection of an epidemic that can helps to save human life or ECG signals or other body sensors for critical patient monitoring to detect critical, possibly life-threatening situations [89]. It is quite challenging to develop generic framework for anomaly detection due to unavailability of labeled data and domain specific normality; thus, most of the anomaly detection methods have been developed for certain domain [90]. It is an important problem and has a significant impact on the efficiency of any data mining system. The importance of its detection is due to the fact that its presence in data can compromise data quality and reduce the effectiveness of learning algorithm. Thus, it is a very critical problem and requires high degree of accuracy. Anomaly detection is the detection of items, events or observations in the data that do not conform to an expected pattern [91]. Broadly speaking, based on their nature, anomalies can be classified into four categories are point (individual data instance is considered as anomalous with respect to the rest of data), contextual (data instance is anomalous in a specific context), collective anomalies (collection of related data instances is anomalous with respect to the rest of dataset).

Based on the available data labels, three broad categories of anomaly detection techniques exist: unsupervised, supervised and semi-supervised anomaly detection. As availability of labeled dataset for anomaly detection is not easily possible; thus, most of the approaches are based on unsupervised or semi-supervised learning methods where the purpose is to find the abnormal behavior. Supervised learning-based methods uses fully labeled data, semi-supervised anomaly detection uses anomaly-free dataset for training and find anomaly if it deviated from trained dataset whereas unsupervised detection method uses intrinsic information to detect anomalous values based on the deviation form majority of data. Why the supervised learning methods is not that successful for anomaly detection? First of all, not all supervised classification method suits for this task, i.e., C4.5 cannot be applied to unbalanced data [92]. Secondly, anomalies are abnormal behavior and they might not have known prior. Semi-supervised is one-class (normal data)-based method that is trained on normal data without anomalies and after that deviations from that data are considered anomalies. Unsupervised is the commonly used method for anomalies detection and it does not require data labeling. Distance or densities are used to estimates the abnormalities based on intrinsic properties of the dataset. Clustering is the simplest unsupervised methods to identify anomaly in data (Fig. 4).

The performance of detection method depends on dataset and parameter selection. Several methods for anomaly detection have been presented in literature, some of most popular anomaly detection are SVM, Replicator neural networks, correlation-based detection, density-based techniques (k-nearest neighbor, local outlier factor), deviations from association rules and ensemble techniques (using feature bagging or score normalization). Output of anomaly detection system could be score or label [93]. Score (commonly used for unsupervised or semi-supervised) is a ranked list of anomalies that are assigned to each instance depending on the degree to which instance is considered anomaly whereas labeling (used for supervised) are binary output (anomaly or not) (Fig. 5).

k-NN (not k-nearest neighbor classification)-based unsupervised anomaly detection is a straightforward way for anomaly detection. It is not able to detect local anomalies and only detect global anomalies. Anomaly score is computed on k-nearest-neighbors distance (computed either single (known as kth-NN) [94] or average (known as k-NN) [95]). *Outlier Factor Local outlier factor* is well know method for finding anomalous data points by measuring the local deviation of a given data point with respect to its neighboring point. LOF works in three steps: k-NN is applied to all data points; estimation of local

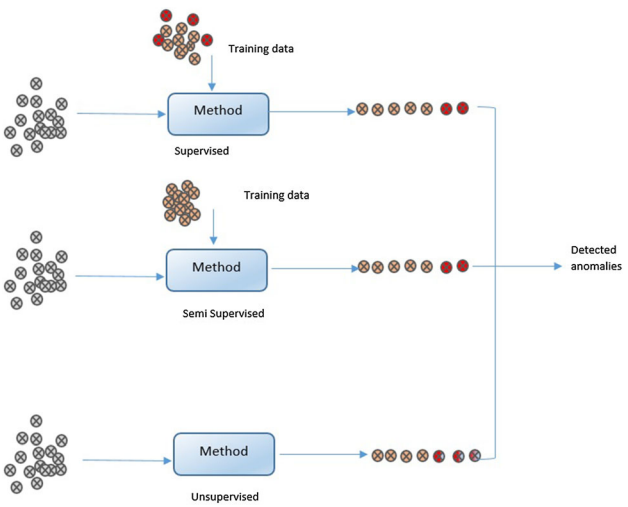

Fig. 4 Anomaly detection methods based on dataset

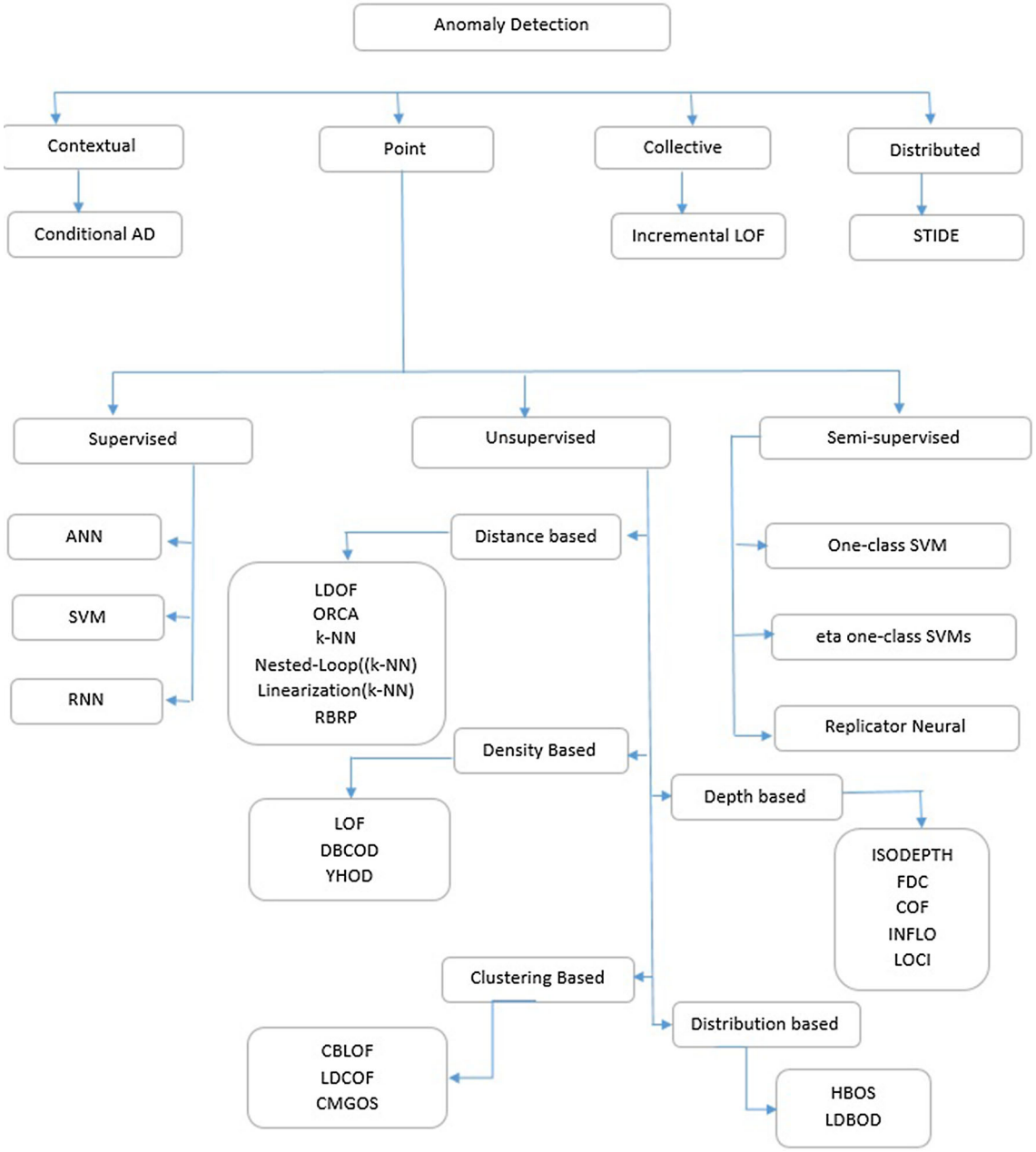

Fig. 5 Anomaly detection methods

density using local readability density; and finally computation of LOF score through LRD comparison with its neighbor LRD. In fact LOF is the ratio of local density. *Connectivity-Based Outlier Factor (COF)* is based on the spherical density computation rather than *k*-NN-based selection. Unlike LOF and COF, *Cluster-Based Local Outlier Factor* (CBLOF) performs density estimation for clusters. First, clustering is performed to find clusters and then anomaly score is calculated based on the distance of each data point with cluster center. *LDCOF* (local density cluster-based outlier factor) is the extension of CBLOF that considers the density of local cluster as well by segmenting into small and large clusters. *LDCOF* (local density cluster-based outlier factor) is the extension of LDLOF to overcome the shortcoming of CBLOF and unweighted-CBLOF by estimating the clusters densities assuming a spherical distribution of the cluster members. LDCOF is local score as it is based on distance of instance to its cluster center. LOF and COF output are scores, but they do not determine the threshold. *Outlier Probability* resolves this issue. *Clustering-based Multivariate Gaussian Outlier Score* (CMGOS) is another extension of cluster-based anomaly detection. First, *k*-means clustering is performed followed by computation of covariance matrix. Local

density estimation is calculated by estimating a multivariate Gaussian model and Mahalanobis distance is used for distance computation. *Histogram-based Outlier Score* (HBOS) is unsupervised, anomaly detection method [96]. It is fast than multivariate approach due to independence of features. For each feature, histogram is computed followed by multiplication of inverse height of the bins it resides for each instance. Due to the feature independence, HBOS can process a dataset under a minute, whereas other approaches may take hours.

One-class SVM [97] is a commonly used semi-supervised anomaly detection method that is trained on anomalous free data [92]. SVM is trained on anomaly-free data whereas normal data is used for classification and once-class SVM classifies into normal or anomaly. In one-class SVM-based supervised anomaly detection scenario, it is trained on dataset and for anomaly detection, each instance in data is scored by normalized distance [98]. *Conditional anomaly detection* is different from standard anomaly detection method [99]. The goal is to detect unusual pattern relating input attributes X an output attributed y in an example a whereas standard anomaly detection identify data that deviate from other data. Amer presented eta one-class SVM to deal with the sensitivity of one-class SVM for outlier data [98]. An improved one-class SVM "OC-SVM" that exploits the inner-class structure of the training set via minimizing the within-class scatter of the training data [100].

Guidelines Not only generalized healthcare data, automated anomaly detection, even for a specific disease like thyroid disease, is not easy to solve. Successful approaches need to use a range of techniques to deal with real-world problem especially when data is quite large and complicated. During recent years, a large number of methods have been presented for anomaly detection, selection of method depends upon the type of anomaly and data itself. Before selecting a method, it is important to know which anomaly detection technique is best suited for a given anomaly detection problem and data. When dimensions are high in number, NN and clustering-based methods are not good option as the distance measures in high number of dimensions are not able to differentiate between anomalous and normal data. To deal with high-dimensional data, spectral techniques could be used as they explicitly address high dimensionality issue through high to low dimensional projection mapping. If fully labeled or partially-labeled data is available then supervised- or semi-supervised-based method could be used even in case of high-dimensional data. Semi-supervised method are preferable over supervised methods as supervised methods are poor to deal distribution imbalance of labels (normal vs specific anomaly) thus data need to be preprocessed to overcome this biasness issue. Due to the sensitivity of healthcare finding, one main challenge is to find best feature vector that provide maximum discrimination power and avoid false positive to minimum and has high accuracy. Thus, it is required to find efficient feature vector that characterizes the anomalous occurrence and its location causing agent in time or frequency domain using some domain transformation or wavelets. Another issue that should be considered is computational complexity as anomaly detection methods have high computational complexity especially when dealing with real-time data, i.e., data generated by ICU monitoring sensors. Most of the anomaly detection methods are unsupervised or semi-supervised and these techniques require expensive testing phases which can limit it in real setting. To overcome this issue, one should consider model size and its complexity. Another big challenge that effects the performance of anomaly detection method is data itself. Data should be accurate, complete and consistent. In healthcare sector, data contains noise which tends to be similar as anomalies or effect detection, Thus before applying anomaly detection, data should be cleaned carefully.

4 Applications

Today, abundant data collected in medical sciences that could be utilized by healthcare organization to get wide range of benefits, i.e., descriptive analytics (What has happened based on symptoms), predictive analytics (What will happen based on current situation) and prescriptive analytics (how to deal with this situation based on best practice and the most effective treatments. As discussed earlier, traditional methods are not to process mounds of data. These days, much of the focus in recent development has been on early detection and prevention of diseases and data mining and data analytics interventions in health sector are increasingly being used for this purpose. By extracting hidden information form medical data using data analytics and machine learning techniques, intelligent system can be designed that besides the physician knowledge and be used as expert system for early diagnose and treatment. To promote early detection and prevention of disease, several intelligent approaches have been presented. Based on the types of applications, we have divided the following discussion into four subsections.

4.1 Examples/case studies

Despite the extensive need of data analytics applications in healthcare, although the work done is not that efficient; however, data analytics is driving vast improvements in healthcare sector and in future, we will see the rapid, widespread implementation across the healthcare

organizations. Here are some case studies for disease prevention in past few years.

ScienceSoft, Truven, Artelnies, TechVantage, Xerox research centre India (XRCI), are some of solution provider companies that are applying data analytics for clinical data efficiently. ScienceSoft provides specific data analysis services to help healthcare providers make efficient and timely decisions and better understand activities. Truven health Analytics provide solution for healthcare data analytics. Artelnies applied artificial intelligence methods to analyze clinical data and provide solutions like medical diagnosis, medical prognosis, microarray analysis and protein design and designed several application for early diagnose and disease prevention, i.e., breast cancer hazard assessment (predict if a patient could have a recurrence of cancer or not), diagnosing heart disease (diagnose a heart problem) and early prognostic in patients with liver disorders (investigating early-stage liver disorder before it becomes serious). XRCI is performing predictive analytics to identify high-risk patients in hospitals and delivering several clinical decision support systems, i.e., ICU admission prediction, complication prediction in ICUs, ICU mortality prediction and stroke severity and outcome prediction.

Kaiser Permanente is one of the first medical organizations to implement an EHR system and is the largest integrated health system serving mainly the western U.S. Kaiser primarily uses SAS analytics tools and SAP's Business Objects business intelligence software to support data analytics activities against the EHR system. Uncertainty of clinical decisions often mislead clinicians to deal with patients (i.e., treatment of newborns using antibiotics: 0.05% newborns having the infection confirmed by blood culture [101] and only 11% of those newborn received antibiotics). Kaiser Permanente system allows clinicians to steer high-cost interventions to deal such issues. System would be able to predict accurately which antibiotics are required for the baby based on the mothers clinical data and the baby's condition immediately after birth that result in cost reduction as well as reduces side effect among newborns. Optum Labs, USA collected EHRs of more than 30 million patients to create database for predictive analytics in order to improve the quality of care. Goal is to use predictive analytics to help physicians to improve patient care by aiding decision power especially for the patients with complex medical histories and suffering from multiple conditions. Healthcare Analytics by Mckinsey & Company combines strategy big data and advanced analytics, and implementation processes.

Prediction of Zika virus has been challenging for public health officials. NASA is assisting public health officials, scientists and communities to limit the spread of Zika virus and identify its causes. Researcher developed computation model based on historical patterns of mosquito-borne diseases (i.e., chikungunya and dengue) to predict the spread of virus and characterized by slow growth and high spatial and seasonal heterogeneity, attributable to the dynamics of the mosquito vector and to the characteristics and mobility of the human populations [102]. Researchers used temperature, rainfall and socioeconomic factors and identified the Zika transmission and predict the number of cases for next year by combining real-world data on population, human mobility and climate.

Harvard Medical School and Harvard Pilgrim Health Care applied analytical methods to EHR data to identify patient with diabetes and classify them into groups (Type 1 and Type II diabetes). Four years worth of data based on numerous indicators from multiple sources have been analyzed. Patient could be grouped into high-risk disease groups and risk could be minimized by preventive care, i.e., new preventive treatment protocols could be introduced among patient groups with high cholesterol. Rizzoli Orthopaedic Institute, Italy, is using advanced analytics to gain more granular understanding of dynamics of clinical variability within families for improving care and reducing treatment costs for hereditary bone diseases. Result showed significant improvement in healthcare efficiency, i.e., 30% reductions in annual hospitalizations and over 60% reductions in the number of imaging tests. Hospital is planning to gain more benefits by insights into the complex interplay of genetic factors and identify cures. Columbia University Medical Center is using analytical technique to analyze brain injured patients to detect complications earlier. To deal with complications proactively rather than reactively, physiologic data form the patient who have suffered form bleeding stroke from a ruptured brain aneurysm is analyzed to provide critical information to medical professionals. Analytical techniques are applied on real-time data as well as persistent data to detect hidden pattern that indicate occurrence of complications. North York general hospital, implemented a scalable, real-time analytics solution to improve patient outcomes and develop deeper understanding of the operational factors driving its business. System provides analytical finding to physicians and administrators to improve the patient outcome.

Intensive Care Unit (ICU) is one of the main section where analytics method could be applied on real-time data for prediction to improve quality care. A number of organizations are working on integration data analytics methods with body sensors and other medical devices to detect plummeting vital signs hours before humans have a clue. SickKids hospital, Toronto, is largest center working on advancing children's health and improving outcomes for infants susceptible to life-threatening nosocomial infections. Hospital applied analytical methods to vital science and other data from real-time monitoring (captured from

monitoring devices up to 1000 times per second) to improve the child health. Potential signs of an infection is predicted before it happens (24 h earlier than with previous methods). Researchers believe that, in future, it will be significantly beneficial for other medical diagnose. University of California, Davis applied analytical methods to EHR to identify sepsis early. QPID analytical system is used by Massachusetts general hospital to predict surgical risk, to help patients with the right course of actions and to ensure that they don't miss critical patient data during admission and treatment.

Even though some of the advanced health organizations have implemented data analytics technologies, however, it still has enough room to grow on analytics.

4.2 Classification

Most important and common use of data analytics and data mining in healthcare probably involves predictive modeling. One key problem is the selection of effective classifier. One general classifier will not work for all types of problems, selection of classifier based on the type of problem and data itself. In the literature, several classification methods is being employed for different disease prediction.

Cardiovascular disease (CVD) can lead to a heart attack, chest pain or stroke due to the blockage of blood vessels. It is one of the leading causes of death (31.5%), whereas 90 % of Cardiovascular diseases are preventable [103]. Lot of research is going on for the early identification and treatment of CVD using data mining tools. Jonnagaddala et al. presented a system for heart risk factor identification and progression for diabetic patients unstructured EHR data [104]. System consist of three modules core NLP, risk factor recognition and attribute assignment module. NLP assigns POS-tags and identifies noun phrases and forwards to risk factor recognition phase that identifies medications, disease disorder mentions, family history, smoking history and heart risk factors. Identified risk factors are then forwards to attribute assignment module assign indicator and time attributes to risk factors. In another study, Jonnagaddala et al. presented rule-based system to extract risk factors using clinical text mining [105]. Risk factor is extracted from unstructured electronic health records. Rules were developed to remove records, which do not contain age and gender information. Alneamy and Alnaish used hybrid teaching learning-based optimization (TLBO) and fuzzy wavelet neural network (FWNN) for identification of heart disease [106]. TLBO is used for training parameter updation used for training FWNN. Training data consists of 13 attributes are forward to FWNN and mean square error is computed that is used for weight updation using TBLO. TLBO-FWNN-based system provided 90.29% accuracy Cleveland heart disease dataset.

Rajeswari et al. [107] presented feature selection using feed forward neural network for ischemic heart disease identification. The feature set is reduced to 12 most discriminant features form 17 that increased the accuracy from 82.2 to 89.4%. In another study, Arslan et al. [108] compared SVM, stochastic gradient boosting (SGB) and penalized logistic regression (PLR) for ischemic stroke identification. SVM provide slightly higher accuracy as compared to SGB and PLR. Anooj [109] used weighted fuzzy rule-based system based on Mamdani fuzzy inference for the diagnosis of heart disease. Attribute selection and attribute weighted method is used for the development of fuzzy rules. Mining procedure is applied to generate appropriate attributes which are use to develop fuzzy rules. Weighted based on frequency is added to attributes in the learning process for effective learning.

As discussed in Sect. 3, decision tree is one of the best methods and is extensively used for disease prevention. [44, 46, 110–113]. NaliniPriya and Anandhakumar [113] extended the decision tree to handle multivariate datasets. Choudhary and Bajaj [112] utilized *k*-NN and decision tree for the prediction of root canal treatment needs. When there is little change in ECG, Alizadehsani et al. [114] presented a system to find a way for specifying the lesioned vessel based on para clinic data, risk factor and physical examination. C4.5, Naive Bayes, and KNN methods are used for coronary arteries stenosis identification on set of 303 random visitor (Z-Alizadeh Sani dataset) with additional effective features (function class, dyspnoea, Q wave, ST elevation, ST depression and t inversion) and no missing value. C4.5 achieved best accuracy 74.20% for left anterior descending, 63.76% for left circumflex and 68.33% for right corona C4.5 and Bagging algorithms [115]. The system achieved 79.54%, 61.46%, and 68.96% left anterior descending, left circumflex and right coronary artery respectively. Karaolis et al. [116] investigated three events myocardial infarction (MI), percutaneous coronary intervention (PCI), and coronary artery bypass graft surgery (CABG) using C4.5 decision tree algorithm on 528 cases. Based on three event using five splitting criteria (age, smoking, hypertension history, family history), most important risk factors extracted are age, smoking, and hypertension history for MI, family history, hypertension history, and history of diabetes for PCI and age, hypertension history, and smoking for CBAG. The system achieved accuracy 66%, 75%, and 75% for MI, PCI, and CABG models, respectively. Alizadehsani [117] compared different classifier SMO, Naive Bayes, neural network for identification of coronary artery disease on Z-Alizadeh Sani dataset, whereas weighted SVM is used for feature selection. Set of 34 of are selected which had the weight

higher than 0.6. The study showed that SMO provides better accuracy 94.08% as compared to 75.51%, 88.11% by Naive Bayes and neural network, respectively. El-Bialy et al. applied fast decision tree and pruned C4.5 tree for the classification of coronary artery disease. Missing, incorrect, and inconsistent data problems is resolved using integration through integration the different datasets [118]. The strategy is to select attributed form different dataset for same medical problem. In this study, four different data coronary artery dataset (Cleveland, Hungarian, V.A. and statlog project heart dataset) are integrated. C4.5 and fast decision tree is applied on each dataset and five common features are selected form all decision tree. These five most common attributes are then used to build new integrated dataset that consist of only selected common features. Result showed that new integrated dataset provided gain in accuracy 78.06% from 75.48%. Fuzzy rules along with decision tree is also used for disease risk prediction [119, 120]. Fuzzy rules with decision tree are applied to overcome problems associated with uncertainty. Kim et al. [119, 121] used fuzzy with decision tree. *Cerebrovascular risk* factor are elder age, hypertension, heart disease, diabetes mellitus, temporal cerebral ischemia seizure and cerebrovascular disease history, and subordinate risk factors: hyperlipidemia, obesity, polycythemia, smoking, drinking, family heredity, oral contraceptive and other medicine. Elder people are vulnerable to cerebrovascular disease. Yeh et al. utilized decision tree, Bayesian classifier and back propagation neural network for the prediction of cerebrovascular disease [122]. In total 29 variables (9 out of 24, 12 out of 29 and 8 out of 10 for physical exam, blood test [101, 123, 124] and disease diagnosis, respectively) are selected based on clinician recommendation. System provided accuracy/sensitivity (%) 98.01/95.29, 91.30/87.10, 97.87/94.82 for decision tree, Bayes and BPNN, respectively, on 493 patients. Study identified eight major factors (diabetes mellitus, hypertension, myocardial infarction, cardiogenic shock, hyperlipidemia, arrhythmia ischemic heart disease and BMI) for accurately predicting cerebrovascular disease. Classification accuracy/sensitivity (%) is increased to 99.59/99.48 based on 16 extracted diagnose rules. Adnan et al. presented hybrid approach using decision Tree, Naive Bayes for Childhood Obesity Prediction [125]. CART is used to select eight important variables. Outputs of hybrid method is clustered into the positive and negative groups.

Diabetes is one of the major health problems in all over the world. It is quite serious disease that can lead to serious complication if not treated on time and there are several other diseases are related to diabetes. The cause of diabetes is mysterious, and it is due to the low insulin production by pancreas or body does not respond to produced insulin properly. As it is lifelong disease, thus, even for individual patient, massive amount of data is available to interpret [126]. Breault et al. applied CART on 15,902 diabetes patients and concluded that age is the most important attribute associated with bad glycemic control [127]. Using the age information, clinician can target specific set of people. Temporal abstraction method can be integrated with data mining algorithms to support data analysis. In another study, Yamaguchi et al. [128] used data forest software for type 1 diabetes mellitus prediction and predict next-morning FBG based on FBG, metabolic rate, food intake, and physical condition. The study concluded that physical conditions are highly correlated with FBG. Complex temporal abstractions are used for diabetes mellitus and blood glucose management [126, 129]. Cho et al. [130] compared different classification methods (logistic regression, SVM, and SVM with a cost sensitive learning method) for diabetic nephropathy and several feature selection methods have been used to remove redundant features. Study showed that linear SVM classifiers with embedded feature selection methods showed the promising result. Huang et al. applied Naive Bayes, IB1 classifier, and CART on 2,064 patients dataset and identify five important factors (age, insulin needs, diagnose duration, diet treatment and random blood glucose) that effect blood glucose control. Using these five attributes, system achieved 95% accuracy and 98% sensitivity [131].

Parkinson disease (PD) is the multisystem neurodegenerative disorder affecting the motor system which result in dopaminergic deafferentation of the basal ganglia, gives rise to characteristic motor disturbances that include slowing of movement, muscular rigidity, and resting tremor [132]. It is the second most common movement disorder. There is no medical treatment [133, 134]; thus, PD patients have to rely on clinical intervention. Das compared different classifiers (DMneural, neural Networks, regression tree and decision Tree) on Max Little dataset [135] for effective diagnosis of Parkinson's diseases. Study showed that neural network obtained best result 92.9% classification accuracy. In an other study, Geetha compared several methods (KNN, SVM, C4.5, Random forest tree Classification and regression Tree, partial least square regression, linear discriminant analysis, etc.) for the classification of PD patient on Max Little dataset [136]. Study showed that Random Tree classifier yields the 100% accuracy. Khan used KNN, AdaBoost and Random forest for the identification of PD patient on Max Little dataset [137]. Result showed that k-NN provided best accuracy 90.26% using $k = 10$ fold validation. Suganya and Sumathi [138] used metaheuristic algorithm for the detection and classification of Parkinson disease. Study reported that ABO algorithm provided best accuracy (97%) as compare SCFW, FCM, ACO and PSO.

End-stage hemodialysis patient care cost is too high thus reducing cost of dialysis is important factor. On regular basis, more than 50 parameters are monitored in kidney dialysis treatment and there are multiple factor that can influence patient survival [139]. To make mining results more likely comprehended, domain knowledge prior to data analysis could be added by combining the temporal abstraction method. Yeh et al. used temporal abstraction with data mining techniques (decision tree and multiple minimum supports association rule) for analyzing biochemical data of dialysis patients to develop a decision support system to find temporal patterns resulting in hospitalization [140]. Association rules and the probability patient hospitalization is used to decrease patient hospitalization. Monthly biochemical test data is transformed to basic TA through the basic TA algorithm that is used as a input to complex TA algorithm to transform into complex TA.

Kusiak et al. used two different methods for knowledge extraction in the form of decision-making rules that are used to predict patient survival [139]. Identifying pressure ulcer patients at early stage is debilitating complication. Raju et al. [141] compared the performance of different classifier (logistic regression, decision trees, random forests, and multivariate adaptive regression splines) risk factor associated with *pressure ulcers*. Decision tree was split based on mobility sub-scale value of 2.5. Compared to the other methods, random forests provide better accuracy.

Stroke ranked third in disease burden and is considered to be the leading causes of hospitalizations, disability and death. Stroke risk and severity prediction can contribute significantly to its prevention and early treatment. Literature shows that there are few post-stroke predictive models and wide range of different techniques has been applied for risk identification related to a particular condition. Strokes factors are age, smoking, diabetes, anti-hypertensive therapy, cardiovascular disease, systolic blood pressure, and left ventricular hypertrophy by electrocardiogram [142]. Khosla et al. [143] presented automatic feature selection method that selects robust features and evaluated three automatic feature extraction approaches forward feature selection (FSS), L1 regularized logistic regression (RLR), and conservative mean feature selection (CM). FSS greedily adds one feature at a time and cross-validation is used to select the best features. CM select relevant and robust to variations features using conservative mean by considering monotonic prediction functions over a single feature. SVM and margin-based censored regression (MCR) is used for classification propose. Study showed that MCR provided better accuracy on set of features selecting using conservative mean method. Sung et al. [144] applied MLR, *k*-nearest neighbor and regression tree on 3577 patients (approximately one-third are prior stroke patient) with acute ischemic stroke. Three-step feature selection (frequency cut-off operation, correlation-based feature selection and expert opinion) process was used and study identified seven predictive features. Panel of stroke neurologists identified the final set of features by consensus. Study identifies stroke severity indices, which represent proxy measures of neurologic impairment that could be used for ischemic stroke patient to adjust strokes severity. Easton et al. [145] developed post-stroke mortality predictive model for or very short term and short/intermediate term mortality. Study identified certain risk factors differentiated between very short term and intermediate term mortality, i.e., age is highly relevant for intermediate term. He et al presented a framework (D-ECG) for accurate detection of cardiac arrhythmia [89] that develop the result regulator using different set of features in order to refine the result.

Liver Disorder also called hepatic diseases is disturbance of liver function that causes illness, normally does not cause any obvious symptoms until liver is damaged and disease is advanced. If it is detected at early stage, acute liver failure caused by an overdose of acetaminophen can sometimes be treated and its effects reversed. Seker et al. applied KNN, SVM, MLP and decision trees. Baitharu and Pani applied decision trees J48, Naive Bayes, ANN, ZeroR, 1BK and VFI algorithm to identify liver disorder [cirrhosis, bile duct, chronic hepatitis, liver cancer and acute hepatitis from liver function test (LFT)] [146]. *Hepatitis* Yasin et al. extracted top seven PCA features (steroid, antiviral, fatigue, malaise, anorexia, liver big and liver firm) form set of 19 obtained using PCA [147]. Zayed et al. [148] used decision tree (C4.5) to predict therapeutic outcome to antiviral therapy in HCV patients.

MERS-CoV is viral airborne disease caused by novel corona-virus (MERS-CoV) which spreads easily and has a high death rate. It developed severe acute respiratory illness, including fever, cough, and shortness of breath. Yang et al. [149] presented non-orthogonal decision trees for mining SARS-CoV protease cleavage data. Bio-mapping is used to transform the *k*-mers to a high-dimensional that is used for construction of decision tree. Prediction accuracy is significantly improved by using bio-mapping-based high-dimensional templates and non-orthogonal decision tree. Lee et al. [150] used SVM (normal, sigmoid, RBF and polynomial) to predict SARS virus and obtained 97.43% accuracy using polynomial. Sandhu et al. [151] presented Bayesian belief network to predicts MERS-CoV-infected patients as well as geographic-based risk assessment. Possibly MERS-CoV infected users are quarantined by using GPS and marked on Google map.

Seminal quality prediction is binary classification problem, and it is quite useful for early diagnosis of seminal disorder and also helps for selection of donor etc. Wang et al. [152] presented three-stage ensemble learning framework called clustering-based decision forests, to tackle unbalanced class learning problem for the prediction of seminal quality. First stage deals with the class imbalancing (Tomek Links method is used for data cleaning) and at second stage, balanced dataset is formed by combining majority subset and bagging minority. Finally, random subspace is used to generate a diverse forest of decision trees. Comparison result on UCI fertility dataset showed clustering-based decision forest outperforms decision tree, SVM, MLP, logistic regression and random forest. To identify lifestyle and environmental features that affect the seminal quality and fertility rate, Sahoo and Kumar [153] applied MLP, SVM, evolutionary logistic regression, PCA, SVM+PSO, chi-square, correlation and *T*-test. To evaluate the seminal quality and fertility rate, decision tree and Naive Bayes, SVM, SVM+PSO and ML is applied on UCI fertility dataset. Result showed that SVM+PSO provides better accuracy as compared to other classifier and also concludes that age, season, surgical interventions, smoking alcoholic habit have more impact on the reduction of fertility rate. In another study, Gil et al. [154] used SVM, MLP and decision tree to predict seminal quality environmental factors and lifestyle data. Naeem [155] used Bayesian belief network (BBN) on UCI fertility dataset to identify effecting parameter. Girela et al. [156] presented MLP based study on data collected by the questionnaire to predict the results of the semen analysis. Sahoo et al. [153], Naeem [155], Gil et al. [154] and Girela et al. [156]studies showed that life style, environmental features and health status strong effect the semen quality and could be used as a measure for early diagnose male fertility issues. *IVF* is one of the most effective and expensive treatments for addressing infertility causes that requires selection of healthy embryos and continuous patient observation. Predictive modeling could help to reduce the chance of multiple pregnancy and increase the success rate by finding the effecting parameters. The first study on IVF outcome prediction was presented in 1997 by Kaufmann et al. [157] using neural network. Corani et al. [158] and Uyar et al. [159] used Bayesian network for predicting pregnancy after IVF. Corani et al. conclude that embryo viability (top graded) is a more critical factor than uterus and helps to decide how many and which embryos to transfer. Simulation result on 5000 IVF cycles shows that decision-based transfer increases both single pregnancy and double-pregnancy rate; moreover, it also reduces the average number of transferred embryos from 3 to 2.8. In another study, Uyar et al. compared different classifier Naive Bayes, decision tree, SVM, KNN, MLP and radial bases function network (RBF) classifier on dataset of 2275 instances (15 variables) in order to discriminate embryos according to implantation potentials [160]. Study reported that Naive bayes and RBF perform better as compared to other methods. Siristatidis et al used neural network based on LVQ to predict outcome early [161] and presented a web-based System to predict IVF outcome [162]. Durairaj and Kumar used MLP [163] and rough set theory [164] to identify influential parameters that effect the success of IVF outcome [163]. Fertility analysis is performed on 27 different attributes of 250 patients. Gvenir et al. presented RIMARC ranking algorithm to predict the chance treatment as probability of success for IVF treatment on dataset of 1456 patients [165]. It assign the score based on the past cases. Results are compared with random forest and Naive Bayes. BPNN [166], SOM [167] based approach to predict IVF treatment result. As in IVF, the success rate is quite low, thus selection of appropriate parameter, precondition, post condition, embryo viability could help to improve the chance of conception. Intelligent techniques could help to increase the chance of conception in IVF treatment. Nevertheless, more efforts are required to improve the efficacy. Moreover, the studies done so far are based on small dataset, that is not sufficient for efficient prediction.

4.3 Clustering

Cluster analysis has been widely applied for such various applications especially in preventive healthcare that helps to reveal hidden structures and clusters found in large datasets. Availability of extensive amount of healthcare data requires powerful data analysis tools to come up with meaningful information. Along with classification techniques, it has been extensively used in healthcare sector and numerous CA applications especially disease prevention have been reported in the literature, such as characterizing diabetic patients on the basis of clusters of symptoms; grouping cardiovascular patients based on their symptom experience and identification of cause of mortality etc.

k-means and its variations have been extensively used for disease prediction. Luo et al. [168] performed cluster analysis for prevention and treatment experience of infectious diseases (influenza, dysentery, tuberculosis, viral hepatitis, and other infectious diseases) from famous herbalist doctors. Study showed that cluster analysis is helpful to summarize the common understanding of experience including concept of syndrome differentiation, regulation of diagnosis and treatment, prescription characteristic that is based on prevention and treatment of influenza, viral hepatitis and other infectious diseases. Association rules are used to find the relationship between etiology, syndrome, symptoms and herbal prescription of

infectious diseases. In another work, Kabir et al. [169] presented the system to detect epileptic seizure from EEG signal using SVM, Naive bayes and logistic regression. Results showed that *k*-mean clustering can handle EEG data efficiently in order to detect epileptic seizure. Kaushik et al. [170] applied enhanced genetic clustering to diagnose and predict diseases based on patient history, symptoms, and existing medical condition. New patients' profile is analyzed against existing patterns and mapped to a particular cluster based on the patients medical parameters followed by prediction of medical condition and probability of being prone to that disease in future. Vijayarani and Sudha [171] applied Fuzzy *C*-means, *k*-means and weight based *k*-means algorithm to predict diseases from hemogram blood test samples. Study result showed that weighted *k*-mean performed better as compared to fuzzy *c*-mean and *k*-means. Norouzi et al. [172] used fuzzy *c*-mean clustering to predict the renal failure progression in chronic kidney based on weight, diastolic blood pressure, diabetes mellitus as underlying disease, and current. The number of fuzzy rules is equal to the number of membership function of input variable. Feng et al. applied *K*-center clustering method to analyze clinical data and syndrome information of 154 cases of prethrombosis state. Study showed that traditional clinical syndrome differentiation presented 12 syndrome patterns (blood stasis, qi deficiency, damp turbidity, yin deficiency, yang deficiency, phlegm turbidity, damp-heat (toxicity), qi stagnation, blood deficiency, phlegm heat, and cold accumulation) [173]. Result showed that blood stasis and qi deficiency are more commonly seen (49.1%) syndromes than other, whereas cold accumulation is the most rarely seen syndrome. Yang et al. [174] performed fuzzy cluster analysis of Alzheimer's disease-related gene sequences. Study result indicates that the gene sequences interrelated within one group is consistently having closer relationship within the group other than in another group. Shaukat et al. [175] compared *k*-means, *K*-medoids, DBSCAN and OPTICS to determine the population of dengue fever infected cases. Results showed that OPTICS outperformed other methods.

As healthcare data is very sensitive thus to overcome the issue of sensitivity, recently hybrid approaches have been considered especially clustering is being used along with classification techniques. Shouman et al. [176] presented hybrid method and applied Naive Bayes and *k*-mean clustering with different initial centroid selection method for the heart disease patient diagnosis. Integration of unsupervised (*k*-mean) and supervised (Naive Bayes) with different initial centroid selection enhanced the prediction accuracy of Naive Bayes as compared to standalone method. For centroid selection different methods (range, inlier, outlier, random attribute values, and random row methods) have been applied. In another work, Zheng et al. [177] presented hybrid of *k*-means and support vector machine (*k*-SVM) algorithms for breast cancer diagnosis. *K*-means is used to recognize the hidden patterns of the benign and malignant tumors separately and membership of each tumor to these patterns is calculated and treated as a new feature for SVM training. Study result showed that *k*-SVM improve the accuracy to 97.38% on WDBC dataset. PCA and LDA are used for feature reduction followed by LS-SVM, PNN and GRNN.

4.4 Association analysis

Association rule mining is to detect factors which contribute to disease outbreak. Moreover, association is also used with classification techniques to enhance the analysis capability. Relational analysis could help to make healthcare data more useful as occurrence of one disease can lead to several other associated diseases. Association rule mining (ARM) has been applied to find link between other associated disease especially cardiac, diabetes, cancer and various form of tumors etc.

Ogasawara et al. performed association analysis to find the relationship between lifestyles, family medical histories and medical abnormalities [178]. Different variable of lifestyle variables (overweight, drinking, smoking, meals, physical exercise and sleeping time), medical history attributes (hypertension, diabetes, cardiovascular disease, cerebrovascular disease, and liver disease) and medical abnormalities (medical abnormalities namely high blood pressure, hypercholesterolemia, hypertriglyceridemia, high blood sugar, hyperuricemia, and liver dysfunction) from 5 year data of 5350 patient were considered for analysis and 4371 rules were extracted. Study concludes that association analysis performed significantly better to predict the risk factor than conventional modeling. Semenova et al developed association rule algorithm with focus on delivering knowledge from large database [179]. Instead of focusing on frequent itemsets, they have considered itemset that provides knowledge and useful insight. Polydict algorithm is used to find frequent itemset patterns. Initial dictionary stores all the variety of sequences with their counts that make search frequent patterns.

Doddi et al. [180] presented random sampling and Apriori to obtain association rules indicating relationships between procedures performed on a patient and the reported diagnoses analyze on large database containing medical record data. Due to the large dataset, random sampling is used to overcome the computation issue. Small sample is collected form large collection of transaction using random sampling and Apriori algorithm was applied on small collected data. Transaction set is partitioned into five disjoint sets each consisting of about 250K transactions and random sampling size of 5K is generated form each

subset of transaction and Apriori algorithm is applied on these selected sample of each subset. Payus et al. used Apriori algorithm for respiratory illness caused by air pollution. Six attributes were considered and 42 rules were generated with minimum support of 0.1 and minimum confidence of 0.1 [181]. Concaro et al. [182] presented temporal association rule mining (TARM) for analysis of care delivery flow of diabetes mellitus by extraction of temporal associations between diagnostics and therapeutic treatments and Apriori variant is used for mining ARs. In another work by Concaro et al. [183], they have applied TARM for analysis of costs related to diabetes mellitus. Jabbaret al. [184] presented heart attack prediction using Boolean matrix (HAPBM) algorithm. HAPBM is a variant of Apriori algorithm with difference in conversion of discretized dataset into Boolean matrix, and then frequent itemset generation from boolean matrix. Raheja and Rajan [185] comparatively analyzed ARM and MiSTIC (Mining Spatio-Temporally Invariant Core Regions) approaches for extracting spatio-temporal of Salmonellosis disease occurrence pattern. Apriori algorithm is used for ARM with minimum support and minimum confidence values of 1 and 40 percent respectively. For efficient rules, economy, demographics and environmental data is also used and Apriori is applied on consolidated dataset. Lee et al. [186] utilized the variant of Apriori algorithm, and standard support-confidence framework for frequent itemsets and ARs generation for discovering medical knowledge from acute myocardial infraction (AMI) patients. Pruning was performed by using lift, leverage, and conviction interestingness measures. Twelve risk factors related to blood factors were selected out of the total 141 risk factors, to find associations between blood factors and disease history in young AMI patients.

Nahar et al. [187] applied Apriori, Predictive Apriori and Tertius for rule generation to investigate the sick and healthy factors which contribute to heart disease for males and females. Attributes indicating healthy and sick conditions were identified. Two experiments (rules to healthy and sick conditions, rules based on gender) have been performed. All sick individuals were regarded to be in one class and healthy individuals to be in another class and Apriori, Predictive Apriori and Tertius is applied. Payus et al. [181] mined air pollution database for extracting reasons behind respiratory illness. Seven attributes were selected and Apriori algorithm was applied with minimum support and minimum confidence value of 0.1. In total, 42 rules were generated (17 were normal hospitalized patients, 24 belonged to moderate hospitalized patients and 1 rule belonged to high hospitalized patient). Anwar and Naseer [188] presented 69 association rules to identify lifestyle and environmental factors on man's seminal fertility and quality. ARs were mined using Apriori algorithm based approach by using XLMiner tool. Experiment and analysis was performed with multiple support and confidence values, and reported to have found interesting association rules. Sharma and Om used applied Apriori algorithm with minimum support value 10% and minimum confidence value 90% [189]. ARM was applied on patients clinical examination and history data for oral cancer detection. Huang presented data cutting and sorting method (DCSM) rather than Apriori algorithm that reduces the time to scan immense sizes database, i.e., health examination and outpatient medical records [131].

Berka and Rauch [190] applied meta-learning techniques to association rules in the atherosclerosis risk domain. They applied ARM to the ARs extracted from atherosclerosis data to generate rules about rules for more effective comprehension. To digitize, Apriori is used and then extracted ARs were post-processed to the rules extracted in prior step by application of ARM. McCormick et al. [191] presented hierarchical association rule mining (HARM) technique to predict sequential events. HARM is applied for automated symptom prediction. Result showed that predictive performance of HARM was the highest with partially observed patients, than with new patients. The idea is based on the history of the patient, the next symptom to be faced by patient could be predicted. Srinivas et al. [192] presented DAST algorithm for mining frequent as well as rare itemsets and positive and negative ARM for disease prediction.

Association rules mining algorithms may results in generation of extremely large number rules, especially in case low minimum support and minimum confidence values, thus reduction technique is necessary for effective analysis. To discover significant association rules, summarize rules having the same consequent and accelerate the search process, Ordonez et al. [193] applied greedy algorithm. Rules were constrained by using association rule size threshold, restriction of itemsets for appearance, restriction of itemset combinations appearance and lift measure usage. For evaluation purpose, 25 out of 113 attributed for 655 individual are used minimum support value of 1%, minimum, confidence 70%, maximum rule size of 4, minimum lift value of 1.20, and minimum lift value of 2.0 for covers rules. Ohsaki et al. [194] performed detailed analysis of practicability of rule interestingness measures. In total 40 interestingness measures were analyzed, the accuracy, uncovered negative, peculiarity, relative risk and chi-square measures are the interestingness measures useful from medical domain experts point of view. Moreover, medical knowledge discovery from databases could be advanced by utilization of other interestingness measures by medical domain experts. Kuo et al. [195] approach reduces the number of generated association rules by considering domain relevant rules selected by

domain expert using. To deal with rule selection, domain expert is required.

Soni and Vyas [196] integrated association rule mining and classification rule mining. Integration is done by focusing on mining a special subset of association rules and then classification is being preformed using these rules. Rules are advanced and different associative classifier (weighted associative classifiers, positive and negative rules, fuzzy and temporal) are used to increase accuracy. Later on, Soni and Vyas [197] extend their work and used weighted associative classifiers for heat disease prediction and presented intelligent heart disease prediction system (IHDPS) using weighted associative classifier and study reported that IHDPS achieved the highest average accuracy as compared to other systems. Concaro et al. [183] presented a method to integrate of both clinical and administrative data and address the issue of developing an automated strategy or the output rules filtering, exploiting the taxonomy underlying the drug coding system and considering the relationships between clinical variables and drug effects. Rajendran and Madheswaran [198] presented hybrid association rule classifier (HARC) based on ARM and decision tree (DT) algorithm. Transactional database containing texture features data was mined using frequent-pattern tree (FP-Tree)-based ARM. Moreover, DT is applied to transactional database instances for each of the rules generated in ARM step. In another work, Shirisha et al. [199] presented Heparin induced thrombocytopenia algorithm based automated weight calculation approach for weighted associative classifier. Based on the different attributes, weights are assigned. Itemsets that are already found to be frequent are added with new items based on the algorithm. Chin et al. [200] presented framework for early RA assessment that integrated efficient associative classifiers to mine RA risk patterns from a large clinical database. Risk patterns that occur frequently, and classify the disease status are extracted. Weighted approach to integrate correlation and popularity information into the measure is used to prevent the ignorance of important rules that the objectivity of the results.

4.5 Anomaly detection

Today, electronic medical records contain vast amount of patient information regarding his conditions along with treatment and procedure records. Anomaly detection in disease prevention and health promotion typically work with patient records and it has high potentials in mining hidden pattern and forecasting of disease. EHR have anomalies due to several reasons such as physician error, instrumentation error and patient condition [91]. One particular advantage health organizers look is hot spotting of data in a timely manner. Thus, it is a key task to support disease prevention and requires high degree of accuracy. Typical anomaly detection approaches focus on detection of data instances that deviate from the majority of data. So far, several anomaly detection techniques have been presented for the detection of disease outbreaks. Beside its critical need, anomaly detection in health record is a challenging task due dataset complexity (non standard and poor quality data, privacy issue etc). Supervised anomaly detection method requires complete labeled data for both anomaly and normal data. However, in healthcare, it is rarely possible to acquire complete labeled data (as labeling is done manually by expert that require substantial efforts; moreover, may occur they can occur spontaneously as novelties at run time). The other issue is data itself, there are few anomalies as compared to normal one. Supervised method is not a preferable choice for anomaly detection; thus, we are not focusing it.

Due to the domain specific normality and unavailability of labeled data, most of the existing studies are based on unsupervised or semi-supervised methods [201] and mainly focus on disease specific anomaly detection; however, there are some generic studies as well. Hu et al. [202] presented generic framework for utilization analysis that can be applied to any patient data. Carvalho et al. [201] presented an effective and generic method for anomaly detection-based score. In the first phase, degree of anomaly is calculated by performing anomaly analysis. Three proximity-based algorithms (KNN, reverse KNN and Local Outlier Factor) are used to for anomaly detection. Detection is based on two steps: score assignment and transference. In first phase, outlier analysis is performed to calculate the degree of anomaly and second phase, score transference assign score. Typical outlier detection methods identify unusual data instances that deviate from the majority of examples in the dataset. Mezger et al. [203] used logistic regression model to detecting deviations from usual medical care. In another work, Hauskrecht et al. [204] presented statistical anomaly detection (evidence-based anomaly detection). Bayesian networks are used as probabilistic models to compute statistics.

Goldstein and Uchida evaluated 19 different unsupervised methods on 10 different datasets from multiple application domains. Lie et al. used three different methods for anomaly detection standard support vector data description, density induced support vector data description and Gaussian mixture on UCI dataset for liver disorder.

Several rule-based anomaly detection haven been developed such as drug-allergy checking, automated dosing guidelines, identifying drug-drug interactions, detecting potential adverse drug reactions, detection of even in chronic condition, i.e., congestive heart failure, diabetes etc.

It is more important in anomaly detection to assure that reported anomalies to a user are in fact interesting. Traditional anomaly detection methods are unconditional that look for outlier with respect to all data. Conditional anomaly detection is a relativity new approach that detects unusual data for a subset of variables. Valko et al. presented instance-based methods for detecting conditional anomalies for two real-world problems (unusual pneumonia patient admission and unusual orders of an HPF4 test) that rely on distance metrics to identify example that are most critical for anomaly detection [205]. To assess the conditional anomaly and analyze the deviations of example a for anomaly detection, one dimensional projection is used. A case is considered to be anomalous if the value (output–input) falls below certain threshold. In another study, Hauskrecht et al. [206] presented conditional-outlier detection method to identify unusual patient management actions (omissions of medication or laboratory orders) based on patent's past data. Rather than only identification of unusual data instance deviation, this approach identify outliers where individual patient management actions strongly depend on patient condition. SVM is used for predicting each type of action, i.e., medication and laboratory orders. Valko et. al. presented non-parametric approach based on the soft harmonic solution for anomaly detection. In an other work, Hauskrecht et al. [207] presented framework for post-cardiac surgical patient anomaly detection that detects conditional anomalies and raises an alert when an anomaly is found. Hong and Hauskrecht [208] presented multivariate conditional model that consider both context and correlations for outliers identification. A tree structured model, multi-dimensional ensemble frameworks [209] is used to build multivariate conditional model. Posterior of individual and posterior joint responses are computed using the decomposable structure of the multi-dimensional models. In order to compute the outlier score for each instance, the data is transformed from original space to d dimensional conditional probability space.

Conventional PCA is used an indicator of the existence or absence of anomalies [120]; however, its disadvantage is the limited effectiveness for small anomaly. Multivariate cumulative sum (MCUSUM) has ability to effectively detect small anomalies with high cross correlated variables and used as alternative to PCA. Harrou et al. [210] integrate PCA and multivariate cumulative sum(MCUSUM) for anomaly detection with high sensitivity to small anomaly. MCUSUM is applied to the PCA output to detect anomalies when data did not fit wit PCA. Evidence showed that PCA-MCSUM integration provide better fault detection on pediatric emergency dataset.

Lee et al. used multilayer perceptron for detection of cardiac anomalies (ventricular tachyarrhythmia, congestive heart failure, malignant ventricular ectopy, supraventricular arrhythmia) form normal cardiac status. Features are extracted form ECG using PCA and histogram of gradient (HOG) [211].

Pattern mining techniques are used to extract frequent patterns and comparison is performed between frequent pattern and domain knowledge for anomaly detection [212]. Frequent-pattern matching identify frequent pattern that deviates from guidelines. It also identifies anomaly that deviates from frequent patterns followed by the comparison between the domain knowledge of disease and the frequent patterns of treatment of that disease. Thus, it allows two types of anomalies detection. The anomalies that comprises the anomalous cases deviate from the frequent patterns and frequent patterns that deviate from the accepted guidelines.

5 Datasets

PhysioBank databases is a large and growing archive of physiologic data [213]. it consists of over 90,000 recordings, or over 4 terabytes of digitized physiologic signals and time series, organized in over 80 databases. PhysioBank archives consist of clinical database, waveform database, multi-parameter database, ECG database, interbeat interval database, other cardiac databases, computing in cardiology challenge datasets, synthetic data, gait and balance databases, neuroelectric and myoelectric databases. EEG, EHG, and more. Each database is placed into a class (completed reference database, archival copies of raw data and other contributed collections of data). Further details of PhysioBank are available at [213].

Breast cancer wisconsin diagnostic (WDBC) Features of the breast cancer dataset are computed from digitized breast mass images of a fine needle aspirate (FNA) describing the characteristics of the cell nuclei [214]. Dataset consists of 32 attribute (30 real-valued input features) of 569 instances (357 benign, 212 malignant). The task of the UCI dataset is to separate cancer from healthy patients. Wisconsin Prognosis Breast Cancer (WPBC) consists of 34 attributes (32 real-valued input features) of 198 instances (151 nonrecur, 47 recur). Each record represents follow-up data for one breast cancer case. *Cardiovascular dataset* Cardiovascular Heart Study (CHS) is a study of risk factors for development and progression of CHD and stroke for cardiovascular diseases in people over the age of 65. It is one of the most widely used benchmark datasets for cardiovascular disease risk factors including stroke [215]. It consists of 5201 instances with significant fraction of missing values and a large number of features. Another dataset, Post-surgical cardiac patients dataset consists of 4486 patients collected from archived EHRs at a

large teaching hospital in the Pittsburgh area [216]. Another multivariate cardiac disease dataset by UCI consists of 76 attributes (categorical, integer, real) and 303 number of instances. The purpose of dataset is to figure-out presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Congenital heart disease (CHD) consists of 30-day outcomes (alive or dead) for congenital heart disease treatment. Figures are provided for patients aged under 16 (pediatric) or over 16 (adult congenital heart disease).

Thyroid disease dataset is another dataset from UCI machine learning repository in the medical domain also known as the annthyroid dataset. It is suited for ANN training having 3 classes (normal: not hypothyroid, hyperfunction and subnormal functioning) 3772 training instances, 3428 testing instances and consist 15 categorical and 6 real attributes. Raw patient measurements contain categorical attributes as well as missing values. For outlier detection, 3772 training instances are used, with only 6 real attributes. Hyperfunction class is treated as outliers class and other two classes are inliers. An other thyroid dataset collected from UCI repository, is binary class either thyroid our non-thyroid class dataset that consist of 7547 instances (776 and 6771 belong to thyroid and non-thyroid respectively) and 29 attributes (mostly numeric or Boolean). It contains both hypothyroid and hyperthyroid data. The raw patient measurements contain categorical attributes as well as several missing values. Another thyroid dataset gathered from Imam Khomeini hospital, consists of 28 attributes. Another dataset released by Intelligent System Laboratory of K. N. Toosi University of Technology consists of 1538 patients of 21 features (sex, age, T3, T3RU, T4, FT4, TSH, palpitation, drowsiness, exophthalmia, diarrhea, constipation, edema, menstruation, diaphoresis, heat intolerance, cold intolerance, weight change, appetite, tremor, and nervousness) each, 15 binary (from x1, x8 x21) and 6 continuous (x2, x3, x4, x5, x6, x7) [55]. It consists of 331 patients belong to Hyper class, 648 patients belong to Hypo class and 559 of them belong to Normal class. *Diabetes* data is multivariate dataset consist of 75 attributes of 100K instances with missing values [217]. It has been prepared to analyze factors related to readmission as well as other outcomes pertaining to patients with diabetes [218, 219].

Diabetes dataset extracted form health facts national database of 130 US hospital to analyze factors related to readmission as well as other outcomes pertaining to patients with diabetes [220]. It consists of 55 (integer) attributes of 100K instance with missing value. *Fertility* dataset consist of 10 real value attributes of 100 instances (18–36 years old) released for classification and regression purpose [154]. Sperm concentration are related to socio-demographic data, environmental factors, health status, and life habits. *IVF* dataset of 1456 patients released by VF Unit at Etlik Zubeyde Hanim Womens Health, Teaching and Research Hospital, Ankara, Turkey that consist of demographic and clinical parameters, as independent features (52 out 64 are related to the female, and 12 are related to the male) [165]. The dataset consists of 43 categorical values features, 21 numerical features and 13.5 % missing values feature. *Hepatitis* UCI hepatitis dataset consist of 155 samples with 19 features (13 binary and 6 are discrete) [221]. The purpose of dataset is to predict the presence of hepatitis virus given the results of various medical tests carried out on a patient. *Liver* ILPD (Indian Liver Patient Dataset) consist of 10 (integer and real) attribute of 583 (416 and 167 liver and no liver patient respectively) instances for classification purpose [222]. *Heparin induced thrombocytopenia* (HIT) dataset collected from 4273 records of post-surgical cardiac patients treated [223].

6 Open research issues and future directions

Data mining and data analytics in the development of preventive healthcare applications have tremendous potential; however, the success hinges on the availability of quality data, but there is no magic recipe to successfully apply data analytics methods in any healthcare organization. Thus, for the successful development of prevention application depends upon how data is stored, prepared and mined. However, healthcare analytics poses a series of challenges when dealing with mound of complex healthcare data. These challenges involved data complexity, access to data, regulatory compliance, information security and efficient analytics methods to successfully analyze this data in a reliable manner.

6.1 Monitoring and incorporating multi-source information

The main goal of preventive healthcare analysis is to take real-world medical data to help in disease prevention. Thus, the successful development of preventive application depends upon how data is stored, prepared and mined. In health sector, we do have large volumes of complex, heterogeneous, distributed and dynamic data coming in, i.e., in US only, healthcare data reached 150 exabytes in 2011 and expected to reach zettabyte scale soon. Despite the rapid increase in EHR adoption, there are several challenges around making that information useful, readable and relevant to the physicians and patients who need it most. One of the key challenges in healthcare industry is how to manage, store and exchange all of this data. Interoperability is considered as one of the solutions to this problem. Another challenge is data privacy that limits the

sharing of data by blocking out significant patient identification information such as MRN, SSN.

Most of the patients visit multiple clinics to try to find a reason for their disease. For example, if a patient visit clinic "A" for urgent care for heart attack and follow-up with his primary care physician at clinic "B" who further refer him to cardiologist at clinic "C". At each clinic, physician looks in depth, go for some test and prescribe some drugs. Thus, healthcare data is often fragmented thus improving interoperability in health sector not only help to improve patient care but also helps to save $30 billion a year. Hospitals have yet to achieve interoperability level, without it, it is almost impossible to improve patient care. US health department want interoperability between disparate EHRs by 2024. All medical stakeholder (physician, administrators, patients etc.) says that interoperability will improve patient care, reduces medical errors and save budget. Imagine having the insight and opinions of hundred of IVF/PGD patients, eases in your decision and satisfaction before going to treatment rather than directly relying on physician recommendations. Based on importance of data integration, healthcare organizations are turning to the implementation of interoperability.

Interoperability is backbone to critical improvements in health sector and yet has to become a reality. Similar to the concept of ATM network, health data should be standardized and shared between providers. Generally, it seems that its quite easy and straightforward to integrate EHRs (integration of different electronic databases). Yes, it could be easy, if EHRs have common structure for data collection but it is not the case in real world. In practice, EHRs are much more complex and were not designed as open system where information can be shared with multiple providers. In fact EHRs were developed with aim to replace paper based work and coordinate patient care within a hospital. It is the most urgent needs in the quest for healthcare improvement. Despite its urgent requirement, the implementation of interoperability is slow due to several factors. One real issue in exchanges of data between providers is that most of the EHRs are unable to interface with each other due to unstructured data. Another issues are data standardization, i.e., varied vocabularies, multiple interpretations. Like banking, medical data need to be standardized so that EHRs can share it automatically and physician can interpret it regardless of EHR. Quality of data, i.e., missing values, multiple records etc. also leads to difficulty in data exchange.

To achieve interoperability level, HL7, HITECH including HIPPA and several other standardization bodies have defined some standards and guidelines. Reviewing process for interoperability level includes standardized test scripts and exchange test of standardized data. The assessment of an organization that how it has achieved the interoperability and security standards, a third party opinion on EHR by the authorized testing and certifying body (ATCB) is considered. CCHIT and ARRA are the two types of certification that are used to evaluate the system.

6.2 P4 medicine

Healthcare practice has largely been reactive where patient have to wait until onset of disease and then treatment and cure of that disease. Our effort to treat disease are often imprecise, ineffective and unpredictable as we do not know genetic and environmental factor of disease. Moreover, drugs and treatment, we are advised, are tested on broad population and prescribed using statistical average; thus, they might work for some population but not all. Overall, P4 medicine (defined as personalized, predictive, preventive and participatory) has the potential to revolutionize care by customizing the healthcare to patient, enabling providers to match drugs to patient based on their profile (right dose of right drug for the right person at the right time), to identify which health condition patient is susceptible to and to determine how patient respond to particular therapy; however, current challenges and concerns need to be addressed to enhance its uptake and funding to benefit patients. One of the important purposes of P4 medicine is to enable disease prevention among healthy individuals through early detection of risk factors. Initiative to P4 care has been taken via personalized genomics, mobile health technology and pilot projects [224]. Near in future, physicians will be able to examine the unique biology of each individual to assess the probability of developing disease such as cancer, diabetes and will be treat perturbations in healthy individuals before symptoms appear, thus optimizing the health condition of individuals and preventing disease.

P4 medicare technologies, however, do not fit exactly into existing healthcare technology assessment and reimbursement process. Future landscape of P4 care is being dictated by current decisions in this space. Currently, this is one of the main challenges for both healthcare service providers and patients reluctant to advancement in healthcare industry. To overcome this issue and provide customized healthcare service, healthcare providers need to build personalized medicare. Rapid growth of p4 care makes it difficult for physicians to understand, interpret and apply new findings. It requires clinician to recognize that patient engagement means much more than their compliance. Medical education curriculum have generally not incorporated personalized medicare concepts. Another biggest challenge is societal acceptance of P4 deployment that is more daunting than scientific and technological challenges facing P4 medicine. These societal challenges include ethics, legal, privacy, data accessibility and data

ownerships issues etc. Moreover, P4 medicare will require new standards and new policies for handling individual's biological and health care information. To deal these issues, there is need to bring industrial partners as part of this consortium to help to transfer P4 medicine to the patient population. Medical colleges must include personalized medicare in curriculum. Patient interest and demand are essential, too.

Asc [225].

6.3 Security and privacy or confidentiality, privacy and security

Every stakeholder in health industry has a role to play in the privacy and security of patient information, In fact it is truly a shared responsibility. Patient privacy and information security are fundamental components of well-functioning healthcare system that helps to achieve better health outcomes, smarter spending, and healthier people. For example, patient may not disclose or may ask physician not to record his health information due to the lack of trust and feel that information might not be confidential. This kind of patient attitude put patient at risk and may deprive physician and researchers to go for important finding as well as put the organization at risk for clinical outcome and operational efficiency analysis. To reap the promise, providers and individuals must trust that an patient health information is private and secure, while on the other hand, providers are facing several challenges in the implementation of privacy and security at patient satisfaction level, i.e., efficient data analysis without providing access to precise data in specific patient records. For example, reminder sent to patient form drug store require access to patient's specific information in EHR that patient does not want to share with drug store. Thus, it this case, does the use of patient information for drug reminder outweigh the potential misuse of patient data ? Security and privacy issues are magnified by data growth, non-standard, variety and diversity of sources such as different formats, multiple sources, nature of data etc. Thus, traditional data security which are actually to secure small scale data are inadequate and fails. Security and privacy is data analytics especially when it draws information from multiple sources poses several challenges.

Privacy of patient data is both the technical issue as well as sociological, which must be addressed jointly from both technical and sociological perspectives. Health Insurance Portability and Accountability Act of 1996 (HIPAA) is world wide accepted set of general requirements and security standards to protect the information in health sector. HIPAA provides the legal rights regarding personally identifiable sensitive patient information and their establish obligations for healthcare providers to protect the use by avoiding its misuse. With the exponential rise in healthcare data, remote monitoring and smart devices, researchers are facing the big challenges to anonymize the patients sensitive information in order to prevent its misuse or disclosure? i.e. discarding of patient sensitive personal information such information SSN(social security number) MRN (medical record number), name, age etc. makes it very complex and challenging to link the patient data up to unique individual. Even hiding such information, still hacker can easily identify some of patient sensitive information through association. The other way to hide patient sensitive information is the use of differential privacy that helps to restrict the data access to the organization based on their requirement. However, these privacy challenges are factors that, in fact leads to the situations where, data analytics researchers are facing the issue from both legal as well as ethical perspective. The main privacy challenges associated with healthcare data analytics, overrunning the privacy concerns of traditional data. For example, how we can share the patient data while limiting the disclosure as well as ensuring the sufficient data utility, i.e., YOB, Gender, 3-digit Zip code unique for 0.04% of US citizens while DOB, Gender, 5-digit Zip code unique for 87% of US citizens [226]. However, limiting the data access results in unavailability of important information contents that could be important for certain data analytics task. Furthermore, the real-time data is not static but changes over period of time thus, earlier defined prevailing techniques may results in blocking some of the most required information or sharing the patient sensitive information.

6.4 Advanced analyzing techniques

Technological advancements (wearable devices, patient centered care etc) are transforming the entire healthcare industry. Nature of health data has evolved, currently, EHRs have simplified the data acquisition process with help of latest technology, but don't have the enough ability to aggregate, transform, or perform analytics on it. Intelligence is restricted to retrospective reporting that is insufficient for data analysis. A plethora of algorithms, techniques and tools are available for analysis of complex data. Traditional machine learning uses statistical analysis based on a sample of a total dataset. Use of traditional machine learning methods against this type of data is not efficient and computationally infeasible. Combination of big healthcare data and computational power lets the analysts to focus on analytics techniques scaled up to accommodate volume, velocity and variety of complex data. During the last decade, there is a dramatic change in the size and complexity of data thus, several emerging data analysis techniques have been presented. At minimum, in healthcare sector, data analytics for big data must have the

support of key functions that are necessary for analyzing the big data in real environment. The criteria for the evaluation of platform may includes, the availability, the scalability, the continuity, quality assurance ease of use, ability to manipulate granularity at different levels, privacy and security enablement [227]. Analyzing healthcare in real time is one of the key requirements, however, facing lot of challenges which required researchers attention, i.e., the lag between data collection and processing, dynamic availability of numerous analytics algorithms etc.

Innovative analytics techniques need to be developed to interrogate healthcare data and gain insight about hidden patterns, trends and associations in the data. Moreover, there will be a shortage of 100K plus-person data analytic researcher through 2020, which could mean 5060% of data analytics positions may go unfilled. Thus, there is need of data analytics scientist with advance technical skill. Deep machine learning, a set of machine learning algorithms based on neural network, is still evolving but has shown great potential for solving complex problems. It deduces the relationship without need of specific model and enables the machine to identify the pattern of interest in huge unstructured data. For example, deep learning-based approach learned on its own from the Wikipedia data that California and Texas are US states. It does not require to understand the concept of country and its states [227]. It reflects that how powerful is deep learning over other machine learning approaches.

6.5 Data quality and dataset

Gone are the days, when healthcare data was small, structured and collected exclusively in electronic health records. Due to the tremendous advancement in IT, wearable technology and other body sensors, increasingly, the data is quite large (moving to big data), unstructured (80% of electronic health data is unstructured), non-standard as well as in multimedia format. This variety in data makes it challenging and interesting for analysis. The quality of healthcare data is concerned for successful prevention system and is due to three main reasons, are incompleteness (missing data), inconsistency (data mismatch between within same or various EHR sources), inaccuracy (non-standards, incorrect or imprecise data) and data fragmentation. Data quality is a group of different techniques that are data standardization, verification, validation, monitoring, profiling and matching. The biggest problem of poor data in health industry has reached at epidemic proportion and introduces several pernicious effects in particular disease prevention. The problem with dirty data are mostly related to missing value, duplication, outliers and stale records. Thus, the first thing before going to data analytics is to perform cleaning for high-quality data and void the analytics until the data is fully clean. Not all healthcare data is collected directly through sensors, large fraction of data is collected through data entry (i.e., patient history, symptoms, physician recommendation, Lab report data entry etc.) that is not prefect as automatics entry and introduces human error in data. Incomplete data and inaccurate data, can also lead to missed or spurious associations that can be wasteful or even harmful to patient care. The opportunities for data error are relatively limited and easily identified by algorithms.

Three million preventable adverse events result 98,000 deaths and extra 17 billion cost per year, whereas most of the causes of medical errors can be overcame by improving the device interoperability. Currently, one-third percent of the hospitals uses various devices (such as defibrillators, electrocardiographs, vital signs monitors, ventilators and infusion pumps) could be integrated with EHRs. This lack of auto data acquisition creates significant sources of waste and risk to patient. Because of incomplete or stale information, data analytics methods may not be reliable for decision making. The solutions to deal with interoperability is not easier as compared to missing and inaccurate data. Since last two decades, little success has been made. Currently, hardware industry lacks the imperative to offer interoperability as most of the hospital have to bear additional cost to integration. They do not want hardware development companies to follow specific standards.

Although real-time data monitors (especially at ICU) are partially used in most of the hospital, however real-time data analytics is not in practice. Hospitals are moving to use real-time data collection and in near future, real-time data analytics will revolutionize the healthcare industry such as early identification of infections, continuous progress of treatment, selection of right drugs etc could helps to reduce the morbidity and mortality. To achieve real-time data processing, we need data standardization and device interoperability.

The other issue common issue is data standardization. Structuring of only 20 percent of data has shown its importance but on the other hand clinical notes are still in practice and created in billions due to the reason that the physician can best explain the clinical encounter. Empower physicians as well as maintaining the data quality is quite challenging. So far, this data is excluded from data analytics as its available in natural language and not discrete. Transforming this unstructured data into discreet form requires efficient intelligent technology and it is hard problem of medical IT until now. The only way we can use this unstructured and non-standard data by using NLP to translate that data using ICD or SNOMET CT into discrete data. Systrue system provides clinical NLP platform with advance intelligence techniques that allows the healthcare organizations to power up the content and data layer of

their applications. AWARE (Ambient Warning and Response Evaluation) is a viewer and clinical decision support tool designed to reduce risk of error. It works with multiple EHR systems and bedside monitors to present only relevant information on a single screen dashboard. Development process of AWARE is focused on reducing information overload, improving efficiency and eliminating medical error in the ICU.

Dataset development in health sector is not simple as it is in other areas. Supervised learning required labeled data. Real problem with dataset is the ground truth, it is more challenging when we are talking about big data and deep learning where we need really big dataset to train. It is very easy to get images of segregating men and women or cars images with different color and shapes however to annotate medical data, we need expert. Thus, expensive medical expertise is needed for high-quality annotation of medical data. Data privacy makes it more challenging as compared to other domain. Furthermore, annotation by single expert is not enough due to human error; thus, it is required to have consensus annotations by multiple expert observers. Whereas in case of unsupervised learning, we need millions of examples that are not easy on the other hand.

7 Conclusion

In health sector, accurate diagnosis/assessment depends upon data acquisition and data interpretation. Data acquisition has improved substantially over recent few years; however, data interpretation process has recently begun to benefits. One area of healthcare that data analytics have major impact is preventive care, helping clinicians to ward off disease before they have change to take hold. To face what we call this "prevention services" organizations must have to access to data analytics experts, healthcare experts, data as well as advanced analytics algorithms and tools. Intelligent healthcare data analytics has the big potential to transform the way the health sector industry uses the sophisticated and state-of-the-art technologies to gain the deeper insight into the data from for disease prevention. In near future, we will be the witness of rapid and widespread implementation as well as the use of P4 medicine across the healthcare organizations. To achieve P4 implementation level, several challenges highlighted above need to be addressed. p4 medicine is at nascent stage of development, rapid advancement in tools, algorithms and wearable healthcare sensors accelerate its maturing process. We have highlighted the challenges and identified a unique assumption that should be considered before implementation of any method. Detail of algorithms with guidelines for implementation of each section is presented.

Funding This work is partially supported by Australian Research Council Linkage Projects under LP170100891 and "Deanship of Scientific Research, King Saud University (Grant No. RG-1435-051)".

References

1. Razzak MI, Naz S (2017) Microscopic blood smear segmentation and classification using deep contour aware cnn and extreme machine learning. In: 2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW). IEEE, pp 801–807
2. Children: reducing mortality. 2015: updated September (2016)
3. Health expenditure, total (% of gdp). World Health Organization Global Health Expenditure Database (2014)
4. National health expenditure projections 2012–2022 (2012)
5. Bosworth H (2010) Improving patient treatment adherence: a clinician's guide. Springer, Berlin
6. Higginbotham EJ, Satcher D (2008) The public health approach to eliminating disparities in health. Am J Public Health 98(3):400–403
7. Patil HK, Seshadri R (2014) Big data security and privacy issues in healthcare. In: 2014 IEEE international congress on big data, pp 762–765
8. Slawson DL, Fitzgerald N, Morgan KT (2013) Position of the academy of nutrition and dietetics: the role of nutrition in health promotion and chronic disease prevention. J Acad Nutr Diet 113(7):972–979
9. World Health Organization (1990) Diet, nutrition, and the prevention of chronic diseases. Report of a WHO Study Group. Diet, nutrition, and the prevention of chronic diseases. Report of a WHO Study Group, p 797
10. Willett WC, Koplan JP, Nugent R, Dusenbury C, Puska P, Gaziano TA (2006) Prevention of chronic disease by means of diet and lifestyle changes. Disease Control Priorities in Developing Countries, pp 833–850
11. Policy brief on ageing no. 3, older persons as consumers (2009)
12. Policy brief on ageing no. 6, health promotion and disease prevention (2010)
13. Rivera DE, Atienza AA, Nilsen W, Allison SM, Mermelstein R, Riley WT (2011) Health behavior models in the age of mobile interventions: are our theories up to the task? Transl Behav Med 1(1):57–71
14. Griffiths F, Munday S, Friede T, Stables D, Holt TA, Thorogood M (2010) Automated electronic reminders to facilitate primary cardiovascular disease prevention: randomised controlled trial. Br J Gen Pract 60(573):137–143
15. Dasah JB, Kuranchie P, Amoah AGB, Adjei DN, Agyemang C (2015) The effect of electronic reminders on risk management among diabetic patients in low resourced settings. J Diabetes Complicat 29(6):818–821
16. Waqialla M, Alshammari R, Razzak MI (2015) An ontology for remote monitoring of cardiac implantable electronic devices. In: 2015 international conference on computer, communications, and control technology (I4CT). IEEE, pp 520–523
17. Waqialla M, Razzak MI (2016) An ontology-based framework aiming to support cardiac rehabilitation program. Procedia Comput Sci 96:23–32
18. Jiang Y, Shepherd M, Maddison R, Carter K, Cutfield R, McNamara C, Khanolkar M, Murphy Dobson R, Whittaker R (2016) Text message-based diabetes self-management support (SMS4BG): study protocol for a randomised controlled trial. Trials 17:179

19. Alaleh Z, Hollmann Markus W, Frits H, Benedikt P, Jeroen H, Polderman Jorinde AW, de Groot FA (2016) An automated reminder for perioperative glucose regulation improves protocol compliance. Diabetes Res Clin Pract 116:80–82
20. Fischer HH, Fischer IP, Pereira RI, Furniss AL, Rozwadowski JM, Moore SL, Durfee MJ, Raghunath SG, Tsai AG, Havranek EP (2016) Text message support for weight loss in patients with prediabetes: a randomized clinical trial. Diabetes Care 39:1364–1370
21. Daghistani T, Al Shammari R, Razzak MI (2015) Discovering diabetes complications: an ontology based model. Acta Inform Med 23(6):385
22. Gerber BS, Stolley MR, Thompson AL, Sharp LK, Fitzgibbon ML (2009) Mobile phone text messaging to promote healthy behaviors and weight loss maintenance: a feasibility study. Health Inform J 15(1):17–25
23. Romanelli RJ, Block TJ, Hopkins D, Carpenter HA, Dolginsky MS, Hudes ML, Palaniappan LP, Block CH, Block G, Azar KM (2015) Diabetes prevention and weight loss with a fully automated behavioral intervention by email, web, and mobile phone: a randomized controlled trial among persons with prediabetes. J Med Internet Res 17(10):e240
24. Steinhubl S, Kim S, Bae WK, Han JS, Kim JH, Lee K, Kim MJ, Kim JY, Oh S (2015) Effectiveness of 6 months of tailored text message reminders for obese male participants in a worksite weight loss program: randomized controlled trial. JMIR mHealth uHealth 3(1):e14
25. Morgan P, Callister R, Collins C, Hutchesson MJ, Tan CY (2015) Enhancement of self-monitoring in a web-based weight loss program by extra individualized feedback and reminders: randomized trial. J Med Internet Res 18(4):e82
26. O'Grady JS, Thacher TD, Chaudhry R (2013) the effect of an automated clinical reminder on weight loss in primary care. J Am Board Fam Med 26(6):745–750
27. Piette JD, List J, Rana GK, Townsend W, Striplin D, Heisler M (2015) Mobile health devices as tools for worldwide cardiovascular risk reduction and disease management. Circulation 132:2012–2027
28. Moreland EC, Volkening LK, Lawlor MT, Chalmers KA, Anderson BJ, Laffel LMB (2006) Use of a blood glucose monitoring manual to enhance monitoring adherence in adults with diabetes: a randomized controlled trial. Arch Int Med 166(6):689–695
29. Bonato P, Chan L, Patel S, Park H, Rodgers M (2012) A review of wearable sensors and systems with application in rehabilitation. J NeuroEng Rehabil 9:21
30. Zhou H, Hu H, Harris ND (2006) Wearable inertial sensors for arm motion tracking in home-based rehabilitation. In: Arai T, Pfeifer R, Balch TR, Yokoi H (eds) IAS. IOS Press, Amsterdam, pp 930–937
31. Ahmed E, Yaqoob I, Hashem IAT, Khan I, Ahmed AIA, Imran M, Vasilakos AV (2017) The role of big data analytics in internet of things. Comput Netw 129:459–471
32. Augustyniak P (2011) Personal wearable monitor of the heart rate variability. Bio-Algorithms Med-Syst 7(1):5–10
33. Parák J, Tarniceriu A, Renevey P, Bertschi M, Delgado-Gonzalo R, Korhonen I (2015) Evaluation of the beat-to-beat detection accuracy of pulse on wearable optical heart rate monitor. In: EMBC. IEEE, pp 8099–8102
34. Villalba E, Salvi D, Ottaviano M, Peinado I, Teresa AM, Akay A (2009) Wearable and mobile system to manage remotely heart failure. IEEE Trans Inf Technol Biomed 13(6):990–996
35. Zhang Z, Zheng J, Ha C (2016) Design and evaluation of a ubiquitous chest-worn cardiopulmonary monitoring system for healthcare application: a pilot study. Med Biol Eng Comput 55:283–294
36. Wang Z, Jiang H, Yang K, Zhang L, Wei J, Li F, Chi B, Zhang C, Wu S, Lin Q, Jia W (2013) Lifetime tracing of cardiopulmonary sounds with ultra-low-power sound sensor stick connected to wireless mobile network. In: NEWCAS. IEEE, pp 1–4
37. Koo HS, Michaelson D, Teel K, Kim D-J, Park H, Park M (2016) Design preferences on wearable e-nose systems for diabetes. Int J Cloth Sci Technol 28(2):216–232
38. Hosseini V (2015) Algorithm and related application for smart wearable devices to reduce the risk of death and brain damage in diabetic coma. J Diabetes Sci Technol 10(3):802–803
39. Esser P, Steins D, Dawes H, Collett J (2014) Wearable accelerometry-based technology capable of assessing functional activities in neurological populations in community settings: a systematic review. J Neuroeng Rehabil 11:36
40. Kausar N, Palaniappan S, Samir BB, Abdullah A, Dey N (2016) Systematic analysis of applied data mining based optimization algorithms in clinical attribute extraction and classification for diagnosis of cardiac patients. In: Hassanien AE, Grosan C, Tolba MF (eds) Applications of intelligent optimization in biology and medicine, vol 96. Intelligent Systems Reference Library. Springer, Berlin, pp 217–231
41. Yoo C, Ramirez L, Liuzzi J (2014) Big data analysis using modern statistical and machine learning methods in medicine. Int Neourol J 18:50
42. Wu X, Zhu X, Wu G-Q, Ding W (2014) Data mining with big data. IEEE Trans Knowl Data Eng 26(1):97–107
43. Joshi S, Nair MK (2015) Prediction of heart disease using classification based data mining techniques. Springer, New Delhi, pp 503–511
44. Smitha T, Sundaram V (2012) Classification rules by decision tree for disease prediction. Int J Comput Appl 43(8):6–12
45. Anooj PKN (2011) Clinical decision support system: risk level prediction of heart disease using weighted fuzzy rules and decision tree rules. Cent Eur J Comput Sci 1(4):482–498
46. Persi Pamela I, Gayathri P, Jaisankar N (2013) A fuzzy optimization technique for the prediction of coronary heart disease using decision tree. Int J Eng Technol 5(3):2506–2514
47. Liu X, Fu H (2014) PSO-based support vector machine with cuckoo search technique for clinical disease diagnoses. Sci World J. https://doi.org/10.1155/2014/548483
48. Anto S, Chandramathi S, Aishwarya S (2016) An expert system based on LS-SVM and simulated annealing for the diagnosis of diabetes disease. IJICT 9(1):88–100
49. Vadicherla D, Sonawane S (2013) Decision support system for heart disease based on sequential minimal optimization in support vector machine. Int J Eng Sci Emerg Technol 2(2):19–26
50. Subbalakshmi G, Ramesh K, Rao MC (2011) Decision support in heart disease prediction system using Naive Bayes. Indian J Comput Sci Eng (IJCSE) 2(2):170–176
51. Tseng W-T, Chiang W-F, Liu S-Y, Roan J, Lin C-N (2015) The application of data mining techniques to oral cancer prognosis. J Med Syst 39(5):1–7
52. Khan SU (2015) Classification of Parkinsons disease using data mining techniques. Parkinsons Dis Alzheimer Dis 2:4
53. Razzak MI, Blumenstein M, Xu G. Robust 2D joint sparse PCA. arXiv preprint arXiv:1001.2019
54. Razzak MI, Saris RA, Blumenstein M, Xu G (2018) Robust 2D joint sparse principal component analysis with f-norm minimization for sparse modelling: 2D-RJSPCA. In: 2018 international joint conference on neural networks (IJCNN). IEEE, pp 1–7
55. Quinlan R (1986) Induction of decision trees. Mach Learn 1:81–106
56. Quinlan JR (1992) C4.5: programs for machine learning. Morgan Kaufmann, Burlington

57. Kong L, Wiederhold BK, Gao K, Wiederhold MD (2013) Clinical experiment to assess effectiveness of virtual reality teen smoking cessation program. Stud Health Technol Inform 19:58–62
58. Valdimir V, Corinna C (1995) Support-vector networks. Mach Learn 20(3):273–297
59. Mangasarian OL, Musicant David R (2000) Lagrangian support vector machine classification. Technical Report 00-06, Data Mining Institute, Computer Sciences Department, University of Wisconsin, Madison, Wisconsin. ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/00-06.ps
60. Luo L, Xie Y, Zhang Z, Li W-J (2015) Support matrix machines. In: International conference on machine learning, pp 938–947
61. Zheng Q, Zhu F, Qin J, Chen B, Heng P-A (2018) Sparse support matrix machine. Pattern Recognit 76:715–726
62. Razzak MI, Blumenstein M, Xu G. Robust support matrix machine. arXiv preprint arXiv:1001.2019
63. Razzak MI, Blumenstein M, Xu G (2019) Multi-class support matrix machines by maximizing the inter-class margin for single trial EEG classification, pp 1–10
64. Razzak MI, Blumenstein M, Xu G. Cooperative evolution multiclass support matrix machines for single trial EEG classification. IEEE J Biomed Health Inform, pp 1–10
65. Kamruzzaman SM, Sarkar AM (2013) A new data mining scheme using artificial neural networks. Sensors 11(5):4622–4647
66. Razzak MI, Naz S, Zaib A (2018) Deep learning for medical image processing: overview, challenges and the future. In: Classification in BioApps. Springer, Cham, pp 323–350
67. Najafabadi MM, Villanustre F, Khoshgoftaar TM, Seliya N, Wald R, Muharemagic E (2015) Deep learning applications and challenges in big data analytics. J Big Data 2(1):1
68. Hinton GE, Osindero S, Teh YW (2006) A fast learning algorithm for deep belief nets. Neural Comput 18(7):1527–1554
69. Phan NH, Dou D, Wang H, Kil D, Piniewski B (2015) Ontology-based deep learning for human behavior prediction in health social networks. In: BCB. ACM, pp 433–442
70. Razzak I, Imran M, Xu G (2018) Efficient brain tumor segmentation with multiscale two-pathway-group conventional neural networks. IEEE J Biomed Health Inform. https://doi.org/10.1109/JBHI.2018.2874033
71. Langley P, Sage S (2013) Induction of selective Bayesian classifiers. CoRR, arXiv:1302.6828
72. Boullé M (2007) Compression-based averaging of selective Naive Bayes classifiers. J Mach Learn Res 8:1659–1685
73. Provost F (2000) Well-trained PETs: improving probability estimation trees, CDER working paper #00-04-is, Stern School of Business, NYU, NY, 10012
74. Boullé M (2006) Regularization and averaging of the selective Naive Bayes classifier. In: IJCNN. IEEE, pp 1680–1688
75. Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Philip SY, Zhou Z-H, Steinbach M, Hand DJ, Steinberg D (2007) Top 10 algorithms in data mining. Knowl Inf Syst 14(1):1–37
76. Wrenn JO, Stein DM, Bakken S, Stetson PD (2010) Quantifying clinical narrative redundancy in an electronic health record. J Am Med Inform Assoc 17:49–53
77. McInnes BT, Melton GB, Zhang R, Pakhomov S (2011) Evaluating measures of redundancy in clinical texts. In: AMIA annual symposium proceedings
78. Cohen R, Elhadad M, Elhadad N (2013) Redundancy in electronic health record corpora: analysis, impact on text mining performance and mitigation strategies. BMC Bioinform 14(1):10
79. Silva HB, Brito P, da Costa JP (2006) A partitional clustering algorithm validated by a clustering tendency index based on graph theory. Pattern Recognit 39(5):776–788
80. Cortijo FJ, Molina R, Garcia JA, Fdez-Valdivia J (1994) A dynamic approach for clustering data. Signal Process 44(2):181–196
81. Self M, Stutz J, Cheeseman P, Kelly J (1988) Autoclass: a Bayesian classification system. In: Proceedings of the fifth international conference on machine learning, Morgan Kaufman, Los Altos, CA, vol 58, issue Supplement, pp 54–64
82. Silva HB, Lerman I, Costa J (2006) Validation of very large data sets clustering by means of a nonparametric linear criterion. Classification, clustering and data analysis. Springer, Berlin, pp 147–157
83. Milligan GW, Cooper MC (1985) An examination of procedures for determining the number of clusters in a data set. Psychometrika 50(1):159–179
84. Cheeseman PC, Stutz JC (1996) Bayesian classification (AutoClass): theory and results. In: Advances in knowledge discovery and data mining, vol 180, pp 153–180
85. Dongkuan X, Tian Y (2015) A comprehensive survey of clustering algorithms. Ann Data Sci 2(2):165–193
86. Cong S, Han J, Padua DA (2005) Parallel mining of closed sequential patterns. In: Grossman R, Bayardo RJ, Bennett KP (eds) KDD. ACM, New York, pp 562–567
87. Ahmed S, Coenen F, Leng PH (2006) Tree-based partitioning of date for association rule mining. Knowl Inf Syst 10(3):315–331
88. Ceglar A, Roddick JF (2006) Association mining. ACM Comput Surv 38(2):5:1–5:42
89. He J, Rong J, Sun L, Wang H, Zhang Y, Ma J (2018) D-ECG: a dynamic framework for cardiac arrhythmia detection from IoT-based ECGS. In: International conference on web information systems engineering. Springer, pp 85–99
90. Ma J, Sun L, Wang H, Zhang Y, Aickelin U (2016) Supervised anomaly detection in uncertain pseudoperiodic data streams. ACM Trans Internet Technol (TOIT) 16(1):4
91. Chandola V, Banerjee A, Kumar V (2009) Anomaly detection: a survey. ACM Comput Surv 41(3):15
92. Goldstein M, Uchida S (2016) A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. PLoS ONE 11(4):e0152173
93. Habeeb RA, Nasaruddin F, Gani A, Hashem IA, Ahmed E, Imran M (2018) Real-time big data processing for anomaly detection: a survey. Int J Inf Manag. https://doi.org/10.1016/j.ijinfomgt.2018.08.006
94. Ramaswamy S, Rastogi R, Shim K (2000) Efficient algorithms for mining outliers from large data sets. In: SIGMODREC: ACM SIGMOD Record, vol 29
95. Angiulli F, Pizzuti C (2002) Fast outlier detection in high dimensional spaces. In: European conference on principles of data mining and knowledge discovery, PKDD, LNCS, vol 6
96. Goldstein M, Dengel A (2012) Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. In: Poster and demo track of the 35th German conference on artificial intelligence (KI-2012), pp 59–63
97. Schölkopf B, Platt JC, Shawe-Taylor J, Smola AJ, Williamson RC (2001) Estimating the support of a high-dimensional distribution. Neural Comput 13(7):1443–1471
98. Abdennadher S, Amer M, Goldstein M (2013) Enhancing one-class support vector machines for unsupervised anomaly detection. In: Proceedings of the ACM SIGKDD workshop on outlier detection and description (ODD13), New York, NY, USA, pp 8–15
99. Song X, Wu M, Jermaine C, Ranka S (2007) Conditional anomaly detection. IEEE Trans Knowl Data Eng 19(5):631–645

100. An W, Liang M, Liu H (2015) An improved one-class support vector machine classifier for outlier detection. Proc Inst Mech Eng C J Mech Eng Sci 229(3):580–588
101. Shirazi SH, Umar AI, Naz S, Razzak MI (2016) Efficient leukocyte segmentation and recognition in peripheral blood image. Technol Health Care 24(3):335–347
102. Monaghan AJ, Morin CW, Steinhoff DF, Wilhelmi O, Hayden M, Quattrochi DA, Reiskind M, Lloyd AL, Smith K, Schmidt CA, Scalf PE, Ernst K (2016) On the seasonal occurrence and abundance of the Zika virus vector mosquito *Aedes Aegypti* in the contiguous United States. PLoS Curr Outbreaks. https://doi.org/10.1371/currents.outbreaks.50dfc7f46798675fc63e7d7da563da76
103. Gidding SS, McGill HC Jr, McMahan CA (2008) Preventing heart disease in the 21st century: implications of the pathobiological determinants of atherosclerosis in youth (PDAY) study. Circulation 117(9):1216–27
104. Jonnagaddala J, Liaw ST, Ray P, Kumar M, Dai HJ, Hsu CY (2015) Identification and progression of heart disease risk factors in diabetic patients from longitudinal electronic health records. BioMed Res Int. https://doi.org/10.1155/2015/636371
105. Jitendra Jonnagaddala, Liaw S-T, Ray P, Kumar M, Chang N-W, Dai H-J (2015) Coronary artery disease risk assessment from unstructured electronic health records using text mining. J Biomed Inform 58(Supplement):S203–S210
106. Alneamy JSM, Alnaish RAH (2014) Heart disease diagnosis utilizing hybrid fuzzy wavelet neural network and teaching learning based optimization algorithm. Adv Artif Neural Syst. https://doi.org/10.1155/2014/796323
107. Makhtar AK, Yussof H, Al-Assadi H, Cheng Yee L, Rajeswari K, Vaithiyanathan V, Neelakantan TR (2012) International symposium on robotics and intelligent sensors 2012 (IRIS 2012) feature selection in ischemic heart disease identification using feed forward neural networks. Procedia Eng 41:1818–1823
108. Arslan AK, Colak C, Sarihan ME (2016) Different medical data mining approaches based prediction of ischemic stroke. Comput Methods Progr Biomed 130:87–92
109. Kunjunninair AP (2012) Clinical decision support system: risk level prediction of heart disease using weighted fuzzy rules. J King Saud Univ Comput Inf Sci 24:27–40
110. Shouman M, Turner T, Stocker R (2011) Using decision tree for diagnosing heart disease patients. In: AusDM, volume 121 of CRPIT. Australian Computer Society, pp 23–30
111. Shouman M, Turner T, Stocker R (2012) Using data mining techniques in heart disease diagnosis and treatment. In: 2012 Japan-Egypt conference on electronics, communications and computers (JEC-ECC), pp 173–177
112. Bajaja P, Choudhary K (2015) Automated prediction of RCT (root canal treatment) using data mining techniques: ICT in health care. Procedia Comput Sci 46:682–688
113. NaliniPriya G, Kannan A, AnandhaKumar P (2012) A knowledgeable decision tree classification model for multivariate heart disease data-A boon to healthcare. In: Li Z, Li X, Liu Y, Cai Z (eds) ISICA, vol 316. Communications in Computer and Information Science. Springer, Berlin, pp 459–467
114. Alizadehsani R, Habibi J, Bahadorian B, Mashayekhi H, Ghandeharioun A, Boghrati R, Sani ZA (2012) Diagnosis of coronary arteries stenosis using data mining. J Med Signals Sens 2(3):153–159
115. Alizadehsani R, Habibi J, Sani ZA, Mashayekhi H, Boghrati R, Ghandeharioun A, Khozeimeh F, Alizadeh-Sani F (2013) Diagnosing coronary artery disease via data mining algorithms by considering laboratory and echocardiography features. Res Cardiovasc Med 2(3):133–139
116. Karaolis MA, Moutiris JA, Hadjipanayi D, Pattichis CS (2010) Assessment of the risk factors of coronary heart events based on data mining with decision trees. IEEE Trans Inf Technol Biomed 14(3):559–566
117. Alizadehsani R, Habibi J, Hosseini MJ, Mashayekhi H, Boghrati R, Ghandeharioun A, Bahadorian B, Sani ZA (2013) A data mining approach for diagnosis of coronary artery disease. Comput Methods Progr Biomed 111(1):52–61
118. Snasel V, Alancar M, Jeloneck D, Salem AB, Valanderau L, Saleh Y, Kumar S, Theodoutou E, El-Bialy R, Salamay MA, Karam OH, Khalifa ME (2015) Feature analysis of coronary artery heart disease data sets. Procedia Comput Sci 65:459–468
119. Lee J, Jaekwon K, Lee Y (2015) Data-mining-based coronary heart disease risk prediction model using fuzzy logic and decision tree. Healthc Inform Res 21(3):167–174
120. Bhatla N, Jyoti K (2012) A novel approach for heart disease diagnosis using data mining and fuzzy logic. Int J Comput Appl 54(17):16–21
121. Kunjunninair AP (2012) Erratum to: "Clinical decision support system: risk level prediction of heart disease using weighted fuzzy rules and decision tree rules". Cent Eur J Comput Sci 2(1):86
122. Yeh D-Y, Cheng C-H, Chen Y-W (2011) A predictive model for cerebrovascular disease using data mining. Expert Syst Appl 38(7):8970–8977
123. Razzak MI, Alhaqbani B (2015) Automatic detection of malarial parasite using microscopic blood images. J Med Imaging Health Inform 5(3):591–598
124. Razzak MI (2015) Malarial parasite classification using recurrent neural network. Int J Image Process (IJIP) 9(2):69
125. Husain W, Adnan MHM, Rashid NA (2012) Hybrid approaches using decision tree, Naive Bayes, means and euclidean distances for childhood obesity prediction. Int J Softw Eng Appl 6(3):99–106
126. Bellazzi R, Abu-Hanna A (2009) Data mining technologies for blood glucose and diabetes management. J Diabetes Sci Technol 3(3):603–612
127. Breault JL, Goodall CR, Fos PJ (2002) Data mining a diabetic data warehouse. Artif Intell Med 26(1–2):37–54
128. Yamaguchi M, Kaseda C, Yamazaki K, Kobayashi M (2006) Prediction of blood glucose level of type 1 diabetics using response surface methodology and data mining. Med Biol Eng Comput 44(6):451–457
129. Shirazi SH, Umar AI, Haq N, Naz S, Razzak MI, Zaib A (2018) Extreme learning machine based microscopic red blood cells classification. Cluster Comput 21(1):691–701
130. Cho BH, Yu H, Kim KW, Kim TH, Kim IY, Kim SI (2008) Application of irregular and unbalanced data to predict diabetic nephropathy using visualization and feature selection methods. Artif Intell Med 42(1):37–53
131. Huang YC (2013) Mining association rules between abnormal health examination results and outpatient medical records. Health Inf Manag J 42(2):23–30
132. Alexander GE (2004) Biology of Parkinsons disease: pathogenesis and pathophysiology of a multisystem neurodegenerative disorder. Dialogues Clin Neurosci 6(3):259–280
133. Singh N, Pillay V, Choonara YE (2007) Advances in the treatment of Parkinson's disease. Prog Neurobiol 81(1):29–44
134. Naseer A, Rani M, Naz S, Razzak MI, Imran M, Xu G (2019) Refining Parkinson's neurological disorder identification through deep transfer learning. Neural Comput Appl. https://doi.org/10.1007/s00521-019-04069-0
135. Roberts SJ, Costello DA, Moroz IM, Little MA, McSharry PE (2007) Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. Biomed Eng Online 6:23
136. Ramani RG, Sivagami G (2011) Parkinson disease classification using data mining algorithms. Int J Comput Appl 32(9):17–22

137. Kunjunninair AP (2015) Classification of Parkinsons disease using data mining techniques. J Parkinsons Dis Alzheimer Dis 2(1):4
138. Suganya P, Sumathi CP (2014) A novel metaheuristic data mining algorithm for the detection and classification of Parkinson disease. Indian J Sci Technol 8:1
139. Kusiak A, Dixon B, Shah S (2005) Predicting survival time for kidney dialysis patients: a data mining approach. Comput Biol Med 35(4):311–327
140. Tsao C-W, Yeh J-Y, Wu T-H (2011) Using data mining techniques to predict hospitalization of hemodialysis patients. Decis Support Syst 50(2):439–448
141. Raju D, Xiaogang S, Patrician PA, Loan LA, McCarthy MS (2015) Exploring factors associated with pressure ulcers: a data mining approach. Int J Nurs Stud 52(1):102–111
142. Belanger AJ, Wolf PA, D'Agostino RB, Kannel WB (1991) Probability of stroke: a risk profile from the Framingham study. Stroke 22:312–318
143. Khosla A, Cao Y, Lin CCY, Chiu HK, Hu J, Lee H (2010) An integrated machine learning approach to stroke prediction. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp 183–192
144. Sung S-F, Hsieh C-Y, Yang Y-HK, Lin H-J, Chen C-H, Chen Y-W, Hu Y-H (2015) Developing a stroke severity index based on administrative data was feasible using data mining techniques. J Clin Epidemiol 68(11):1292–1300
145. Easton JF, Stephens CR, Angelova M (2014) Risk factors and prediction of very short term versus short/intermediate term post-stroke mortality: a data mining approach. Comput Biol Med 54:199–210 (2014)
146. Baitharu TR, Pani SK (2016) Analysis of data mining techniques for healthcare decision support system using liver disorder dataset. Procedia Comput Sci 85:862–870
147. Yasin H, Jilani TA, Danish M (2011) Hepatitis-C classification using data mining techniques. Int J Comput Appl 24(3):1–6
148. Zayed N, Awad AB, El-Akel W, Doss W, Awad T, Radwan A, Mabrouk M (2013) The assessment of data mining for the prediction of therapeutic outcome in 3719 Egyptian patients with chronic hepatitis c. Clin Res Hepat Gastroenterol 37(3):254–261
149. Yang ZR (2005) Mining SARS-CoV protease cleavage data using non-orthogonal decision trees: a novel method for decisive template selection. Bioinformatics 21(11):2644–2650
150. Lee S, Kim Y, Kang J, Oh J, Baek J, Yoon T (2014) Prediction of SARS coronavirus main protease by support vector machine. In: International proceedings of computer science and information technology, vol 59, p 185
151. Sandhu R, Sood SK, Kaur G (2016) An intelligent system for predicting and preventing MERS-CoV infection outbreak. J Supercomput 72(8):3033–3056
152. Xu Q, Wang H, Zhou L (2014) Seminal quality prediction using clustering-based decision forests. Algorithms 7:405–417
153. Sahoo AJ, Kumar Y (2014) Seminal quality prediction using data mining methods. Technol Health Care 22(4):531–545
154. Gil D, Girela JL, De Juan J, Gomez-Torres MJ, Johnsson M (2012) Predicting seminal quality with artificial intelligence methods. Expert Syst Appl 39(16):12564–12573
155. Maeemr M (2014) Etiological evaluation of seminal traits using Bayesian belief. Int J Biosci Biotechnol 6(6):79–86
156. Johnsson M, Gomez-Torres MJ, Girela JL, Gil D, De JJ (2014) Semen parameters can be predicted from environmental factors and lifestyle using artificial intelligence methods. Biol Reprod 88(4):99
157. Snowden S, Smye SW, Sharma V, Kaufmann SJ, Eastaugh JL (1997) The application of neural network in predicting the outcome of in-vitro fertilization. Hum Reprod 12(7):1454–1457
158. Corani G, Magli C, Giusti A, Gianaroli L, Gambardella LM (2013) A Bayesian network model for predicting pregnancy after in vitro fertilization. Comput Biol Med 43(11):1783–1792
159. Ciray HN, Bahceci M, Uyar A, Bener A (2009) Predicting implantation outcome from imbalanced IVF dataset. In: Proceedings of the World Congress on engineering and computer science, San Francisco, USA. WCECS, pp 562–567
160. Uyar A, Bener A, Ciray HN, Bahceci M (2010) ROC Based evaluation and comparison of classifiers for IVF implantation prediction. Springer, Berlin, pp 108–111
161. Chrelias C, Kassanos D, Siristatidis C, Pouliakis A (2011) Artificial intelligence in IVF: a need. Syst Biol Reprod Med 57(4):179–185
162. Pouliakis A, Trivella M, Papantoniou N, Bettocchi S, Siristatidis C, Vogiatzi P (2016) Predicting IVF outcome: a proposed web-based system using artificial intelligence. Vivo 30(4):07–08
163. Durairaj M, NandhaKumar R (2013) Data mining application on ivf data for the selection of influential parameters on fertility. Int J Eng Adv Technol (IJEAT) 6(2):262–266
164. Durairaj M, Nandhakumar R (2014) An integrated methodology of artificial neural network and rough set theory for analyzing IVF data. In: 2014 international conference on intelligent computing applications (ICICA), Coimbatore. IEEE, pp 126–129
165. Dilbaz S, Ozdegirmenci O, Demir B, Dilbaz B, Güvenir HA, Misirli G (2015) Estimating the chance of success in IVF treatment using a ranking algorithm. Med Biol Eng Comput 53:911–920
166. de Barros GP, de Paula LS, Bartmann AK, Faria M, Bettini NR (2013) The number of embryos obtained can offset the age factor in IVF results according to an artificial intelligence system. Womens Health Gynecol 2(5):17–19
167. Ziniewicz P, Milewska AJ, Czerniecki J, Woczyski S, Malinowski P, Milewski R (2014) The use of data mining methods to predict the result of infertility treatment using the IVF ET method. Stud Logic Gramm 39(52):67–74
168. Luo Y, Li J-Q, Zheng D-W, Tan Z-P, Zhou H, Deng Q-P, Liu Y-T, Ou A, Yin J (2011) Application of data mining technology in excavating prevention and treatment experience of infectious diseases from famous herbalist doctors. In: 2011 IEEE international conference on bioinformatics and biomedicine workshops (BIBMW). IEEE, pp 784–790
169. Kabir E, Siuly S, Cao J, Wang H (2018) A computer aided analysis scheme for detecting epileptic seizure from EEG data. Int J Comput Intell Syst 11:663–671
170. Kaushik K, Kapoor D, Varadharajan V, Nallusamy R (2014) Disease management: clustering-based disease prediction. Int J Collab Enterp 4(1–2):69–82
171. Vijayarani S, Sudha S (2015) An efficient clustering algorithm for predicting diseases from hemogram blood test samples. Indian J Sci Technol 8(17):1
172. Norouzi J, Yadollahpour A, Mirbagheri SA, Mazdeh MM, Hosseini SA (2016) Predicting renal failure progression in chronic kidney disease using integrated intelligent fuzzy expert system. Comput Math Methods Med. https://doi.org/10.1155/2016/6080814
173. Feng Y, Wang Y, Guo F, Xu H (2014) Applications of data mining methods in the integrative medical studies of coronary heart disease: progress and prospect. Evid Based Complement Alternat Med. https://doi.org/10.1155/2014/791841
174. Yang J, Si J, Gu X, Shi O (2013) Fuzzy cluster analysis of Alzheimer's disease-related gene sequences. Engineering 5(10):530
175. Shaukat K, Masood N, Shafaat AB, Jabbar K, Shabbir H, Shabbir S (2015) Dengue fever in perspective of clustering algorithms. arXiv preprint arXiv:1511.07353

176. Shouman M, Turner T, Stocker R (2012) Integrating Naive Bayes and *k*-means clustering with different initial centroid selection methods in the diagnosis of heart disease patients. In: CS & IT-CSCP, pp 125–137
177. Zheng B, Yoon SW, Lam SS (2014) Breast cancer diagnosis based on feature extraction using a hybrid of *k*-means and support vector machine algorithms. Expert Syst Appl 41(4):1476–1482
178. Ogasawara M, Sugimori H, Iida Y, Yoshida K (2005) Analysis between lifestyle, family medical history and medical abnormalities using data mining method—association rule analysis. In: International conference on knowledge-based and intelligent information and engineering systems. Springer, pp 161–171
179. Semenova T, Hegland M, Graco W, Williams G (2001) Effectiveness of mining association rules for identifying trends in large health databases. In: Workshop on integrating data mining and knowledge management, vol 1. ICDM, p 19
180. Ravi SS, David C, Torney S, Doddi AM (2001) Discovery of association rules in medical data. Med Inform Internet Med 26(1):25–33
181. Payus C, Sulaiman N, Shahani M, Bakar AA (2013) Association rules of data mining application for respiratory illness by air pollution database. Int J Basic Appl Sci 13(3):11–16
182. Concaro S, Sacchi L, Cerra C, Stefanelli M, Fratino P, Bellazzi R (2009) Temporal data mining for the assessment of the costs related to diabetes mellitus pharmacological treatment. In: AMIA
183. Concaro S, Sacchi L, Cerra C, Fratino P, Bellazzi R (2009) Mining healthcare data with temporal association rules: improvements and assessment for a practical use. In: Conference on artificial intelligence in medicine in Europe. Springer, pp 16–25
184. Jabbar MA, Chandra P, Deekshatulu BL (2012) Knowledge discovery from mining association rules for heart disease prediction. J Theor Appl Inf Technol 41(2):45–53
185. Raheja V, Rajan KS (2012) Comparative study of association rule mining and MiSTIC in extracting spatio-temporal disease occurrences patterns. In: 2012 IEEE 12th international conference on data mining workshops. IEEE, pp 813–820
186. Lee DG, Ryu KS, Bashir M, Bae J-W, Ryu KH (2013) Discovering medical knowledge using association rule mining in young adults with acute myocardial infarction. J Med Syst 37(2):1–10
187. Nahar J, Imam T, Tickle KS, Chen Y-PP (2013) Association rule mining to detect factors which contribute to heart disease in males and females. Expert Syst Appl 40(4):1086–1093
188. Anwar MA, Ahmed N (2014) Analyzing lifestyle and environmental factors on semen fertility using association rule mining. Inf Knowl Manag 3(4):79–86
189. Sharma N, Om H (2014) Significant patterns for oral cancer detection: association rule on clinical examination and history data. Netw Model Anal Health Inform Bioinform 3(1):1–13
190. Berka P, Rauch J (2010) Mining and post-processing of association rules in the atherosclerosis risk domain. In: Information technology in bio-and medical informatics, ITBAM 2010. Springer, pp 110–117
191. McCormick T, Rudin C, Madigan D (2011) A hierarchical model for association rule mining of sequential events: an approach to automated medical symptom prediction. SSRN Electron J. https://doi.org/10.2139/ssrn.1736062
192. Srinivasan S, Ramakrishnan S (2011) Evolutionary multi objective optimization for rule mining: a review. Artif Intell Rev 36(3):205–248
193. Ordonez C, Ezquerra N, Santana CA (2006) Constraining and summarizing association rules in medical data. Knowl Inf Syst 9(3):1–2
194. Ohsaki M, Abe H, Tsumoto S, Yokoi H, Yamaguchi T (2007) Evaluation of rule interestingness measures in medical knowledge discovery in databases. Artif Intell Med 41(3):177–196
195. Kuo Y-T, Lonie A, Pearce AR, Sonenberg L (2014) Mining surprising patterns and their explanations in clinical data. Appl Artif Intell 28(2):111–138
196. Soni S, Vyas OP (2010) Using associative classifiers for predictive analysis in health care data mining. Int J Comput Appl 4(5):33–37
197. Soni J, Ansari U, Sharma D, Soni S (2011) Intelligent and effective heart disease prediction system using weighted associative classifiers. Int J Comput Sci Eng 3(6):2385–2392
198. Rajendran P, Madheswaran M (2010) An improved image mining technique for brain tumour classification using efficient classifier. arXiv preprint arXiv:1001.1988
199. Shirisha Y, Rao SSS, Sujatha D (2012) Data mining with predictive analysis for healthcare sector: an improved weighted associative classification approach. Glob J Comput Sci Technol 11(22):31–36
200. Chin YC, Weng MY, Lin TC, Cheng SY, Yang YHK, Tseng VS (2015) Mining disease risk patterns from nationwide clinical databases for the assessment of early rheumatoid arthritis risk. PLoS ONE 10(4):e0122508
201. Carvalho FM, Teixeira HC, Dias EC, Meira W, Carvalho O (2015) A simple and effective method for anomaly detection in healthcare. In: 4th workshop on data mining for medicine and healthcare, 2015 SIAM international conference on data mining, Vancouver, Canada
202. Hu J, Wang F, Sun J, Sorrentino R, Ebadollahi S (2012) A healthcare utilization analysis framework for hot spotting and contextual anomaly detection. In: AMIA. AMIA
203. Mezger J, Visweswaran S, Hauskrecht M, Clermont G, Cooper GF (2007) A statistical approach for detecting deviations from usual medical care. In: AMIA annual symposium proceedings, p 1051
204. Hauskrecht M, Valko M, Kveton B, Visweswaran S, Cooper GF (2007) Evidence-based anomaly detection in clinical domains. In: Annual American medical informatics association symposium, pp 319–324
205. Valko M, Cooper G, Seybert A, Visweswaran S, Saul M, Hauskrecht M (2008) Conditional anomaly detection methods for patient-management alert systems. In: Proceedings of the international conference on machine learning, vol 2008. NIH Public Access
206. Hauskrecht M, Visweswaran S, Cooper GF, Clermont G (2013) Conditional outlier approach for detection of unusual patient care actions. In: AAAI (late-breaking developments), volume WS-13-17 of AAAI workshops. AAAI
207. Milos H, Michal V, Iyad B, Gilles C, Shyam V, Gregory C (2010) Conditional outlier detection for clinical alerting. In: AMIA annual symposium proceedings 2010, pp 286–90
208. Hong C, Hauskrecht M (2016) Multivariate conditional outlier detection and its clinical application. In: Schuurmans D, Wellman MP (eds) AAAI. AAAI Press, Menlo Park, pp 4216–4217
209. Batal I, Hong C, Hauskrecht M (2015) A generalized mixture framework for multi-label classification. In: SIAM data mining conference (SDM). SIAM
210. Harrou F, Kadri F, Chaabane S, Tahon C, Sun Y (2015) Improved principal component analysis for anomaly detection: application to an emergency department. Comput Ind Eng 88:63–77
211. Kwon C-B, Lee H-J, Oh J, Ban S-W (2016) Emergent cardiac anomaly classification using cascaded autoassociative multilayer perceptrons for bio-healthcare systems. Int J Bio-Sci Bio-Technol 8:351–362

212. Antonelli D, Bruno G, Chiusano S (2013) Anomaly detection in medical treatment to discover unusual patient management. IIE Trans Healthc Syst Eng 3(2):69–77
213. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ch IP, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE (2000) PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 101(23):e215–e220
214. Frank A, Asuncion A (2010) UCI machine learning repository. http://archive.ics.uci.edu/ml. University of California, Irvine, CA. School of Information and Computer Science, 213
215. Enright P, Furberg CD, Gardin JM, Kronmal RA, Kuller LH, Manolio TA, Mittelmark MB, Newman A, OLeary DH, Psaty B, Rautaharju P, Tracy RP, Fried LP, Borhani NO, Weiler PG (1991) The cardiovascular health study: design and rationale. Ann Epidemiol 1(3):263–276
216. Kveton B, Visweswaran S, Cooper GF, Hauskrecht M, Valko M (2007) Evidence-based anomaly detection. In: Proceedings of annual American medical informatics association symposium). AAAI, pp 319–324
217. Strack B, DeShazo JP, Gennings C, Olmo JL, Ventura S, Cios KJ, Clore JN (2014) Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. BioMed Res Int. https://doi.org/10.1155/2014/781670
218. Ghamdi HA, Alshammari R, Razzak MI (2016) An ontology-based system to predict hospital readmission within 30 days. Int J Healthc Manag 9(4):236–244
219. Al-Qarny ZA, Alshammari R, Razzak MI (2015) Impact of sharing health information related to diabetes through the social media network: ontology. Int J Behav Healthc Res 5(3–4):162–171
220. Gennings C, Olmo JL, Ventura S, Cios KJ, Clore JN, Strack B, DeShazo JP (2014) Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. BioMed Res Int. https://doi.org/10.1155/2014/781670
221. Asuncion A, Newman D (2007) UCI machine learning repository
222. Babu MSP, Ramana BV, Venkateswarlu NB (2012) A critical comparative study of liver patients from USA and India: an exploratory analysis. Int J Comput Sci 9:506
223. Valko M, Hauskrecht M (2008) Distance metric learning for conditional anomaly detection. In: FLAIRS conference, pp 684–689
224. Green S, Vogt H (2016) Personalizing medicine: Disease prevention in silico and in socio. Humana Mente J Philos Stud 9:105–145
225. Hood L, Galas D (2008) P4 medicine: personalized, predictive, preventive, participatory a change of view that changes everything. Computing Community Consortium, Washington
226. Sweeney L (2002) Achieving *k*-anonymity privacy protection using generalization and suppression. Int J Uncertain Fuzziness Knowl Based Syst 10(5):571–588
227. Raghupathi W, Raghupathi V (2014) Big data analytics in healthcare: promise and potential. Health Inf Sci Syst 2(1):3