A Review Paper on Scope of Big Data Analysis in Heath Informatics

Kazi Md Shahiduzzaman
Assistant Professor, Dept. of Electrical and Electronic Engineering
Jatiya Kabi Kazi Nazrul Islam University, Trishal, Mymensingh
Bangladesh
shahiduzzaman@jkkniu.edu.bd

Lusekelo Kibona
Department of Computer Science
Ruaha Catholic University (RUCU)
Tanzania
lusekelo2012@gmail.com

Hassana Ganame
Department of Information and Telecommunication
School of Engineering of Bamako
Mali
ganame_hassana@yahoo.fr

Abstract— *The term Health Informatics represent a huge volume of data that is collected from different source of health sector. Because of its' diversity in nature, quite a big number of attributes, numerous amount data, health informatics can be considered as Big Data. Therefore, different techniques used for analyzing Big Data will also fit for Health Informatics. In recent years, implementation of Data Mining on Health Informatics brings a lot of fruitful outcomes that improve the overall healthcare system both in analyzing disease and improving healthcare services which eventually reduce expenses. This paper will define the term the health informatics with a detail discussion about different source of heath informatics. Finally, some case study will be illustrated as examples where data mining techniques are applied to produce more efficient, in depth outcomes in analyzing disease.*

Keywords— Big Data, Health Informatics, Data Mining, Electronic Health Record, Gene Signature.

1. INTRODUCTION

Health informatics is an interesting field of research due to its' multi-dimensional attribute and concern of interest. In this digital world the data become an integral part of health informatics. The amount of healthcare data collected continues to grow at unprecedented rates, already reaches yottabyte (10^{24} bytes) levels [5]. No matter how different authors define health informatics it is undeniable that it has entered the Big Data era. Effectively understanding and building knowledge from healthcare data requires developing advanced analytical techniques that can effectively transform data into meaningful and actionable information. The technologies like classification, association analysis, cluster analysis, anomaly detection, can be used to handle big data are very efficient to bring the unlimited potential information from health informatics. Analyzing health data will allow us to understand the patterns that are hidden in the data. Also, it will help the clinicians to build an individualized patient profile and can accurately compute the likelihood of an individual patient to suffer from a medical complication in the near future [2]. Data Mining and Big Data Analysis are the most useful tools for realizing the aim of diagnosis, treatment, helping, and healing all patients in healthcare, improvement in health care output (HCO) [1].

Health informatics is a combination of information science and computer science in the area of healthcare. There are different research areas within the field of health informatics like Bioinformatics, Image Informatics (e.g. Neuro- informatics), Clinical Informatics, Public Health Informatics, and also Translational Bio Informatics (TBI). Health Informatics along with its' subfields, is an enrich data which is derived from different type of source such as sensors, images, text (biomedical literature, clinical notes etc), electronic health records (EHD), omics [3] etc. This heterogeneity in the data collection and representation process leads to numerous challenges in both the processing and analysis of the underlying data. Big data in healthcare and the life sciences is complex, voluminous, and diverse in the data types and speed at which it is generated. These data sets are considered too large and complicated for typical data processing methods, and thus, the term big data was coined. However, when these data are able to be synthesized and analyzed with sophisticated methodology, patterns, trends, and associations can be understood and used to help inform healthcare decisions that may potentially improve patient care, lower costs, and save lives.

The field of healthcare has often seen advances coming from diverse disciplines such as databases, data mining, information retrieval, medical researchers, and healthcare practitioners. While this interdisciplinary nature adds to the richness of the field, it also adds to the challenges in making significant advances. So application of technologies related Big Data Analysis and Data Mining in Health Informatics can be more realistic and efficient solution for getting desire outcome. Outcomes like deliver precision medicine [4] and to create a Learning Healthcare System (LHS). These data are very useful for the clinician to adopt treatment and prevention therapies for an individual patient. Evidence generated from that individual's therapy can serve as data to feed into the LHS. One aim of the LHS is to create an evidence cycle where evidence is generated, applied, and refined. This cycle will create a healthcare system that develops, providing custom-made therapies to individual patients, incorporating knowledge gained from each experience, and ultimately improving the health of persons and populations. Translation becomes bidirectional in LHS [4]. Again, Healthcare organizations and governing bodies can use Big Data analytics to improve the efficiency of healthcare delivery and to reduce costs while improving outcome. For the clinical researcher, the increasing availability of data of all different types presents new opportunities to conduct research to identify important areas where patient care may be improved. So the focus of this paper is on the study of different Big Data techniques which are used in health informatics and their outcomes. For this reason, in this paper, several case study will be illustrated to demonstrate the performance of Big Data analytic and Data Mining.

This paper is organized as follows: Section II provide an introductory idea and basic definitions about Big data and Big Data Analytics (Data Mining). In Section III, different sources of health informatics are describe along with basic analytics technique. In next Section IV, some of the illustration of data mining in health informatics will be presented along with their positive outcomes. In Conclusion the overall summery will be presented along with future possibilities.

2. Big Data and Big Data Analytics

"Big Data" is a new buzzword without a consensus on its definition, and the properties of Big Data have changed from 3Vs, Volume, Velocity, and Variety to 5Vs adding Value and Veracity [8]. Now let's come to the concept of data mining. Data mining is the process of automatically discovering useful information in large data repositories to find novel and useful patterns that might otherwise remain unknown [7]. It is a fundamental part of knowledge discovery in databases (KDD) which converts the raw data into useful information, as shown in the following Fig.1.

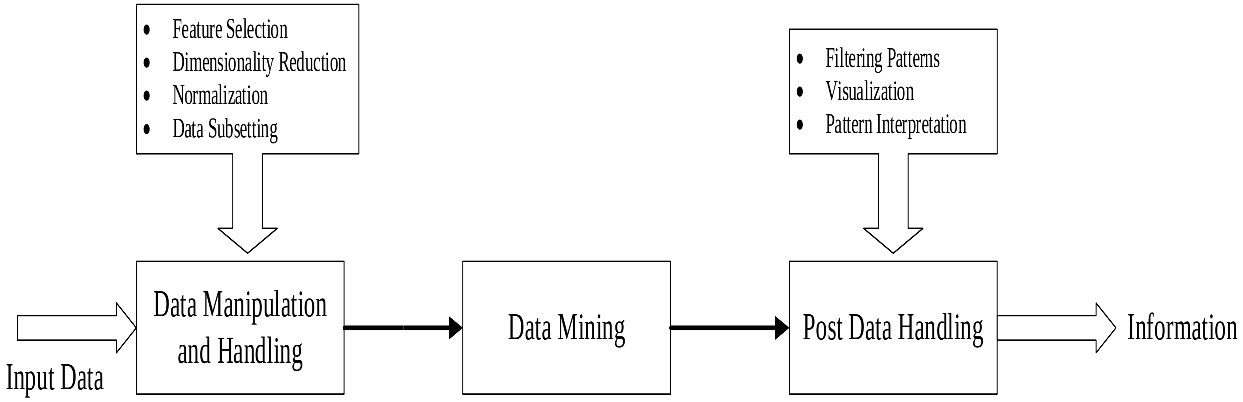
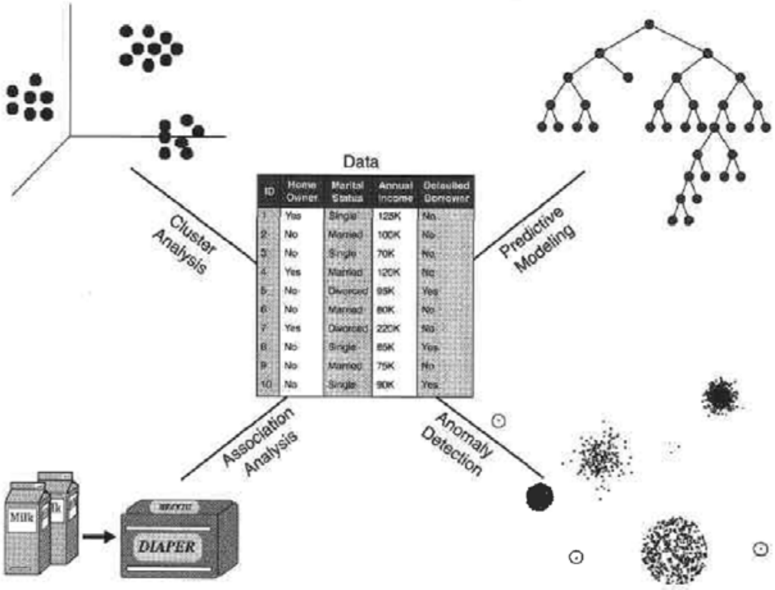


Fig. 1. *Representation of KDD process*

Data mining tasks are generally divided into two major categories, Predictive task which has an objective to predict the value of a particular attribute based on the values of other attributes and Descriptive task which is used to derive patterns (correlations, trends, clusters, trajectories, and anomalies) that summarize the underlying relationships in data. The following Fig.2. illustrate the four core data mining tasks.



ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Fig. 2. *Four Core Data Mining Tasks [p-7,7]*

First consider the predictive modeling which refers to the task of making a model for the target attribute as a function of the explanatory attribute. There are two types of predictive modeling tasks: Classification, which is used for discrete target variables, and Regression, which is used for continuous target variables. The aim of the both task is to make a model that limits the error between the predicted value and true value of the concerned attribute. As the Descriptive modeling is associated with association analysis, cluster analysis and anomaly analysis, so in the remaining paragraph will contains literature review on them. Association analysis is used to discover patterns that describe strongly associated features in the data. The goal of this method is to extract the most interesting patterns in an efficient manner. Next, Cluster analysis seeks to find groups of closely related observations so that observations that belong to the same cluster are more similar to each other than observations that belong to other clusters. Finally, Anomaly detection is the task of identifying observations whose characteristics are significantly different from the rest of the data. The goal of this detection algorithm is to discover the real anomalies and avoid falsely labeling normal objects as anomalous [p 7-11,7].

3. Source of Heath Informatics Data

This section of this paper covers the different health informatics data sources and their impact on analytical algorithms. The heterogeneity of the sources for medical data mining is rather broad, and this creates the need for a wide variety of techniques drawn from different domains of data analytics.

3.1 Electronic Health Record (EHR)

An Electronic Health Record (EHR) is a digital version of a patient’s medical history. It encompasses a full range of data relevant to a patient’s care such as health conditions, medical history, medication, immunizations, radiology reports, laboratory data, any progress report, billing information etc. It enables the administrator to utilize the data for billing purposes, the physician to analyze patient diagnostics information and treatment effectiveness, the nurse to report adverse conditions, and the researcher to discover new knowledge [9]. The different component of EHR is depicted in the Fig.3.

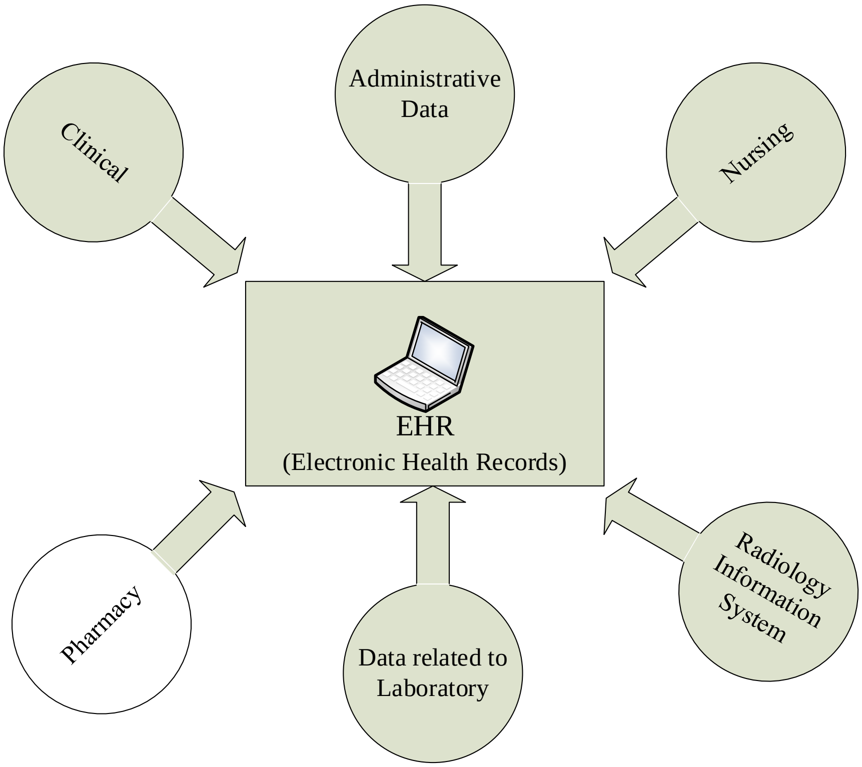

Administrative Data
Clinical
Nursing
EHR
(Electronic Health Records)
Pharmacy
Data related to Laboratory
Radiology Information System

Fig. 3. *Various component of EHR*

Administrative data such as patient registration, admission, discharge, and transfer data are one of the components of the HER. This data allows the aggregation of a person's health information for clinical analysis and research. Next, the data related to Nursing allows a medical practitioner to enter medical orders and instructions for the treatment of a patient. Radiology Information Systems (RIS) are used for managing medical imagery and associated data. RIS is the core database to store, manipulate, and distribute patient radiological data. Laboratory data is a structured data that can be expressed using standard terminology and stored in the form of a name-value pair. Lab data plays an extremely important part in the clinical care process, providing professionals the information needed for prevention, diagnosis, treatment, and health management. Point should be noted the most of the medical decisions are taken based on this data. The pharmaceutical data contains the medication history of a patient such as drug name, dosage, route, quantity, frequency, start and stop date, prescribed by, allergic reaction to medications, source of medication, etc. Finally, the Clinical Documentations are important because documentation is critical for patient care, serves as a legal document, quality reviews, and validates the patient care provided [9].

The primary purpose of EHR data is to support healthcare and administrative services. Information is produced as a byproduct of routine clinical services. They are not a suitable format for performing research tasks. So, to build an algorithm to process this data, first we need to select the interest, followed by the identification of key clinical elements. It may contain billing codes, laboratory and test results, radiology reports, medication history. The gathered information may be combined with a machine learning method. Classification models like Support Vector Machine (SVM), Gaussian Naïve model, Artificial Neural Network (ANN) can be very useful for making true prediction regarding the mater of interest.

3.2 Biomedical Image

Biomedical images are another source of health informatics. These images can be categorized as magnetic resonance imaging (MRI), computed tomography (CT), positron emission tomography (PET), ultrasound (US), or a wide range of microscopy modalities such as fluorescence, bright field, and electron microscopy [10]. Due to the multi-dimensionality and huge amount of data contained in these images, they can be considered as big data and Biomedical image analysis methods enable the extraction of quantitative measurements and inferences from images. Hence, it is possible to detect and monitor certain biological processes and extract information about them. Image segmentation is a process to divide a digital image into separate parts or regions in such a manner that the regions have a strong correlation with objects or areas of the real world contained in the image. This is used to locate objects and boundaries in images. Dividing the image into meaningful regions simplifies the representation of an image into something that is more meaningful and easier to analyze. Segmentation is one of the most important steps leading to the analysis of image data because it enables the further analysis of individual objects [10]. Clustering is an effective way to segment a biomedical image. Through clustering, several discriminative features can be computed on an image, and each pixel can be represented as a point in an N-dimensional space, where the number of dimensions corresponds to the number of features chosen. Again, by using different data mining techniques like Apriori algorithm any desire feature can be extracted from the Bio Medical image.

3.3 Sensor Data in Healthcare

With the advancement in the sensor and wearable technologies several new data sources are available to provide insights on patients beside EHR. Some examples can be heartbeat rate monitor, blood pressure cuffs and Bluetooth enabled different types of devices. Information from these type of device can be very useful for early diagnosis, remote health monitoring, chronic disease management etc. The sensor data can be categorized as physical sensor data, wearable activity sensor data, human sensor data and contextual sensor data [11]. But the sensor data also creates data overload problem. That is why it is essential to complement such sensing capabilities with data mining and analytical capabilities to transform the large volumes of collected data into meaningful and actionable information. The main challenge associated with sensor data is applying models to data. Modeling problems can be classified into six broad categories: (i) anomaly detection to identify statistically deviant data, (ii) association rules to find dependencies and correlations in the data, (iii) clustering models to group data elements according to various notions of similarity, (iv) classification models to group data elements into predefined classes, (v) regression models to fit mathematical functions to data, and (vi) summarization models to summarize or compress data into interesting pieces of information [11]. Again Reality mining [12], is an emerging field of research that uses statistical analysis and machine learning methods on these digital traces to develop comprehensive pictures of our lives, both individually and collectively. Computational models based on this data, combined with any physiological information collected from body sensors and smart environments, can dramatically transform both individual as well as community health [11].

3.4 Other Sources of Heath Informatics

The above sources describe can be considered as the main source of Health informatics. Besides them, there are some sources which need to be mentioned because they are also important sources. First, Biomedical Signal Analysis consists of measuring signals from biological sources, the origin of which lies in various physiological processes. Electroencephalogram (EEG), electrogastrogram (EGG), phonocardiogram (PCG), electroneurogram (ENG), electromyogram (EMG) and electrocardiogram (ECG) are some example of Biomedical signals [2], [13]. A wide variety of methods are used for filtering, noise removal, compact method, Principal Component Analysis (PCA), Singular Value Decomposition (SVD), wavelet transformation etc. are used to analyze the signal to extract the concerned information. Next the Omic Data which is related to our genetic structure of any healthy person or any definite diseases, the genome-wide responses to certain genetic and chemical perturbations, and the large-scale molecular changes that are associated to various disease phenotypes [14]. Various computational methods and information mining tools are used for analyzing the data, particularly focusing on relations between various biological entities and phenotypes. More often such data analyses will lead to novel discoveries and testable hypotheses. Box plots can be used to represent this data to show the distribution of the values of a single numerical attribute [14]. In order to identify meaningful expression patterns from the microarray, clustering methods can be applied to identify if some genes show correlated expression across the given set of biological groups or if some samples share similar gene expression profiles. Like the clustering strategy for identifying gene expression patterns, classification methods can be used to identify gene signatures, which represent a set of genes that can differentiate different biological groups based on the gene expression. Finally, there are some more source like clinical text, biomedical literatures are also source of health informatics [2].

4. Illustration of Data Mining in Health Informatics

In this section some application of the data mining in health informatics are listed.

4.1 Data Mining in EHR data to evaluate the risk factor of IM (Intestinal Metaplasia) in two province of Southeastern China

In this case study, a long term EHR data is considered which is taken from 127,173 [6] upper endoscopies taken place in two provinces named Jiangsu and Anhui in Southeastern China in between 2004 to 2011. The motivation behind this work is that China has a standardized mortality rate of gastric cancer was almost three times than the average global value in 2008 [6] and intestinal metaplasia (IM) is the precancerous lesions of gastric carcinoma (GC) and prevention of GC is very challenging in this area. Therefore, based on the data from endoscopies, data mining algorithm is used to find correlations between IM status and H. pylori infection, atrophic gastritis (AG), dysplasia, age, gender, peptic ulcer, bile reflux, chronic inflammatory severity, degree of acute inflammation, and lymphoid follicle number. Before going for further analysis, it must be mentioned that, from several study it is found that age, smoking, obesity, drinking, H. pylori infection, and bile reflux are risk factors for increasing IM. For example, smokers infected with high-virulence H. pylori strains, the risk of IM was further increased. Similarly, obesity, which was BMI >25 kg/m^2 in males and BMI >27 kg/m^2 in females, was also one of the risk factors of IM [6].

In this big data analytics, the presence of IM is the attribute of interest. The status of IM is scored from 0 (absent) to 1 (mild), 2 (moderate), or 3 (marked). It includes information like the patients' age, gender, images of endoscopic examinations, endoscopic and pathological findings, and the results of rapid urease tests. Odds ratio (OR), chi-square test, and t-test were carried out to examine correlations between IM status and all parameters describe in the above paragraph. The Cochran– Armitage trend test of IM was also carried out. A linear regression analysis was applied to geographical information. All p-values calculated were two-tailed and at a significance level of 0.01 [6].

The results of this data mining shows that the exceptional population size, and the large sample size gave us sufficient statistical power to not only validate conclusions of previous studies but also to reveal some new discoveries. Age, gastric ulcer (GU), H. pylori infection, AG, dysplasia, and GC has positive correlation to IM incidence. Duodenal ulcer (DU) is the only factor with a significant negative relation to IM. From this analysis, bile reflex not only showed a positive relation with IM incidence, but also had an increase of IM level; furthermore, the percentage of patients with bile reflex increased as well [6].

4.2 Application of Association rule mining in Health Informatics

Association rule mining (ARM) is a method to discover meaningful relations between variables in databases which is applicable to application domains such as bioinformatics and medical diagnosis [p-328, 7]. By implementing this method on MATLAB a highly compatible GUI is designed with comprehensive clinical setting to analyze healthcare data [20] and applied on predictive health (PH) model to analyze the mental illness.

PH is a new and innovative healthcare model that focuses on maintaining health rather than curing diseases. Computer-based decision support systems may benefit this domain by providing more quantitative health assessment, enabling more objective advice and action plans from predictive health providers. However, data mining for predictive health is more challenging compared to that for diseases. Because the phenotype of health relies on interactions not limited to biology. Again, PH data also contains measurements from multiple disciplines to provide a comprehensive description of human health which implies information heterogeneity among measurements. This is a common challenge in healthcare data mining. This is a reason why decision support systems are rare in the domain of predictive health.

The developed ARM system is used to generate quantitative and objective rules for health assessment and prediction. The dataset used for this purpose contains 2,637 de-identified health reports from 696 healthy participants with 906 measurement variables. Twelve (12) resulting rules can predict mental illness based on five (5) psychological factors. This study provides vital knowledge to prevent the development of mental illness. For example, if a person has developed depressive symptoms (BDI), providers need to offer proactive advice for the prevention of the potential development of disorders especially in perceived empathic self-efficacy (PSSE) and family functioning (FAD) because they are associated with mental illness risk if comorbid with BDI [20].

4.3 Classification Method Used to Discover Biomarker and Molecular Signature

The large quantities of omics data open the opportunity for developing molecular-level signatures for each known disease phenotype, which could potentially lead to more accurate classifications of the disease into subtype, stage, or grade for the purpose of improved treatment-plan development and prognosis evaluation. With the capability of measuring thousands to millions of parameters, such as gene expression, protein and metabolite abundance, or DNA copy numbers, on a biological sample, the questions of which such features to use for designing the predictive models are crucial. The feature selection is an important step due to the following reasons:

a. A predictive model with many parameters from a limited size of observed samples always lead to poor model in terms of ability to predict for future samples, often referred to as the curse of dimensionality.
b. Smaller numbers of markers used for the predictive model allow the design of the diagnostic devices for cheaper and faster prediction.
c. Predictive models with fewer markers can suggest biological interpretation that may potentially lead to better understanding of the molecular underpinnings of prognosis.

The problem of feature selection has been solved with SVM classifier. All combination of N-genes goes through human genes until the classifier found the molecular signature achieve the desired accuracy. Actually it does not need to search through all the genes encoded in the human genome since the majority of the human genes are not expressed for any specific tissue type. To get a sense about the amount of computing time it may need to exhaustively search through gene combinations. So one needs to go through all possible combination which is too large for a desktop workstation to handle. SVM recursive feature elimination (RFE) method [18], [19] which is very efficient for extracting feature from gene is used here. This model starts from all features and iteratively removes features with small weights estimated by a linear SVM. Basically, one can search for the most N informative genes with a specified feature, to solve the classification problem using a heuristic approach to achieve a desired computational efficiency. The basic idea is to start with a list of genes, each having some discerning power in distinguishing the two classes of samples, and train an optimal classifier with all the genes, followed with a procedure that repeatedly removes genes from the initial gene list as long as the classification accuracy is not affected until the desired number of N genes are left.

In this way numerous N-gene signatures have been identified for diagnostic use, mostly used in cancer cases, including a 70-gene panel for predicting the potential for developing breast cancers, a 21 gene panel, termed Oncotype DX for making the portfolio of Breast, colon and prostate cancer, a 71-gene panel for identification of cancers that are sensitive to TRAIL (TNF-related apoptosis inducing ligand)-induced apoptosis, a 31-gene panel used to predict the metastasis potential of breast cancer and a 16-gene panel for testing for non-small-cell lung cancer against other lung cancer types. Having a test kit for a specific cancer type, e.g., metastasis prone or not, can help surgeons to make a quick and correct decision regarding what surgical procedures to take on the spot. One challenge in identifying signature genes for a specific type of disease using gene expression analysis lies in the proper normalization of transcriptomic data collected by different research labs using different platforms, to ensure that the identified signature genes are generally applicable. Some carefully designed normalization may be needed to correct any systematic effects on gene expression levels caused by different sample preparation and data-collection protocols.

5. CONCLUSION

This review paper discusses about the research that undergoes in Health Informatics sector using Big Data Analytics or Data Mining in recent years. With the advance hardware and efficient algorithms, now data mining is able to handle and produces meaningful outcomes in Big Volume, Velocity, Variety, Veracity, and Value of the data generated by Health Informatics. In this way many unanswered questions in medical sector now come under light which eventually leads to improve the healthcare system. It is clearly demonstrated that how predictive rule like linear regression method correlates the status of Intestinal Metaplasia (IM) with risk factors like age, gender, H. pylori infection etc. Similarly, implementation of Association rule mining on predictive health model to study mental illness which shows interesting and meaningful results, makes a relationship between the factors causing this illness. Again, SVM recursive model is used for the predicting the nature of cancer using genetic informatics. Different number of gene signature is capable of predicting different type of cancer.

The Scope of Big Data Analytics or Data Mining is not only limited to find relationship between disease and its' risk factors, gene structure and so on, but also it can be implemented in improving clinic workflow, patient care, resource management, fleet planning, demand analysis and dynamic deployment strategies, readmission preventive analysis in hospital. All these will reduce the cost for healthcare, increase the healthcare level, provide more satisfactions to patients. So in the end, it is appropriate to conclude that more and more researches like edge computing in healthcare, more efficient and accurate predictive algorithm in different sector, the prospective of Internet of Things (IoT) and IoT enabled device in healthcare can be conducted for more efficient, correlated and improved healthcare.

REFERENCES

[1] Matthew Herland, Taghi M Khoshgoftaar and Randall Wald, 'A review of data mining using big data in health informatics' *Journal of Big Data*, vol. 1, no. 2, p-p. 1-35, 2014.

[2] Chandan K. Reddy, Charu C. Aggarwal '*An Introduction to Healthcare Data Analytics*', in *Healthcare Data Analytics*, Chandan K. Reddy, Charu C. Aggarwal E.d, CRC Press, 2015, pp. 1-18.

[3] Pamela A. Tamez and Mary B. Engler, '*Empowering Clinician Scientists in the Information Age of Omics and Data Science*' in *Actionable Intelligence in Healthcare*, Jay Liebowitz, Amanda Dawson, E.d, CRC Press, 2017, pp. 1-18.

[4] Collins, F. S. and H. Varmus, 'A new initiative on precision medicine', *N Engl J Med, vol.* 372, no. 9, pp no. 793–795. doi: 10.1056/NEJMp1500 523, 2015.

[5] Minjae Kim, '*Data Analytics for the Clinical Researcher* 'in *Actionable Intelligence in Healthcare*, Jay Liebowitz, Amanda Dawson, E.d, CRC Press, 2017, pp. 51-62.

[6] Chao Zhang, Shunfu Xu and Dong Xu, '*Big Health Data Mining*', in '*Health Informatics Data Analysis Methods and Examples*', Dong Xu, May D. Wang, Fengfeng Zhou, Yunpeng Cai, E.d, Springer, 2017.

[7] Pang Ning Tan, Michaelsteinbac, Vipin Kumar, '*Introduction to Data Mining*', Pearson Education Inc, 2006.

[8] Y. Demchenko, P. Grosso, C. de Laat, P. Membrey, Addressing big data issues in scientific data infrastructure, in 2013, International Conference on Collaboration Technologies and Systems (CTS), IEEE, 2013.

[9] Rajiur Rahman and Chandan K. Reddy, '*Electronic Health Records: A Survey*', in *Healthcare Data Analytics,* Chandan K. Reddy, Charu C. Aggarwal E.d, CRC Press, 2015, pp. 21-51.

[10] Dirk Padfield, Paulo Mendonca, and Sandeep Gupta, '*Biomedical Image Analysis*', in *Healthcare Data Analytics*, Chandan K. Reddy, Charu C. Aggarwal E.d, CRC Press, 2015, pp. 61-85.

[11] Daby Sow, Kiran K. Turaga, Deepak S. Turaga, and Michael Schmidt, '*Mining of Sensor Data in Healthcare: A Survey*', in *Healthcare Data Analytics*, Chandan K. Reddy, Charu C. Aggarwal E.d, CRC Press, 2015, pp. 91-117.

[12] *A. Pentland, D. Lazer, D. Brewer, and T. Heibeck*, '*Improving Public Health and Medicine by Use of Reality Mining*', Studies in Health Technology Informatics, vol. 149, pp. 93-102, 2009.

[13] Abhijit Patil, Rajesh Langoju, Suresh Joel, Bhushan D. Patil, and Sahika Genc, '*Biomedical Signal Analysis*', in *Healthcare Data Analytics*, Chandan K. Reddy, Charu C. Aggarwal E.d, CRC Press, 2015, pp. 127-176.

[14] Juan Cui, ‘*Genomic Data Analysis for Personalized Medicine*’, in *Healthcare Data Analytics*, Chandan K. Reddy, Charu C. Aggarwal E.d, CRC Press, 2015, pp. 187-207.

[15] Kalpana Raja and Siddhartha R. Jonnalagadda, ‘*Natural Language Processing and Data Mining for Clinical Text*’, in *Healthcare Data Analytics*, Chandan K. Reddy, Charu C. Aggarwal E.d, CRC Press, 2015, pp. 219-242.

[16] Claudiu Mihaila, Riza Batista-Navarro, Noha Alnazzawi, Georgiosˇ Kontonatsios, Ioannis Korkontzelos, Rafal Rak, Paul Thompson, and Sophia Ananiadou, ‘*Mining the Biomedical Literature*’, in *Healthcare Data Analytics*, Chandan K. Reddy, Charu C. Aggarwal E.d, CRC Press, 2015, pp. 251-284.

[17] Alexander Kotov, ‘*Social Media Analytics for Healthcare*’, in *Healthcare Data Analytics*, Chandan K. Reddy, Charu C. Aggarwal E.d, CRC Press, 2015, pp. 309-333.

[18] Xue-wen Chen and Jong Cheol Jeong, ‘*Enhanced Recursive Feature Elimination*’ presented in *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*, Dec. 13-15, 2007.

[19] Joanna Goscik, Tomasz Łukaszuk, ‘Application of The Recursive Feature Elimination and The Relaxed Linear Separability Feature Selection Algorithms to Gene Expression Data Analysis’, Advances in Computer Science Research, vol. 10, pp. 39-52, 2013.

[20] Chih-Wen Cheng and May D. Wang, ‘*Healthcare Data Mining, Association Rule Mining, and Applications*’, in ‘*Health Informatics Data Analysis Methods and Examples*’, Dong Xu, May D. Wang, Fengfeng Zhou, Yunpeng Cai, E.d, Springer, 2017.