Large Scale Medical Data Mining for Accurate Diagnosis: A Blueprint

Md. Sarwar Kamal, Nilanjan Dey and Amira S. Ashour

Abstract Medical care and machine learning are associated together in the current era. For example, machine learning (ML) techniques support the medical diagnosis process/decision making on large scale of diseases. Advanced data mining techniques in diseases information processing context become essential. The present study covered several aspects of large scale knowledge mining for medical and diseases investigation. A genome-wide association study was reported including the interactions and relationships for the Alzheimer disease (AD). In addition, bioinformatics pipeline techniques were implied for matching genetic variations. Moreover, a novel ML approaches to construct a framework for large scale gene-gene interactions were addressed. Particle swam optimization (PSO) based cancer cytology is another discussed pivotal field. An assembly ML Random forest algorithm was mentioned as it was carried out to classify the features that are responsible for Bacterial vaginosis (BV) in vagina microbiome. Karhunen-Loeve transformation assures features finding from various level of ChIP-seq genome dataset. In the current work, some significant comparisons were conducted based on several ML techniques used for diagnosis medical datasets.

Keywords Medical data mining · Machine learning · Particle swarm optimization · Alzheimer disease · Cancer · Bacterial vaginosis · Karhunen-Loeve transformation · Random–forest algorithm

Md.Sarwar Kamal
Computer Science and Engineering, East West University Bangladesh, Dhaka, Bangladesh
e-mail: sarwar.saubdcoxbazar@gmail.com

N. Dey (✉)
Department of Information Technology, Techno India College of Technology, Kolkata, India
e-mail: neelanjan.dey@gmail.com

A.S. Ashour
Faculty of Engineering, Department of Electronics and Electrical Communications Engineering, Tanta University, Tanta, Egypt
e-mail: amirasashour@yahoo.com

S.U. Khan et al. (eds.), *Handbook of Large-Scale Distributed Computing in Smart Healthcare*, Scalable Computing and Communications,
DOI 10.1007/978-3-319-58280-1_7

1 Introduction

Biological processing is considered the imperative part of the world computing. Thus, biological research has a great impact in data arrangement, analysis and measurement due to its robust mechanical and automated techniques. Mining techniques are significant for retrieving meaningful information from the biological data. It is a dynamic and systematic demonstration. However, more powerful methods, algorithms, software and integrated tools are required for the biological processing. Machine learning is one of the key methods for handling biological datasets and very large DNA (Deoxyribonucleic acid) sequences [1–4]. Computers and other digital systems assist large biological data processing, thus the discovery and development of new systems is essential with the rapid growth of large biological dataset. New research and computations are generating huge volume of datasets in each and every moment. Moreover traditional approaches are unable to manage very large biological data with accurate and fast computations. Consequently, biological mining techniques which are hybrid mechanisms with computer science, physics, chemistry, biology, mathematics, statistics, genetic engineering, molecular biology and biochemistry; become indispensable. Furthermore, sets of evolutionary computing algorithms can govern the large biological dataset processing [5–10]. These techniques achieve faster biological data processing with accuracy and perfections. Moreover, cloud computing, data sciences and bioinformatics are examples for popular new fields for assisting biological data processing. Furthermore, due to the massive amount of the biological information, big data in biological processing become a common phenomenon in current industry and laboratories. Organizing and arranging information from these big dataset is a challenging issue as well as a key factor in knowledge mining. Statistical and mathematical illustrations are supportive for retrieving meaningful and hidden information. Data mining techniques are equally important for information exaction [11–15].

In the age of wireless communications and faster digitization, very large bio centric information have been growing in exponential manner due to the rising of faster processing on microarray datasets. Moreover, DNA sequencing, RNA (Ribonucleic acid) synthesis, protein-protein interactions are also some prime factors that increase the datasets volume. Big data analysis techniques support these datasets to obtain meaningful information from the human and animal dataset. The growth of the data volume in the recent era is significantly huge compared to few years back. Recent growth is so rapid, which is almost twenty times more than three years back collections as reported in Fig. 1. The primary assessments are done for 12 years as 2000–2015.

In Fig. 1, the X-axis and Y-axis illustrated the years, and the datasets outcome for each year; respectively. All the four lines are merged due to the similarities among the datasets collections. There are very less changes from 2009 to 2015,

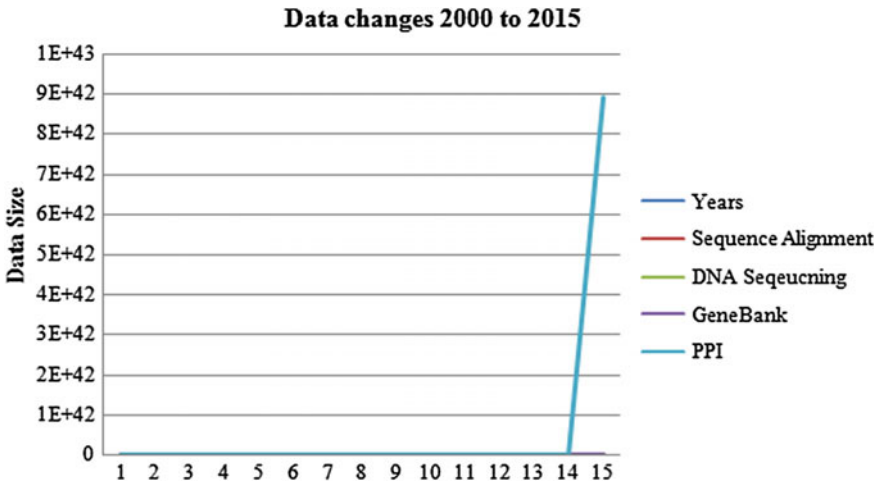

Fig. 1 Various datasets growth from 2000 to 2012 [16]

but there are sudden changes from 2011 to 2015. Currently, these changes are automated and reached the apex points without return [17–29]. One of the pivotal reasons behind the large biological datasets is the diversity of human, animal and plant life. Despite the variations, the biological details interrelated with the universal sets of living organs.

The main contribution of the current work is to highlight the machine learning role in the large scale medical data mining. Innovative data mining techniques in diseases information processing was addressed due to the excessive data generation in the medical experiments and very big scopes in human centric diseases. In addition, consecutive tissue networks have been discussed to identify the AD existence in the tissue. The PSO role to assess the infected parts of cells was addressed. The current work included also extensive discussion related to the statistical epitasis networks for checking gene-gene interactions for obesity. Furthermore, the concept and applications of big data in health informatics was included.

The chapter is organized as follows. Background study is demonstrated at Sect. 2 with delineating the AD related associations including the gray matter functionalities of this disease as well as introducing cancer based computational analysis. Big data and its impacts on health informatics are narrated in Sect. 3. In Sect. 4, the machine learning techniques based supervised learning analysis is addressed. Web semantics is demonstrated in Sect. 5, while the HIV-1 computational classifications, as well as the findings of obesity under gene-gene interaction, Microbial communities, and Bacterial Vaginosis, are addressed in Sect. 6. Finally, the conclusion is given in Sect. 7.

2 Background

General mining can be groped in several ways; however most popular mining processes can be categorized as predictive -or descriptive-data mining as reported in Table 1. Image processing, signal processing, business data processing, DNA sequencing or protein interactions are easily manageable using one of these techniques. Predictive data mining techniques are used frequently to get faster and accurate data. Most of the predictive techniques are based on statistical processing as well as mathematical analysis. There are lots of mathematical models and techniques that are under predictive models. Irrespective of areas and subjects, predictive data mining approaches are imperative to obtain exact information from lots of datasets. Moreover, simulations and other computing are also easily adjustable by predictive mining. It enables ML approaches to learn and train datasets based on the historical data demonstrations and computing. These analysis and synthesis bridge biological mining from present to upcoming future.

There are set of predictive approaches which are frequently used in the digital era, such as the neural networks, principal component analysis, independent component analysis, particle swarm intelligence, self-organization map, regressions, support vector machine (SVM), classification and regression tree (CART), decision tree (DT), deep neural networks (DNN), discriminate analysis (DA), Bayesian Network (BN), Boosting (BT) and Random Forest (RF). In addition, several experiments have been crowned using various methods of bioinformatics related to identification of gene factors behind enormous devastating diseases. Nonetheless, all the previous works transpire only a low segment of gene-gene

Table 1 Sources of data mining

Predictive biological analysis	Descriptive biological analysis
The primary goal is to separate the large dataset into small groups	This process generates set of rules for controlling whole dataset
Set of statistical analysis support to get the exact meaning of the desired items	There are set of descriptive analysis that indirectly used the predictive mining algorithms. As for example Smith waterman and Needleman Wanch algorithms were applied to find the sequences under predictive environments
All features associated in the training datasets are equally important for all data levels	Group of training data clustering is used to handle the large volume of biological datasets. Popular clustering processes are frequently using to get the meaningful ideas
Training and testing features determine the overall outcomes of the experiments	Features are collected in a group rather being individual
In each experiment, there must get some outcomes. If outcomes have probabilities greater than 70% it is said to be acceptable	Experimental outcomes are determined by group results. Sometimes single results are not measured due to the excessive volumes of datasets exist in a group

interaction or difference due to DNA sequence switch, which is liable for rapidly increasing rate of demolishing disease like Alzheimer's. The AD is one kind of neurological disturbance, which occurs due to damages of brain cells. Though Alzheimer's diseases start lightly, it widens rapidly causing short time memory loss and cognitional degeneration. Long term process of this disease leads to dementia, which is responsible for enormous damage of human brain's usual activities [30]. Various algorithms belong to bioinformatics can be applied to identify the principal genetic codes behind some crucial diseases like Alzheimer. Various studies regarding AD have been illustrated that almost 70% damage is associated with genetics for this diseases [31–33]. Even one of the significant gene APOE (Apo lipoprotein E), which is the principal cholesterol server to human brain is engaged with the genes accused of spreading AD.

Basically, developed NetWAS which is one of the ML based algorithms is considered to recognize the symphony among the associated genes. Furthermore, the Network Interface Miner for Mutagenic Interactions (NIMMI) combines protein-protein interaction data and GWAS data for better and reliable performance [34]. So, a new ML approach for tissue-specific internal reaction to override the previous findings from the method GWAS was implemented. Specific tissue features that may play prime role for determining the root cause for every devastating disease, such as the AD and can overcome various critical challenges were reported in [35]. Thus, the source code and overall findings of this work can simplify the way to develop approaches of various methods for best outcome and better efficiency.

Several studies were conducted related to the genetics computing for grey matter density in Alzheimer's disease, where the AD has no exact known cure [36–38]. In [37], a full concentration on bioinformatics approach to the genetic analysis of grey matter density to result in deceased outset of the AD was provided. Various kinds of ML were carried out by assembling them together for better execution than previous works. Full concentration on gene factors movement and internal reaction behind them along with functional genomics data for entrancing biological relationship was specified. Considering all the undiscovered facts, this study is based on genome wide association study (GWAS) and applied on the datasets which belong to the AD Neuroimaging Initiative (ADNI) using grey matter density methods implemented process. A new method was implemented to cope with polymorphisms and their regression to make an obstacle for rapidly growing AD. Functional magnetic resonance imaging (fMRI) methods on approximately 818 peoples as well as 733 genetic data categories for experiment were applied. After that both fMRI and GWAS has been embedded successfully to enhance the possibility of bringing to pass voxel-wise genome-wide association studies (vGWAS) for managing better opportunity to generate various mapping based problems. In the first stage, the Quantitative Multifactor Dimensionality Reduction (QMDR) process was engaged to classify the total number of genes along with SNPs, which can overcome the requirement of the first stage till execution. Basically, the QMDR helps to detect non-linear SNP–SNP interactions [39]. In the second phase, bioinformatics approach was applied on genes enrolled in the first phase to diminish

the number of genes factors as in [40]. Furthermore, gene factor evolution using microarray proves the complication of breast cancer disease. A lot of methods have been proposed by various researchers and scientists to identify the main culprit factors behind this devastating disease.

The most liable genes for this heinous disease are BRCA1 and BRCA2 [41]. Various recent studies represent several microarray resolution to get a better way to identify the sub graphs for the cure of breast cancer [42]. From the perspective of clinical science, one of the heterogeneous diseases is breast cancers which obstacle for improving the diagnosis of tumors classifications clinically [43]. Currently, multi-gene lists and single sample predictor models provide better performance to reduce the multidimensional complexity level of this disease. The incapability of some established model to deal with high dimensional data limits the opportunity of gaining desired result, however various new studies contributing a great role to compete with this mysterious disease. A new iterative powerful strategy for computably biased subtypes and enhancing class prediction while using METABRIC dataset were performed.

Typically, the traditional methods help largely for clinical decision making creating various discoveries. The PAM50 methods are used for assigning the molecular subtypes based on various gene expressions. Other various methods are also correlated to clinical diagnosis [44, 45]. All the methods work on low dimensional datasets as well as individual sets of data.

From bioinformatics, the ensemble learning with various algorithms results in a great extent. Actually the main advantage of ensemble learning is it can easily comprehend decreased over fitting as well as improvise performance of classification. There's a lots of ensemble approaches among which select-bagging and select-boosting are the main approaches to work efficiently and in a faster way. Although, the iterative approach was used alone or along with CM1 score then the outcome was quite disappointed, whereas the iterative approach with combination of an ensemble learning mechanism provides faster and efficient performance than others [46]. Besides, practically this work's improvised methods titled iterative method with CM1 score and ensemble learning approach represents a great effectively for foretelling more accurate and exact sample subtypes in the METABRIC breast cancer dataset.

Generally, several cancer diseases are identified by the human cell which is primarily affected [47–49]. Meanwhile, next-generation sequencing and microarrays already have disclosed enormous number of genomic features like DNA copy number alterations (CNA), mRNA expression (EXPR), microRNA expression (MIRNA), and DNA somatic mutations (MUT). Therefore, lots of exploration for a particular type of this genomic data produces various types of prediction biomarkers in cancer. Various predictive biomarkers have mentioned for research basis on various number of biological components simply, such as genomic, proteomic, metabolomics, pathological, imaging and psychological features. Thus, the genomic biological features have been used in a great extent although here National Cancer Institute and the National Human Genome Research Institute plays the main role [49].

In [49], a Cox proportional hazard model was proposed for feature selection algorithm. In addition, the constraint PSO process was also interpreted based on biological behavior. Completing various iterations using bootstrap beta coefficients have been detected going through the log-likelihood (NFS and CPSO) or a penalized maximum likelihood function. The used CPSO was basically associated with biological behavior of flock. Therefore, these particles belong to the swarm of particles representing the positions and velocity. This CPSO randomly set the positions vectors and velocity together. The velocities and positions are updated automatically updated analyzing their performance and types of elements. The authors have used 4 types of datasets which are online based associated with almost 100 types of elements including EXPR, MIRNA, CNA, and MUT in the TCGA assortment while accession. The CPSO was used to evaluate basis components of the sophisticated dataset using its ability to produce various user defined elements that is used as the basic survival model.

The network feature selection model is also used to detect protein-protein interaction network for evaluation of feature selection. Therefore, it is also used for producing poly-cancer biomarkers. This NFS established best performance, while the datasets were more complicated and sophisticated to deal with. It evaluated any kinds of data and represented the desired result [50] and also successfully explored the complicated data with noticeable success rate. This algorithm performed the best efficient process of finding the basic features for detecting cancer disease. Using similar algorithms all the models were evaluated and were compared using the concordance index (c-index) for obtaining the better performance.

This work differentiated the predictive level of various features genomic for characterize genomic related data. The integration process of enormous genomic datasets produced higher class models of datasets which are more powerful than that from single survival data simply like mRNA. From the source of genomic data, the mRNA gene reproduced stronger highly preserved models for integration of cancer genes data sets.

From the preceding studies, it is clear that the collected data from various scientific experiments or movements are massive which require efficient data management, analyze, providing, manipulating to reach in a goal [51]. Thus, big data analytics become essential to handle such huge data amount.

3 Big Data Computing

Big data computing is an overgrowing technique for mining multidimensional data from scientific discovery along with various large-scale structures [52]. Big data analysis system is designed in such a way that it can identify any meaningful data from a vast crowd of data. Big data technologies are currently gaining the opportunities in medical science, bioinformatics, health informatics, computer science, management system and lots of fields. Various recent articles have been reported statistical information for the big data computing benefits in several applications

such as mobile devices, tablet computers, internet of things, and cloud computing [53, 54]. The most significant fields are scientific exploration, health care, governance, financial and business management analysis, web analytics, internet of things along with mobile health informatics, and bioinformatics. Compared to big data computing, service-oriented technologies, such as cloud computing is also capable of storing data, analyzing and manipulating large size of data but sometimes it becomes challenging working with cloud computing, that time big data computing is must. The overall architecture of big data was demonstrated in Fig. 2.

Figure 2 illustrated the big data computing along with closer innovation or discussion on architecture, technologies, tools, mobile technologies, health technologies and many more paradigms working behind big data computing. Typically, big data technologies have been used widely medical science and health-care informatics research. A huge numbers of data sets have been gathered and generated using various bioinformatics approach for the sake of research in the fields of medical science [56, 57]. In order to compete with this increasing amount of dataset, algorithms of informatics science to explore the hidden discoveries from them were performed. In spite of all this there's a lots of obstacle with big data among which store, search, analysis, processing, sharing, viewing, discovering knowledge from those data though exploring knowledge from these type of big data has become burning question for the scientists and researchers of last decades [58].

Traditionally, there are various types of sources for genomics and proteomics information. Each and every source has its own styles, mythology though most of source use ontology like genome ontology. Precision medicine asserts the entire requirement needed to acquire the best clinical outcome. Therefore, precision of medical data refers to analyze, to interpret and to integrate the increasing number of unequal sources [59]. The benefits of machine learning, supervised learning, and data mining were used to replace the current traditional use of various algorithms. Informatics approaches for analyzing, integrating, and detaching the medicine legibility for better performance of current medical science were introduced in [59].

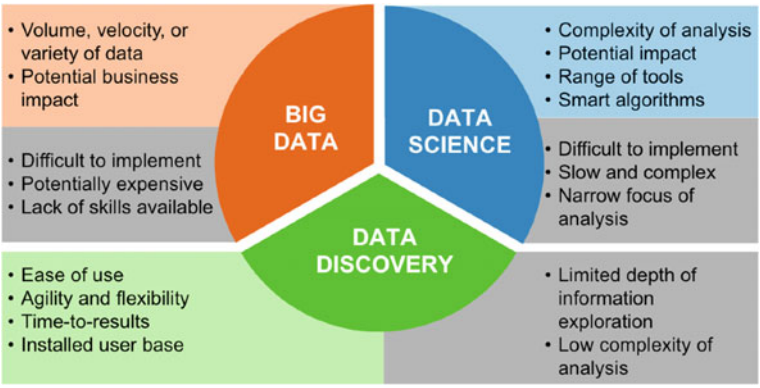

Fig. 2 Overview architecture of big data computing [55]

The authors differentiated between the traditional strategies and next generation informatics approaches to evaluate data of medical science for making a pathway to cope with any kind of diseases and explore new drugs analyzing the output produced using various modern informatics approaches.

Consequently, for going towards the goal of precise, medicine and healthcare in the current systems require bioinformatics approaches along with developed accurate storing capability for information regarding genomic data. Additionally, more emphasizes are compulsory to characterize genetic factors for accurate assessment for developing the health outcomes [60]. Health informatics is associated with health care technologies to develop more reliable health care system. Its basic elements are basically engaged with information science, computer science, social science, behavioral science, management science, and others [61]. Actually, health informatics is involved with the resources, devices, storages, backups with other computers, clinical guidelines, and information/communication system [62, 63]. For gaining knowledge from this vast amount of data, health informatics will require a potential limitless process to cope with this kind of data. The most challenging consequence working with vast data is investigating this data in a reliable manner. The main use of health informatics is to develop various bioinformatics process to illustrate vast medical data in an easy way. In addition, the health informatics cooperates with the population data which is executed in a particular subfield [64]. A big velocity for big data is happen when new update of current population's health care system is providing to data center in a great speed by means of health sensor or mobile devices [65].

The basis goal of health informatics is to provide answer of any patient's frequently asked question regarding health issue in a constant possible time via mobile devices, internet services, and tablet computers. Furthermore, mobile health monitors refers the use of improved modern technologies like mobile device, computers for monitoring health issue and providing alarm for risky situations [66]. Presently, the mobile health has risen as sub-segment of eHealth, use of information technologies like mobile phones, tablets, communications technologies, and patients monitoring [67]. It monitors the symptoms, signs of various diseases via mobile phone technologies [68, 69]. Sometimes this issue becomes challenging to tackle the big data problems. The spreading of mobile health technologies spectrum is shown in Fig. 3 [70].

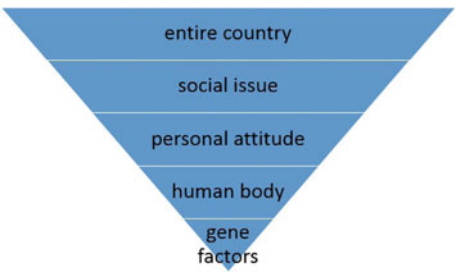

Fig. 3 The overall mobile health technologies process [70]

Such biomedical big data is considered the most challenging task for the active researchers and it indicates investigating, analyzing, storing data, visualizing in lower dimension of higher data. It generally extracts desired value from data [71, 72]. Therefore, big data requires a set of new techniques with new forms of integration to form better performance [73]. Recently, parallel computing which was proposed by Google has created a wide vision of working with big data. Cloud computing is also an approach of dealing with biological big data because of its ability to develop system ability, and to speed up the system. Cloud computing also diminishes system requirements. Biomedical research indicates analyzing the biological data in molecular or macromolecular stages. Next generation sequencing technologies and big data techniques in bioinformatics assist to access into biomedical research data sets. The Hadoop and MapReduce technologies are playing a great role recent years for dealing with the vast amount of data. Big data has great significant efforts in clinical informatics. Clinical informatics refers to health care technologies in medical science along with all kinds of health related issues. Overall there's a lot of use of big data in various fields like clinical science, biomedical research, and imaging informatics. By the sake of various ML algorithms, all the liable gene factors or symptoms as well as risk factors can be determined. Clinical and practically observed data actually helps to accomplish precautionary steps for type 2 diabetes [74–82].

4 A Supervised Learning Process to Validate Online Disease Reports

All the bioinformatics algorithms that determine the main risk gene factors or gene variations behind each and every disease of medical science must require a large number of data to manipulate. For this situation, the pathogen distribution model is widely used for its high predictive ability of any factors. It also demonstrates largely to create diseases maps depending on each disease's variation and liable gene factors. Production of data for these methods comes from online health based data system such as Genbank. One of the major problem of online data base is all the data may not be valid and there is no such method for validation of dynamically providing online based data. Depending on environmental and socio-economic condition the occurrence possibility of each disease in a particular position is defined by the location of disease occurrence.

4.1 Targeted Learning in Healthcare Research

From the beginning of availability of big data electronic health care technologies and claims data sets are arising in a great extent for answering the drug safety

measures all over the world. The ever-growing rate of these current technologies towards investigational data sets such as genomic information, laboratory results, and radiologic images [83]. These big data sets in health care technologies are increasing day by day and the questions are widening. Compared to this trend the new approaches are not improved enough to get rid of this trend. The old parametric modeling approaches are inadequate for analyzing the coefficients of the big data source. Therefore, the illustration of these coefficients is largely depends on various covariant. All the traditional approaches feel obstacle for the higher dimensionality of big data. They can't convert this higher dimension to its lower one [84]. Realizing this context, big data problems can be solved using a new approach namely targeted learning (TL).

Targeted learning (TL) is a current approach for dealing with big data problems which is implemented using semi parametric approach along with supervised machine learning mechanisms. This approach targets on higher dimensional data and helps previewing lower dimension of higher dimensional data. The specific focus of targeted learning algorithms is to minimize the bias of any targeted parameter and make it simple to reach in a discovery point. Basically the targeted learning (TL) algorithm is the combination of two bioinformatics approaches which are Super learning (SL) and targeted minimum loss- based estimation (TMLE). Actually, the TMLE is applied on data which is analyzed by Super learning (SL) before. First, the big clinical data is manufactured by SL, after that TMLE is used. Because of this combination the targeted learning (TL) approach outputs sound findings analyzing big health data. Nowadays, the TL is basically applied in a wide range of spheres like genomics, precision medicine, health policy, and drug safety. Therefore, this paper illustrated the significant contribution because of the combination of two bioinformatics approach named Super learning (SL) and targeted minimum loss- based estimation (TMLE).

4.2 Lumping Versus Splitting for Mining in Precision Medicine

In recent years, the rise of data severe biology, advancement in molecular biology and technological biology along with the way the health care is delivered to the patient of current world paves the biologists working on these new diseases to find a particular cure using precision medicine models [85]. Medical science is utilizing the advantage of modern improvised data mining methodologies in various aspects of detection of liable DNA codes or gene factors for better outcome of devastating diseases which are invented recently. It is one of the most challenges for medical science finding prevention or cure from such kinds of recent invented diseases. Consequently, thousands of people die because of not finding the exact gene factors or DNA codes working behind the disease. With the renovation of bioinformatics, a new movement has been seen in medical science because of getting the related

genes factor corresponding to each and every disease [86]. Therefore, nowhere the data mining is needed for precision medicine. The capability of representing each and every disease's risk for treatment purpose is one of the correlating factors for precision medicine. This achieved a great measure from the last era for ongoing technology improvement. A lot of current powerful approaches rely on unilabiate and linear evaluations that can be easily avoid the structure of complicated criterion [87, 88]. One of the successful example for the precision medicine model is to drug development involving the drug Crizotinib, a vanquisher for the MET and ALK kinases which started the practical improvement with a widen number of population in a great extent.

Several studies tried to prove that using accurate types of data mining and making working portion for each disease will be fruitful for the scientists and researchers determining subtype graphs in low cost along with less time consumption. With the determination of subtypes for various diseases, a great opportunity can be found to keep compete with the increasing rate and approaches of development of diseases along with accurate liable factors of that particular disease and also if this happens then it will be start of a new era from the perspective of medical science. Using enough specific small groups and perfect small types of data can improve the precision medicine models. Thus, the biological data mining process has a great effect in biological large data to play a sensitive role in finding responsible culprit accurately. It also can be used for many clinical and practical contexts to ensure better treatment for the patients and also for exploring reason behind recent invented diseases along with devastating diseases like diabetes or cancer.

5 Mining Drug-Drug Interactions on Semantic Web Technology

Drug-drug attraction has the most priority for showing its tremendous effect on patients [89–93]. A sufficient knowledge should be mandatory for prescribing or taking the medicine for both the clinical association and patients. The new improved process regarding investigation of various diseases risk factor for patients should be broadening to decline the death rate for such diseases. In order to improve genetic tests for finding the suspicion of various diseases, DDI-induced ADEs is required to diminish the risk of prescribing harmful medication [93]. Providing information investigating the DDIs data is full of challenge for medical science. However, using improved informatics approach can easily easy to beat the challenge.

A vast number of new drugs have been invented using various bioinformatics still processing. One of the main significant discoveries is taxon related drugs. Paclitaxel and docetaxel are considered the best performance giving anticancer taxon related drugs though they have lots of bad side effects [94]. In clinical

statistics, singular machinery to protest cancer gene for diminishing growth rate makes the Texans related drugs differ from the other related drugs. Besides it is considered most hopeful treatment of cancer disease over worldwide [95, 96].

6 Machine Learning Based Health Problems Manipulation

The HIV (human immunodeficiency virus) is one of the deadly disease from the last three or four decades. The prevention process for the HIV can be improved either scanning the responsible mutants or scanning the resistive capabilities of resistive drugs [97]. Several expert groups have been working on both genotypic and phenotypic consequence of genes for HIV although genotypic is faster and cheaper than phenotypic. In [98], protease and reverse transcriptase cross-resistance information for improved drug resistance prediction by means of multi-label classification were introduced. Therefore, 2 protease sequence and 715 reverse transcriptase sequences along with specific genotypic and phenotypic data have been manipulated using various ML techniques including binary relevance classifiers, classifier chains, and ensembles of classifier chains. Multi-level classification models along with cross-counteraction intelligence were applied to portend the two of the best resistive drug classes used for antiretroviral therapy for the disease HIV-1. The two basic drugs are named as protease inhibitors (PIs) and non-nucleoside reverse transcriptase inhibitors (NNRTIs). Completing overall process the authors have successfully achieve a stage that is quite predicted to be accurate compared to other investigation till today.

Another health problem that obstacle the individuals is the obesity. Recently, obesity becomes a common problem for almost everyone both in developed and developing countries as well [99, 100]. Recent studies have proven the necessity of epistasis or gene-gene interactions for explaining the harmful factor of overweight and obesity. Various network-based models play a great role to explain the basis reason. It is also challenging for the researchers to analyze the data pairwise. The network based algorithms play a great tribute as it has the ability continuing pairwise operation. A network based approach was implemented in [100] which is called statistical epistasis network (SEN) to classify the SNP-SNP interaction in every gene associated with obesity or overweight. The interactions which exceed a specific terminate point build the SEN network. This study represented a traditional properties for identification of genetic interaction from genome based arrays and also invention of various nodes of biological substances to diminish the obesity problem specifying the accurate gene responsible for overweight or obesity problems.

Bacterial vaginosis which is known as vagina microbiomes consists of various types of biological bacteria although all of them a few are quite dangerous for human body [101]. In [102], a relationship between bacterial viagnosis and a

microbial community was extracted. For microbial community, the subsets of every community are justified with various bioinformatics algorithms. For classification and identification of relationship between bacterial vaginosis and microbial communities the logistic regression (LR) and random forests (RF) have been used. The authors successfully represented the relationships between BV and MC by adding some features to machine learning methods of bioinformatics. The ML achieved great accuracy to determine similar patterns; therefore it can be used to detect the similar patterns of bacterial vaginosis and microbial community.

From the preceding survey, it is established that big data mining has a significant role in clinical medicine for prediction of disease development, guidance of rational use of drugs, medical management, and evidence-based medicine as well as disease risk assessment, and clinical decision support. Big data is generated everywhere during monitoring, and medical healthcare and imaging processes as well as from social media applications [102–114]. Machine learning proved its efficiency in predicting and classifying several disease, however big data mining techniques including Dempster–Shafer theory, rough set theory [115], fuzzy theory, artificial neural network [116], cloud theory, inductive learning theory, genetic algorithm, decision tree, Bayesian network, pattern recognition, statistical analysis, and high-performance computing can be studied in future work.

7 Conclusion

Health informatics ensures faster and accurate medical data and symptoms processing. There are several analysis that support information mining from large volume of raw data. Drug design, big data analyses, diabetes factors predictions, cancer gene analyses, machine learning based scoring system, semantic web synthesis, Epigenetic internal functionalities, type-1 HIV intersections, computational obesity simulations and Microbial Communities and Bacterial picture are some areas that are sketched. Each and every area is the key filed that control the better life of human being. Adverse Event Report System (AERS) can be applied to assess the drugs-drugs interactions that help for perfect drug design simulation. For recent adverse disease, the HIV (human immunodeficiency virus) datasets are also verified by binary relevance classier, which dynamically categorized the infected data.

Computational mining and simulations help to get new dimensions in these sectors. Accurate measurements are vital for better health. Now-a-days, robots are frequently used to complete the critical operations of the human body. Moreover, lots of devices are employed to control the exact amount of chemicals during drug design, disease identifications, pathological interactions for HIV monitoring and organic chemical reactions. Subsequently, large data mining approaches for scientific measurement are essential for ever.

References

1. Aenenhaus Arthur, Philippe Cathy, Guillemot Vincent, Cao A. Kim, and Frouin Vincent. 2014. *Variable selection for generalized canonical correlation analysis.* Biostatistics. 15(3): 569–83.
2. Alon, 2003. *Biological networks: the tinkerer as an engineer.* Science, 301:1866–1867.
3. Al-Shahrour Fatema, Minguez Pabel, Vaquerizas M. Jaun, and Dopazo Jaoquin. 2005. *BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments*. Nucleic Acids Res. 33(Web Server issue): W460–4.
4. Alexander Stojadinovic, Anton Bilchik, and Smith David. 2013. *Clinical decision support and individualized prediction of survival in colon cancer: Bayesian belief network model*. Ann Surgery Oncology 20(1):161–74.
5. Ahn Yoel, Bagrow P. James, and Lehmann Sune. 2011. *Link communities reveal multiscale complexity in networks.* Nature 20, 466:761–764.
6. Ashburner Michael, Ball A. Catherine, Blake A. Judith, Botstein David, Butler Heather, and Eppig T. Midori. 2000. *Gene ontology: tool for the unification of biology.* The Gene Ontology Consortium. Natural. Genetics. 25, 25–29.
7. Batagelj Valadimir, and Mrvar Andrej. 1998. *Pajek—Program for Large Network Analysis*. Connections, 21:47–57.
8. Banks A. Charles, Kong E. Stephen, and Washburn P. Michael. 2012. *Affinity purification of protein complexes for analysis by multidimensional protein identification technology*. Protein Expression and Purification, 86:2, 105–119.
9. Bader D. Gray and Hogue W. Christopher. 2003. *An automated method for finding molecular complexes in large protein interaction networks*. BMC bioinformatics, 4:1.
10. Bader D. Gray, and Hogu W. Crishtopher. 2003. CWV: *An automatedmethod for findingmolecular complexes in large protein interaction networks.* BMC Bioinformatics, 4:2.
11. Breitkreutz Bobby, Stark Chris, and Tyers Mike. 2003. *Osprey: a network visualization-system.* Genome Biology, 4(3):R22.
12. Crippen Gordan, Havel F. Timothy. 1988. *Distance Geometry and Molecular* Conformation. New York: Wiley.
13. Cao A. Kim, Rossouw Debra, Robert-Granié Chiristele, and Besse Philippe. 2008. *A sparse PLS for variable selection when integrating omicsdata. Stat Application of Genetic* Molecular Biology, 7(1):35.
14. Cao A. Kim, Martin G. Pascal, Robert-Granié Christele, and Besse Phillippe. 2009. *Sparse canonical methods for biological data integration: application to a cross-platform study*. BMC Bioinformatics.10:34.
15. Chung Dongjun, Chun Hyonho and KelesSunduz. *Sparse Partial Least Squares (SPLS) Regression and Classification.*
16. Zou Dong, Ma Lina, Yu Jun, and Zhang Zhang, 2015. *Biological Databases for Human Research*. Genomics Proteomics Bioinformatics, 13,55–63.
17. Chuang Han, Lee Eunjun, Liu Yu, and Ideker Trey. 2007. *Network-based classification of breast cancer metastasis.* Molecular System Biology, 3:140.
18. Chintapalli R. Venkateswara, Wang Jing, and Dow A. Julian. 2007. *Using FlyAtlas to identify better. Drosophila melanogaster models of human disease*. Natural Genetic. 39:7, 15–20.
19. Chatr-Aryamontri Andrew, Breitkreutz J. Bobby, and Heinicke Sven. 2013. *The BioGRID interaction database: 2013 update.* Nucleic Acids Research, 41:1, D816–D823.
20. Chen S. Ming, Han Jaiwei, and Yu S. Philip. 1996. *Data mining: An overview from a database perspective*. IEEE Trans. Knowledge and Data Engineering, 8:866–883.
21. Chopra Pankaj, Kang Jaewoo, Yang Jiang, and Lee M. Goo. 2008. *Microarray data mining using landmark gene-guided clustering*. BMC Bioinformatics, 9:92.

22. Costa G. Ivan, Krause Roland, Opitz Lennart, and Schliep Alexnder. 2007. *Semi-supervised learning for the identification of syn-expressed genes from fused microarray and in situ image data*. BMC Bioinformatics, 8:10, S3.
23. Croft David, Mundo F. Antonio, Haw Robin, Milacic Marija, Weiser Joel, and Wu Guanming. 2014. *The Reactome pathway knowledgebase*. Nucleic Acids Res. 42:D472–7.
24. Cserháti Tibor, Kósa Agnes, and Balogh Sandor. 1998. *Comparison of partial least-square method and canonical correlation analysis in a quantitative structure-retention relationship study.* Journal of Biochemistry Biophysics Methods. 36(2–3):131–141.
25. D'andrade Roy. 1978. *U-Statistic Hierarchical Clustering*. Psychometrika. 4:58–67.
26. Dahlquist D. Kam. 2004. *Using GenMAPP and MAPPFinder to view microarray data on biological pathways and identify global trends in the data*. Bioinformatics, Chap. 7, Unit 75.
27. Fern X. Zhang Fern and Brodley E. Carla. 2003. *Solving cluster ensemble problems by bipartite graph partitioning*. In Proceedings of the 21st International Conference on Machine Learning: 2003; Banff, Alberta. New York, NY: ACM Press; 182–189.
28. Fruchterman M. Thomas, Reingold M. Edward. 1991. *Graph Drawing by Force-Directed Placement*. Software. Practice and Experience, 21:1129–1164.
29. Frey J. Brendan, and Dueck Delbert. 2007. *Clustering by passing messages between data points*. Science, 315(5814):972–976.
30. Corder, E.H., A.M. Saunders, W.J. Strittmatter, D.E. Schmechel, P.C. Gaskell, GWet Small, A.D. Roses, J.L. Haines, and Margaret A. Pericak-Vance. "Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families." *Science* 261, no. 5123 (1993):921–923.
31. Goate A, Chartier-Harlin MC, Mullan M, Brown J, Crawford F, Fidani L, Giuffra L, Haynes A, Irving N, James L, Mant R. *Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease*. Nature. 1991 Feb 21; 349(6311): 704–6.
32. Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet. 2013; 45(12):1452–8.
33. Younghee Lee, Haiquan Li, Jianrong Li, Ellen Rebman, Ikbel Achour, Kelly E Regan, Eric R Gamazon, James L Chen, Xinan Holly Yang, Nancy J Cox, and Yves A Lussier, Network models of genome-wide association studies uncover the topological centrality of protein interactions in complex diseases, J Am Med Inform Assoc. 2013 Jul; 20(4): pp. 619–629. Published online 2013 Jan 25. doi:10.1136/amiajnl-2012-001519.
34. N. Akula, A. Baranova, D. Seto, Jeffrey, M.A. Nalls, A. Singleton, L. Ferrucci, T. Tanaka, S. Bandinelli, Y.S. Cho, Y.J. Kim, Jong-Young Lee, Bok-Ghee Han, J. McMahon, A Network-Based Approach to Prioritize Results from Genome-Wide Association Studies, Published: September 6, 2011.
35. Goate A, Chartier-Harlin MC, Mullan M, Brown J, Crawford F, Fidani L, Giuffra L, Haynes A, Irving N, James L, Mant R. Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. *Nature*. 1991 Feb 21; 349(6311): 704–6.
36. Rogaev EI, Sherrington R, Rogaeva EA, Levesque G, Ikeda M, Liang Y, Chi H, Lin C, Holman K, Tsuda T, Mar L. Familial Alzheimer's disease in kindreds with missense mutations in a gene on chromosome 1 related to the Alzheimer's disease type 3 gene. *Nature*. 1995 Aug 31; 376(6543):775–8.
37. Bertram L, McQueen MB, Mullin K, Blacker D, Tanzi RE: Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. Nat Genet 2007, 39:17–23.
38. Mullan M, Crawford F, Axelman K, Houlden H, Lilius L, Winblad B, Lannfelt L. A pathogenic mutation for probable Alzheimer's disease in the APP gene at the N–terminus of β–amyloid. Nature genetics. 1992 Aug 1; 1(5):345–7.
39. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH, Multifactor-dimensionality reduction reveals high-order interactions among

estrogen-metabolism genes in sporadic breast cancer, Am J Hum Genet. 2001 Jul; 69 (1): pp. 138–47. Epub 2001 Jun 11.
40. Kononen J, Bubendorf L, Kallionimeni A, Bärlund M, Schraml P, Leighton S, Torhorst J, Mihatsch MJ, Sauter G, Kallionimeni OP. Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nature medicine. 1998 Jul 1; 4(7):844–7.*
41. Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000; 406(6797):747–52. doi:10.1038/35021093.
42. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J ClinOncol. 2009; 27(8):1160–1167. doi:10.1200/JCO.2008.18.1370.
43. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. The Lancet. 2005; 365(9460):671–9. doi:10.1016/S0140-6736(05)17947-1.
44. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012; 486(7403):346–52. doi:10.1038/nature10983.
45. David J. Dittman, Taghi M. Khoshgoftaar, Amri Napolitano, Selecting the Appropriate Ensemble Learning Approach for Balanced Bioinformatics Data, Proceedings of the Twenty-Eighth International Florida Artificial Intelligence Research Society Conference, pp. 329–334.
46. K. Gao, T. Khoshgoftaar, R. Wald, Combining Feature Selection and Ensemble Learning for Software Quality Estimation, Proceedings of the Twenty-Seventh International Florida Artificial Intelligence Research Society Conference, pp. 47–52.
47. Heim S, Mitelman F. Cancer cytogenetics: chromosomal and molecular genetic aberrations of tumor cells. John Wiley & Sons; 2015 Aug 17.
48. Folkman J. Angiogenesis in cancer, vascular, rheumatoid and other disease. Nature medicine. 1995 Jan 1; 1(1):27–30.
49. Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, Bernabé RR, et al. International network of cancer genome projects. Nature. 2010; 464(7291):993–8. doi:10.1038/nature08987.
50. Qian Wang, Jiaxing Zhang, Sen Song, Zheng Zhang, Attentional Neural Network: Feature Selection Using Cognitive Feedback, arXiv:1411.5140v1[cs.CV] 19 Nov 2014.
51. Dean J, Ghemawat S. MapReduce: simplified data processing on large cluster. Communications of the ACM 2008; 51(1): 107–113.
52. IDC, The digital universe in 2020: big data, bigger digital shadows, and biggest growth in the Far East, www.emc.com/leadership/digital-universe/index.htm [last accessed 20 November 2014].
53. Chen C.L.P, Zhang C.Y. Data-intensive applications, challenges, techniques and technologies: a survey on big data, Inform. Sci, doi:10.1016/j.ins.2014.01.015.
54. Chen M, Mao S, Liu Y. Big data survey. Mobile Networks and Applications 2014; 19(2): 171–209.
55. Chen M, Mao S, Liu Y. Big data: a survey. Mobile Networks and Applications. 2014 Apr 1; 19(2):171–209.
56. Jake Luo, Min Wu, Deepika Gopukumar and Yiqing Zhao, Big Data Application in Biomedical Research and Health Care: A Literature Review, libertas academia, freedom to research, published on 19 Jan 2016. doi:10.4137/BII.S31559.
57. Dr. Xin Deng, Dr. Donghui Wu, Big Data Analytic Technology for Bioinformatics and Health Informatics, Call for Papers: Special Session at 2015 IEEE Symposium on Computational Intelligence in healthcare and e-health (IEEE CICARE 2015).
58. Emdad Khan, Addressing Bioinformatics Big Data Problems using Natural Language Processing: Help Advancing Scientific Discovery and Biomedical Research, Modern Computer Applications in Science and Education, pp. 221–228.
59. Cambridge Healthtech Institute's Eighth Annual, Integrated Informatics Driving Translational Research & Precision Medicine, March 7–9, 2016, Moscone North Convention

Center, San Francisco, CA, Part of the 23rd International Molecular Medicine Tri-Conference.
60. Tonia C. Carter and Max M. He, Challenges of Identifying Clinically Actionable Genetic Variants for Precision Medicine, Journal of Healthcare Engineering Volume 2016 (2016), Article ID 3617572.
61. Coiera E. Guide to health informatics. CRC press; 2015 Mar 6.
62. O'donoghue, John; Herbert, John (2012). "Data management within mHealth environments: Patient sensors, mobile devices, and databases". Journal of Data and Information Quality (JDIQ). 4 (1):5.
63. Mettler T, Raptis DA (2012). "What constitutes the field of health information systems? Fostering a systematic framework and research agenda". Health Informatics Journal. 18 (2): 147–56. doi:10.1177/1460458212452496. PMID 22733682.
64. Chen J, Qian F, Yan W, Shen B (2013) Translational biomedical informatics in the cloud: present and future. BioMed Res Int 2013, 8.
65. Brown-Liburd H, Issa H, Lombardi D. Behavioral implications of Big Data's impact on audit judgment and decision making and future research directions. Accounting Horizons. 2015 Jun; 29(2):451–68.
66. Adibi, Sasan, ed. (February 19, 2015). Mobile Health: A Technology Road Map. Springer. ISBN 978-3-319-12817-7.
67. Sohn H, Farrar CR, Hemez FM, Shunk DD, Stinemates DW, Nadler BR, Czarnecki JJ. A review of structural health monitoring literature: 1996–2001. Los Alamos National Laboratory, USA. 2003.
68. Frank J, Di Ruggiero E, Mowat D, Medlar B. Developing knowledge translation capacity in public health. Canadian Journal of Public Health. 2007 Jul; 98(4).
69. Schatz B, Marsh C, Patrick K, et al. Research challenges in measuring data for population health to enable predictive modeling for improving healthcare. ACM SIGHIT Rec. 2012; 2:36–41.
70. Jiang Y, Liao Q, Cheng Q, et al. Designing and evaluating a clustering system for organizing and integrating patient drug outcomes in personal health messages. AMIA Annu Symp Proc. 2012; 2012: 417–426.
71. Becker T, Curry E, Jentzsch A, Palmetshofer W. New Horizons for a Data-Driven Economy: Roadmaps and Action Plans for Technology, Businesses, Policy, and Society. In New Horizons for a Data-Driven Economy 2016 (pp. 277–291). Springer International Publishing.
72. Magoulas, Roger; Lorica, Ben (February 2009). "Introduction to Big Data". Release 2.0. Sebastopol CA: O'Reilly Media (11).
73. Snijders, C.; Matzat, U.; Reips, U.-D. (2012). "'Big Data': Big gaps of knowledge in the field of Internet". International Journal of Internet Science. 7: 1–5.
74. Wild S, Roglic G, Green A, et al. Global prevalence of diabetes estimates for the year 2000 and projections for 2030. Diabetes Care 2004; 27:1047–1053.
75. Centers for Disease Control and Prevention. Estimates of diabetes and its burden in the United States. National Diabetes Statistics Report. Atlanta, GA: US Department of Health and Human Services. 2014.
76. Lindstro M. J, Louheranta A, Mannelin M, et al. The Finnish Diabetes Pre-vention Study (DPS): Lifestyle intervention and 3-year results on diet and physical activity. Diabetes Care 2003; 26:3230–3236. 20.
77. Li G, Zhang P, Wang J, et al. The long-term effect of lifestyle interventions to prevent diabetes in the China Da Qing Diabetes Prevention Study: A 20-year follow-up study. Lancet 2008; 371:1783–1789. 21.
78. Ramachandran A, Snehalatha C, Mary S, et al. The Indian Diabetes Pre-vention Programme shows that lifestyle modification and metformin prevent type 2 diabetes in Asian Indian subjects with impaired glucose tolerance (IDPP-1). Diabetologia 2006; 49:289–297.
79. Kahn HS, Cheng YJ, Thompson TJ, et al. Two risk-scoring systems for predicting incident diabetes mellitus in US adults age 45 to 64 years. Ann Intern Med 2009; 150:741–751. 24.

80. Stern MP, Williams K, Haffner SM. Identification of persons at high risk for type 2 diabetes mellitus: Do we need the oral glucose tolerance test? Ann Intern Med 2002; 136:575–581. 25.
81. Chen L, Magliano DJ, Balkau B, et al. AUSDRISK: An Australian Type 2 Diabetes Risk Assessment Tool based on demographic, lifestyle and simple anthropometric measures. Med J Aust 2010; 192:197–202. 26.
82. Lindstro M J, Tuomilehto J. The diabetes risk score: A practical tool to predict type 2 diabetes risk. Diabetes Care 2003; 26:725–731.
83. Groves P, Kayyali B, Knott D, et al. The 'big data' revolution in healthcare. Boston: McKinsey Quarterly. 2013.
84. Bellman RE. The theory of dynamic programming. Rand Corporation technical report, 1957.
85. van Regenmortel MH, Fauquet CM, Bishop DH, Carstens EB, Estes MK, Lemon SM, Maniloff J, Mayo MA, McGeoch DJ, Pringle CR, Wickner RB. Virus taxonomy: classification and nomenclature of viruses. Seventh report of the International Committee on Taxonomy of Viruses. Academic Press; 2000.
86. National Institutes of Health. Precision medicine initiative cohort program, 2015.
87. National Institutes of Health. Precision Medicine Initiative Cohort Program. 2016.
88. Francis S. Collins, M.D., Ph.D., and Harold Varmus, M.D, A New Initiative on Precision Medicine, the new England journals of medicine, N England J Med 2015; 372:793–795 February 26, 2015 doi:10.1056/NEJMp1500523.
89. Seripa D, Panza F, Daragjati J, Paroni G, Pilotto A. Measuring pharmacogenetics in special groups: geriatrics. Expert opinion on drug metabolism & toxicology. 2015 Jul 3; 11(7): 1073–88.
90. Siobhan Dumbreck, Angela Flynn, Moray Nairn, Martin Wilson, Shaun Treweek, Stewart W Mercer, Phil Alderson, Alex Thompson, Katherine Payne, Guthrie, Drug-disease and drug-drug interactions: systematic examination of recommendations in 12 UK national clinical guidelines, *BMJ* 2015; 350.
91. Caterina Palleria, Antonello Di Paolo, Chiara Giofrè, Chiara Caglioti, Giacomo Leuzzi, Antonio Siniscalchi, Giovambattista De Sarro, and Luca Gallelli, Pharmacokinetic drug-drug interaction and their implication in clinical management, journal of research in medical science, J Res Med Sci. 2013 Jul; 18(7): 601–610.
92. Becker ML, Kallewaard M, Caspers PW, Visser LE, Leufkens HG, Stricker BH. Hospitalisations and emergency department visits due to drug-drug interactions: a literature review. Pharmacoepidemiol Drug Saf. 2007; 16(6):641–51.
93. Daly AK. Pharmacogenomics of adverse drug reactions. Genome med. 2013; 5(1):5.
94. Verma RP, Hansch C, QSAR modeling of taxane analogues against colon cancer, Eur J Med Chem. 2010 Apr; 45(4):1470–7. doi:10.1016/j.ejmech.2009.12.054. Epub 2010 Jan 13.
95. Fauzee NJS, Dong Z, Wang YI. Taxanes: promising anti-cancer drugs. Asian Pac J Cancer Prev. 2011; 12:837–51.
96. Song L, Chen QH, She XE, Chen XG, Wang FP. Conversional synthesis and cytotoxic evaluation of novel taxoid analogs. J Asian Nat Prod Res. 2011; 13(9):787–98.
97. Niko Beerenwinke, Martin Däumer, Mark Oette, Klaus Korn, Daniel Hoffmann, Rolf Kaiser, Thomas Lengauer, Joachim Selbig, and Hauke Walter, Geno2pheno: estimating phenotypic drug resistance from HIV-1 genotypes, Nucleic Acids Res. 2003 Jul 1; 31(13):3850–3855. PMCID: PMC168981.
98. Soo-Yon Rhee, Jonathan Taylor, Gauhar Wadhera, Asa Ben-Hur, Douglas L. Brutlag, Robert W. Shafer, Genotypic predictors of human immunodeficiency virus type 1 drug resistance, vol. 103 no. 46, Soo-Yon Rhee, 17355–17360.
99. Kelly T, Yang W, Chen C-S, Reynolds K, He J. Global burden of obesity in 2005 and projections to 2030. Int J Obes (Lond). 2008; 32:1431–7.
100. Ogden CL, Carroll MD, Kit BK, Flegal KM. Prevalence of childhood and adult obesity in the United States, 2011– 2012. JAMA. 2014; 311:806–14.

101. Ravel J, Gajer P, Abdo Z, Schneider GM, Koenig SS, McCulle SL, et al. Vaginal microbiome of reproductive-age women. Proc Natl Acad Sci. 2011; 108(Supplement 1):4680–687.
102. Kamal S, Ripon SH, Dey N, Ashour AS, Santhi V. A MapReduce approach to diminish imbalance parameters for big deoxyribonucleic acid dataset. Computer methods and programs in biomedicine. 2016 Jul 31; 131:191–206.
103. Kamal S, Dey N, Nimmy SF, Ripon SH, Ali NY, Ashour AS, Karaa WB, Nguyen GN, Shi F. Evolutionary framework for coding area selection from cancer data. Neural Computing and Applications.:1–23.
104. Acharjee S, Dey N, Samanta S, Das D, Roy R, Chakraborty S, Chaudhuri SS. Electrocardiograph Signal Compression Using Ant Weight Lifting Algorithm for Tele-Monitoring. Journal of Medical Imaging and Health Informatics. 2016 Feb 1; 6(1): 244–51.
105. Dey N, Das P, Chaudhuri SS, Das A. Feature analysis for the blind-watermarked electroencephalogram signal in wireless telemonitoring using Alattar's method. In Proceedings of the Fifth International Conference on Security of Information and Networks 2012 Oct 25 (pp. 87–94).
106. Cinque M, Coronato A, Testa A. Dependable services for mobile health monitoring systems. International Journal of Ambient Computing and Intelligence (IJACI). 2012 Jan 1; 4(1):1–5.
107. Van Hoof J, Wouters EJ, Marston HR, Vanrumste B, Overdiep RA. Ambient assisted living and care in The Netherlands: the voice of the user. Pervasive and Ubiquitous Technology Innovations for Ambient Intelligence Environments. 2012 Sep 30:205.
108. Odella F. Technology Studies and the Sociological Debate on Monitoring of Social Interactions. International Journal of Ambient Computing and Intelligence (IJACI). 2016 Jan 1; 7(1):1–26.
109. Tapia DI, Corchado JM. An ambient intelligence based multi-agent system for Alzheimer health care. International Journal of Ambient Computing and Intelligence (IJACI). 2009 Jan 1; 1(1):15–26.
110. Favela J, Tentori M, Segura D, Berzunza G. Adaptive awareness of hospital patient information through multiple sentient displays. International Journal of Ambient Computing and Intelligence (IJACI). 2009 Jan 1; 1(1):27–38.
111. Baumgarten M, Mulvenna M, Rooney N, Reid J. Keyword-Based Sentiment Mining using Twitter. International Journal of Ambient Computing and Intelligence. 2013; 5(2):56–69.
112. Odella F. Technology Studies and the Sociological Debate on Monitoring of Social Interactions. International Journal of Ambient Computing and Intelligence (IJACI). 2016 Jan 1; 7(1):1–26.
113. Araki T, Ikeda N, Dey N, Chakraborty S, Saba L, Kumar D, Godia EC, Jiang X, Gupta A, Radeva P, Laird JR. A comparative approach of four different image registration techniques for quantitative assessment of coronary artery calcium lesions using intravascular ultrasound. Computer methods and programs in biomedicine. 2015 Feb 28; 118(2):158–72.
114. Araki T, Ikeda N, Dey N, Acharjee S, Molinari F, Saba L, Godia EC, Nicolaides A, Suri JS. Shape-based approach for coronary calcium lesion volume measurement on intravascular ultrasound imaging and its association with carotid intima-media thickness. Journal of Ultrasound in Medicine. 2015 Mar 1; 34(3):469–82.
115. Roy P, Goswami S, Chakraborty S, Azar AT, Dey N. Image segmentation using rough set theory: a review. International Journal of Rough Sets and Data Analysis (IJRSDA). 2014 Jul 1; 1(2):62–74.
116. Dey N, Ashour AS, Chakraborty S, Samanta S, Sifaki-Pistolla D, Ashour AS, Le DN, Nguyen GN. Healthy and Unhealthy Rat Hippocampus Cells Classification: A Neural Based Automated System for Alzheimer Disease Classification. Journal of Advanced Microscopy Research. 2016 Jun 1; 11(1):1–0.