International Journal of Recent Engineering Research and Development (IJRERD)
ISSN: 2455-8761
www.ijrerd.com || Volume 02 – Issue 08 || August 2017 || PP. 20-27

Data Mining Techniques for Building HealthCare Information System (HCIS)

[1]Andemariam Mebrahtu, [2]Balu Srinivasulu

[1]Lecturer, Department of Computer Science, Eritrea Institute of Technology, Eritrea, Northern East Africa
[2]Lecturer, Department of Computer Science, Eritrea Institute of Technology, Eritrea, Northern East Africa

Abstract: The successful application of data mining in highly visible fields like e-business, marketing and retail have led to the popularity of its use in knowledge discovery in databases (KDD) in other industries and sectors. Among these sectors that are just discovering data mining are the fields of medicine and public health. Use of information technology enables automation of data mining and knowledge that help bring some interesting patterns which means eliminating manual tasks and easy data extraction directly from electronic records, electronic transfer system that will secure medical records, save lives and reduce the cost of medical services as well as enabling early detection of infectious diseases on the basis of advanced data collection. Health Care Information System which provide an effective way to solve the problem of managing clinical data is stressed. In this paper it is shown how the health care industry is solving their problems through data mining techniques. Various learning methods in data mining, data mining tasks, importance of data mining in health care industry, forecasts and issues in health care industry are discussed.

Keywords: Data Mining, Health care, Clustering, Classification, Extraction, Predictions, KDD

I. INTRODUCTION

Today, Health organizations are capable of generating and collecting a large amount of data. This increase in data volume automatically requires the data to be retrieved when needed. With the use of data mining techniques is possible to extract the knowledge and determine interesting and useful patterns. The knowledge gained in this way can be used in the proper order to improve work efficiency and enhance the quality of decision making. Above the foregoing is a great need for new generation of theories and computational tools to help people with extracting useful information from the growing volume of digital data [1]. Health Care Information System (HCIS) has in fact been playing a minor role in the industry for many years but has yet to be implemented successfully end-to-end because of the many hurdles it has faced such as privacy concerns, cost and simply the lack of technology. Health care delivery system adopts information technology; vast quantities of health care data become available to mine for valuable knowledge. Healthcare data includes patient centric data, their treatment data and resource management data. It is very massive and information rich. Health care organizations generally adopt information technology to reduce costs as well as improve efficiency and quality. Medical researchers hope to exploit clinical data to discover valuable knowledge i.e. hidden relationship lying implicitly in individual patient health records. The trends in data can be discovered from the application of data mining techniques on healthcare data. These new uses of clinical data potentially affect healthcare because the patient-physician relationship depends on very high levels of trust. To operate effectively physicians need complete and accurate information about the patient for clinical outcome and operational efficiency analysis.

Data mining techniques have been used in healthcare research and known to be effective. It is especially used when it draws information from multiple sources posing special problems. Data mining is the application of algorithms for extracting patterns from large volume of data. There is a wealth of data available within the healthcare systems [2]. For example, hospitals and physicians are commonly required to report certain information for a variety of purposes from census to public health to finance. This often includes patient number, Postal code, sex, date of birth, age, service date, diagnoses codes, procedure codes (CPT), as well as physician identification number, physician postal code, and total charges. Compilations of this data have been released to industry and researchers. The healthcare environment is information rich yet knowledge poor.

Thanks to this technique, it is possible to predict trends and behavior of patients or diseases. This is done by analyzing data from different perspectives and finding connections and relationships between seemingly unrelated information. In the process of data mining previously unknown trends and patterns from a database of information are discovered and transform information into meaningful solutions [3].

International Journal of Recent Engineering Research and Development (IJRERD)
ISSN: 2455-8761
www.ijrerd.com || Volume 02 – Issue 08 || August 2017 || PP. 20-27

II. LEARNING METHODS IN DATA MINING

Data mining is one among the most important steps in the knowledge discovery process. Various steps involved in KDD iterative process is given below [1]:

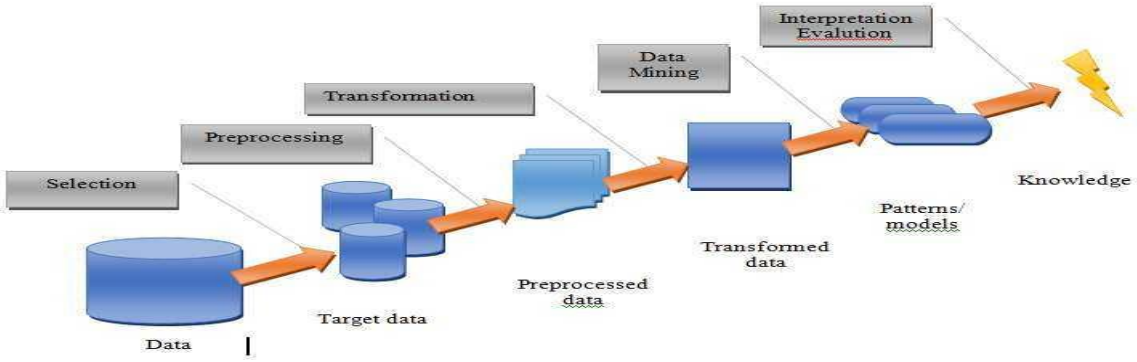

Fig.1: KDD Process [1]

Data mining techniques fall into two broad categories: unsupervised and supervised. Unsupervised learning refers to the technique that is not guided by any particular variable or class label. In the unsupervised learning, we do not create a model or hypothesis prior to the analysis. We apply the algorithm directly to the data and observe the results. A model will then be built according to the results. Thus, unsupervised leaning is used to define class for data without class assignments. Clustering is one of the common unsupervised techniques.

In contrast, for supervised learning, a model is built prior to the analysis. We then apply the algorithm to the data in order to estimate the parameters of the model. The objective of building models using supervised learning is to predict an outcome or category of interest. Classification, statistical regression and association rules building are very common supervised learning techniques used in medical and clinical research. Table 1 is the summary comparing the characteristics and the techniques used for the two different learning methods. Followed is a brief explanation of the four learning techniques.

TABLE I : COMPARING THE CHARACTERISTICS AND THE TECHNIQUES OF THE UNSUPERVISED AND SUPERVISED LEARNING

	Characteristics	Techniques
Unsupervised Learning	□ No guidance □ Use to Define the class □ Seldom utilized (until recently)	□ Clustering □ Association Rule
Supervised Learning	□ With guidelines □ Class defined □ Common with vast literature and application	□ Classification □ Statistical Regression □ Artificial neural networks

A. Clustering

Clustering is an unsupervised learning technique revealing natural groupings in the data. Cluster analysis refers to the grouping of a set of data objects into clusters. A cluster is a collection of data objects which are similar to one another within the same cluster but not similar to the objects in another cluster. Clustering is also called unsupervised classification where no predefined classes are assigned.

B. Association Rule

Association rule discovery is to find the relationships between the different items in a data base. It is normally express in the form X => Y, where X and Y are sets of attributes of the dataset which implies that transactions that contain X also contain Y.

C. Classification

Classification is a supervised learning method. It is a method of categorizing or assigning class labels to a pat-tern set under the supervision. The object of classification is to develop a model for each class. Classification methods can usually be categorized as follows:

1) Decision Tree: Decision tree classifiers divide a decision space into piecewise constant regions. It splits a dataset on the basis of discrete decisions, using certain thresholds on the attribute values. It is one of the most widely used classification method as it is easy to interpret and can be represented under the If-then-else rule condition.

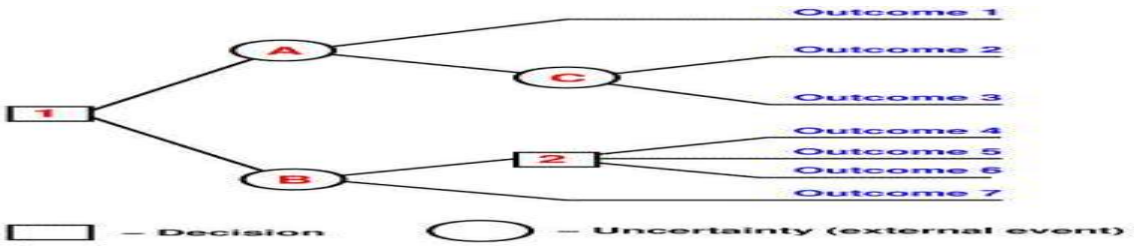

Fig.2 : Decision tree analysis

2) Nearest-Neighbor: Nearest-neighbor classifiers [4] typically define the proximity between instances, find the neighbors if a new instance, and then assign to it the label for the majority class of its neighbors.

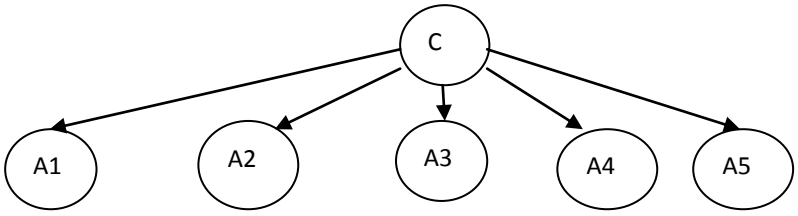

Fig.3: Nearest-Neighbor Algorithm

3) Probabilistic Models: Probabilistic models are models which calculate probabilities for hypotheses base on Bayes' theorem [4]. A Naïve Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. For example, a fruit may be considered to be a tomato if it is red in color, round in shape, and about 3" in diameter. This classifier takes all these features to contribute independently to the probability that this fruit is a tomato, whether or not they're in fact related to each other or to the existence of the other features.

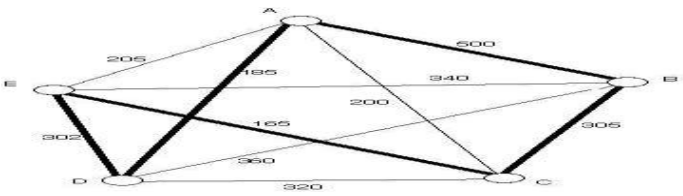

Fig. 4: Representation of a Bayesian Classifier Structure

D. Statistical Regression

Regression models are very popular in the biomedical literature and have been applied in virtually every sub-specialty of medical research. Before computers were widely used, linear regression was the most popular model to find solutions of the problem of estimating the intercept and coefficients of the regression question. It has solid foundation from the statistical theory. Linear regression is similar to the task of finding the line that minimizes the total distance to a set of data. That is find the equation for line Y = a + bX. With the help of computers and software package, we can calculate the high complex models.

E. Artificial Neural Networks

Artificial neural networks [5] are signal processing systems that try to emulate the behavior of human brain by providing a mathematical model of combination of numerous neurons connected in a network. It learns through examples and discriminate the characteristics among various pattern classes by reducing the error and automatically discovering inherent relationships in a data-rich environment. No rules or programmed information is need beforehand. It composes of many elements, called nodes which are connected in between. The connection between two nodes is weighted and by the adjustment of these weights, the training of the

network is performed. The weights are network parameters and their values are obtained after the training procedure. There are usually several layers of nodes. During the training procedure, the inputs are directed in the input layer with the desirable output values as targets. A comparison mechanism will operates between the out and the target value and the weights are adjusted in order to reduce error. The procedure is repeated until the network output matches the targets. There are many advantages of neural networks like adaptive learning ability, self-organization, real-time operation and insensitivity to noise. However, it also has a huge disadvantage that it is highly dependence on the training data and it does not provide an explanation for the decisions they make, just like working in the 'black box'.

F.Time - Series Analysis

Time series is a sequence of events where the next event is determined by one or more of the preceding events. Time series reflects the process being measured and there are certain components that affect the behavior of a process. Time series analysis includes methods to analyze time-series data in order to extract useful patterns, trends, rules and statistics. Stock market prediction is an important application of time- series analysis.

G. Summarization

Summarization is the generalization of data. A set of relevant data is summarized which result in a smaller set that gives aggregated information of the data. For example, the shopping done by a customer can be summarized into total products, total spending, offers used, etc. Such high level summarized information can be useful for sales or customer relationship team for detailed customer and purchase behavior analysis. Data can be summarized in different abstraction levels and from different angles.

H. Advanced Data Mining Techniques

During the past few years, researchers have tried to combine both unsupervised and supervised methods for the analysis [6]. Some examples of advanced unsupervised learning models are hierarchical clustering, c-means clustering self-organizing maps (SOM) and multidimensional scaling techniques. Advanced examples of the supervised learning models classification and regression trees (CART) and support vector machines [7].

III. THE DATA MINING TASK

The data mining tasks are different types depending on the use of data mining result the data mining tasks are classified as:

A. Exploratory Data Analysis

In the repositories vast amount of information's are available .This data mining task will serve the two purposes

(i)Without the knowledge for what the customer is searching, then

(ii) It analyze the data these techniques are interactive and visual to the customer.

B. Descriptive Modeling

It describe all the data, it includes models for overall probability distribution of the data, partitioning of the dimensional space into groups and models describing the relationships between the variables.

C. Predictive Modeling

This model permits the value of one variable to be predicted from the known values of other variables.

D. Discovering Patterns and Rules.

This task is primarily used to find the hidden pattern as well as to discover the pattern in the cluster. In a cluster a number of patterns of different size and clusters are available .The aim of this task is "how best we will detect the patterns". This can be accomplished by using rule induction and many more techniques in the data mining algorithm like K-Means. These are called the clustering algorithm.

E. Retrieval by Content

The primary objective of this task is to find the data sets of frequently used in the for audio/video as well as images It is finding pattern similar to the pattern of interest in the data set.

IV. BUILDING HEALTH CARE INFORMATION SYSTEM

Healthcare is a very research intensive field and the largest consumer of public funds. With the emergence of computers and new algorithms, health care has seen an increase of computer tools and could no

longer ignore these emerging tools. This resulted in uniting of healthcare and computing to form health care information system. This is expected to create more efficiency and effectiveness in the health care system, while at the same time, improve the quality of health care and lower cost.

Health informatics is an emerging field. It is especially important as it deals with collection, organization, storage of health related data. With the growing number of patient and health care requirements, having an automated system will be better in organizing, retrieving and classifying of medical data. Physicians can input the patient data through electronic health forms and can run a decision support system on the data input have an opinion about the patient's health and the care required. An example in the advances in health informatics can be the diagnosis of a patient is health by a doctor practicing in another part of the world. Thus healthcare organizations can share information regarding a patient which will cut costs for communication and at the same time be more efficient in providing care to the patient.

There are other issues like data security and privacy, which is equally important when considering health related data. Thus HCIS "deals with biomedical information, data, and knowledge--their storage, retrieval, and optimal use for problem solving and decision making". This is a highly interdisciplinary subject where fields in medicine, engineering, statistics, computer science and many more come together to form a single field. With the help of smart algorithms and machine intelligence we can provide the quality of healthcare by having, problem solving and decision-making systems. Information systems can help in supporting clinical care in addition to helping administrative tasks. Thus the physicians will have more time to spend with the patients rather than filling up manual forms.

V. THE IMPORTANCE AND USES OF DATA MINING IN HEALTH CARE

Despite the differences and clashes in approaches, the health sector has more need for data mining today. There are several arguments that could be advanced to support the use of data mining in the health sector, covering not just concerns of public health but also the private health sector.

Data overload-There is a wealth of knowledge to be gained from computerized health records. Yet the overwhelming bulk of data stored in these databases makes it extremely difficult, if not impossible, for humans to sift through it and discover knowledge. In fact, some experts believe that medical breakthroughs have slowed down, attributing this to the prohibitive scale and complexity of present-day medical information. Computers and data mining are best-suited for this purpose.

Evidence-based medicine and prevention of hospital errors-When medical institutions apply data mining on their existing data, they can discover new, useful and potentially life-saving knowledge that otherwise would have remained inert in their databases. By mining hospital records, such safety issues could be flagged and addressed by hospital management and government regulators.

Policy-making in public health **Lavrac et al.** (2007) combined GIS and data mining using among others, Weka with J48 (free, open source, Java-based data mining tools), to analyze similarities between community health centers in Slovenia. Using data mining, they were able to discover patterns among health centers that led to policy recommendations to their Institute of Public Health. They concluded that "data mining and decision support methods, including novel visualization methods, can lead to better performance in decision making." Applying data mining in the medical field is a very challenging undertaking due to the idiosyncrasies of the medical profession. Shillabeer and Roddick's work (2007) cite several inherent conflicts between the traditional methodologies of data mining approaches and medicine. In medical research, data mining starts with a hypothesis and then the results are adjusted to fit the hypothesis. This diverges from standard data mining practice, which simply starts with the data set without an apparent hypothesis.

EMR – *Electronic medical Records used in Health Care information system.* Also, whereas traditional data mining is concerned about patterns and trends in data sets, data mining in medicine is more interested in the minority that do not conform to the patterns and trends. What heightens this difference in approach is the fact that most standard data mining is concerned mostly with describing but not explaining the patterns and trends. In contrast, medicine needs those explanations because a slight difference could change the balance between life and death.

For example, anthrax and influenza share the same symptoms of respiratory problems. Lowering the threshold signal in a data mining experiment may either raise an anthrax alarm when there is only a flu outbreak. The converse is even more fatal: a perceived flu outbreak turns out to be an anthrax epidemic (Wong et al 2005). It is no coincidence that we found that, in most of the data mining papers on disease and treatment, the conclusions were almost-always vague and cautious.

VI. FORECASTING IN HEALTH SECTOR

Health forecasting is predicting health situations or disease episodes and forewarning future events. It is also a form of preventive medicine or preventive care that engages public health planning and is aimed at

facilitating health care service provision in populations [8]. Health forecasting has been commonly applied to emergency department visits, daily hospital attendance and admissions [9].

There are important terms in forecasting that are worth noting because of the way in which they are used across various fields. The term *prediction* is mainly used across several fields of study to mean an opinion-based speculation with no explicit causal assumptions [10]. In the health forecasting literature, however, the terms *prediction* and *prognosis* could mean different things, even though they are sometimes used interchangeably and without clarity. The term *prognosis* refers to a forecasting of outcomes under no intervention, whilst *prediction* is used to mean forecasting health outcomes that are associated with some health-related intervention [11]. *Syndromic surveillance* is another closely related concept that is well known in disease surveillance literature. This concept focuses on case detection and events that lead to/precede an outbreak, and involves detecting aberrations in the patterns of diseases and using this information to determine future outbreaks [12, 13]. Syndromic surveillance was initially developed as an innovative electronic surveillance system and was aimed at improving early detection of outbreaks attributable to biologic terrorism or other causes [12].

A. Align Staffing with Predicted Patient Demand

One of the best ways to sustainably reduce labor costs is to forecast demand far enough in advance to match staff and resources without incurring last-minute expenses. How can you ensure you have the right mix of staff on any given day unless you can accurately predict the number and types of patients that staff will be caring for?A reliable system provides the tools required to monitor progress against the forecast and adjusts as necessary to stay on course. It quickly enables you to see where you will be understaffed or over bedded and plan accordingly.

B. Planning Across Horizons When Forecasting in Healthcare

A patient demand forecasting tool must be accurate enough to help drive strategic, budgetary, scheduling, patient flow and staffing decisions. An optimal forecast enables you to plan continuously over multiple time horizons: [14].

Strategic (3–5 years) and **budgetary** (1–2 years) forecasts are used for long-term planning based on historical trends

Scheduling (1–4 months) and **operational** (today to 7 days out) projections enable resource managers to adjust near-term plans by also accounting for current hospital status

Even a 3–5 day view into future demand and resource requirements gives leaders more time to consider a variety of less expensive options for workforce optimization. For example, rather than adding last-minute staff to cover a spike in demand, you may be able to delay elective surgeries or re-sequence scheduled tests. If you still need to increase staff, you can do so based on straight time.

C. Forecasting Hospital Demand to Anticipate Peaks and Valleys

Demand forecasting also improves patient throughput and operational efficiency. When you can anticipate peaks and valleys, you can schedule the right staff and even flex units up and down well in advance. Armed with consistently accurate demand forecasts, planners can notify resource managers of predicted low census periods in time for them to reconfigure units and redeploy staff. Units with seasonal occupancy can be closed and repurposed as an observation unit or another specialty unit. Units with consistently low censuses can be merged to create economies of scale.

D. Workforce Optimization through Flexible Scheduling

Moving to staggered shifts based on forecasted patient demand is one of the most effective ways to eliminate inefficiencies, such as routine overstaffing, that are inherent in budget-driven schedules with pre-set intervals [14].

An effective hospital capacity planning tool should be able to forecast patient arrival times and flow throughout their stay and translate this anticipated activity into facility and staffing requirements:

- Planners can staff accordingly, moving from 8-hour to 4-hour shifts where appropriate to match resources more closely to fluctuating demand
- Shifts changes can be designed to avoid peak arrival and discharge patterns
- In addition to forecasts, the system should provide nurse managers with information on real-time patient arrivals by the hour, with adjustments to the staffing plan already incorporated.

E. Data Accuracy: Essential to Building a Proactive Culture

Some accuracy analyses compare the average monthly census of the model to the average actual census. This aggregation trick can make a model appear much more accurate than it will be in practice. To be statistically valid, accuracy calculations must be based on an absolute daily comparison of the forecast to what actually happened [14].

Your forecasting methodology should be transparent to all stakeholders so that they trust the data and resulting projections. It should enhance the skill of clinical leaders by pointing out needed actions in advance and expanding the options available to solve operational problems. As staff see the forecast tracking to actual demand and begin to experience a less chaotic environment, a proactive culture built on healthcare analytics can take hold.

VII. CONCLUSION

This research briefly reviewed the various data mining techniques in clinical data. The study would be helpful to researchers to focus on the various issues of data mining. In further research direction will be the various classification algorithms and significance of evolutionary computing approach in designing of efficient classification algorithms for data mining. Most of the previous studies on data mining applications in various fields use the variety of data types range from text to images and stores in variety of databases and data structures. The different methods of data mining are used to extract the patterns and thus the knowledge from this variety databases. Selection of data and methods for data mining is an important task in this process and needs the knowledge of the domain. Several attempts have been made to design and develop the generic data mining system but no system found completely generic. Thus, for every domain the domain expert's assistant is mandatory. The domain experts shall be guided by the system to effectively apply their knowledge for the use of data mining systems to generate required knowledge. The domain experts are required to determine the variety of data that should be collected in the specific problem domain, selection of specific data for data mining, cleaning and transformation of data, extracting patterns for knowledge generation and finally interpretation of the patterns and knowledge generation. Most of the domain specific data mining applications show accuracy above 90%. The generic data mining applications are having the limitations. From the study of various data mining applications it is observed that, no application called generic application is 100 % generic. The intelligent interfaces and intelligent agents up to some extent make the application generic but have limitations. The domain experts play important role in the different stages of data mining. The decisions at different stages are influenced by the factors like domain and data details, aim of the data mining, and the context parameters. The domain specific applications are aimed to extract specific knowledge. The domain experts by considering the user's requirements and other context parameters guide the system. The results yields from the domain specific applications are more accurate and useful. Therefore it is conclude that the domain specific applications are more specific for data mining. From above study it seems very difficult to design and develop a data mining system, which can work dynamically for any domain.

REFERENCES

[1]. Fayyad, U., Shapiro, G. P., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence, 1-54.

[2]. BanuRahaman S. and Shashi M., "Sequential mining equips e-Health with knowledge for managing diabetes, "4th International Conference on New Trends in Information Science and Service Science(NISS), 2010, pp. 65-71.

[3]. boirefillergroup.com. (2010). Data Mining Methodology. Retrieved 06 12, 2012,from Boire Filler Group: http://www.boirefillergroup.com/methodology.php.

[4]. J. T. Tou and R. C. Gonzalez, "Pattern recognition principles," Addison-Wesley, London, 1974.

[5]. R. O. Duda, P. E. Hart, and D. G. Stork, "Pattern classification," Wiley, 2001.

[6]. T. Hastie, R. Tibshirani, and J. Friedman, "The elements of statistical learning: Data mining, inference, and prediction," Springer, New York, 2001.

[7]. J. W. Lee, J. B. Lee, M. Park, and S. H. Song, "An extensive comparison of recent classification tools applied to microarray data," Computational Statistics & Data Analysis, Vol. 48, No. 4, pp. 869–885, 2005.

[8]. Met-Office. The Met Office health forecasting services, Exeter. 2009. http://www.metoffice.gov.uk/health/. Accessed 01 Jan 2009.

[9]. Boyle J, Jessup M, Crilly J, Green D, Lind J, Wallis M, et al. Predicting emergency department admissions. Emerg Med J. 2011

[10]. Gentry L, Calantone RJ, Cui SA. The forecasting classification grid: a typology for method selection. J Glob Bus Manag. 2006; 2(1):48–60.

[11]. Poikolainen K. A comment diagnosis as a means of health forecasting. SocSci Med Part A Med Psychol Med Sociol. 1979;13:165–166. [PubMed]
[12]. Henning KJ. What is syndromic surveillance? MMWR Morb Mortal Wkly Rep. 2004; 53(Suppl):5–11. [PubMed]
[13]. Unkel S, Farrington CP, Garthwaite PH, Robertson C, Andrews N. Statistical methods for the prospective detection of infectious disease outbreaks: a review. J R Stat SocSer A. 2011; 175(1):49–82.
[14]. http://sites.mckesson.com/enterpriseintelligence/staff-to-forecasted-demand.htm.

AUTHOR BIBLIOGRAPHY

Mr. Andemariam Mebrahtu, Currently I am working as Lecturer and Head of Computer Science Department , College of Science in Eritrea Institute of Technology, Asmara, Eritrea, Northern East Africa. I have sound experience in teaching, academic Administration activities and research in field of Computer Science. I published a number of international journal papers related to the Computer Science. My research area includes Cloud Computing, Data Mining and Big Data Management.

Mr. Balu Srinivasulu currently I am working as a Lecturer in the Department of Computer Science, Eritrea Institute of Technology, Asmara, Eritrea. I have wide experience of teaching and research in field of Computer Science. I have published a number of international journal papers related to the Computer Science. My areas of research are Wireless Networks, Communication Networks, Big Data, Remote Processing automation(RPA) using Blue Prism and Cloud Computing.