JES Journal of Engineering Sciences

Vol 11, Issue 5,May/2020
ISSN NO:0377-9254

Survey on using Electronic Medical Records (EMR) to Identify the Health Conditions of the Patients

J. Jannathul Firthous[1] and M. Mohamed Sathik[2]

[1]Research Scholar, Sadakathullah Appa College, Affiliated to Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India

[2]Prinicipal and Research Supervisor, Sadakathullah Appa College, Tirunelveli, Affiliated to Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India

***Abstract*— In medical industry, the research on computer aided disease diagnosis is growing expeditiously. The identification of diseases is not as easy as the pathologies of different diseases cannot be examined quickly. Usually the disease can be diagnosed based on the symptoms caused to the patients but for few diseases the symptoms will be common and differentiation of diseases cannot be determined easily until upon doing several tests. Apart from this, the medical errors may occur due to several reasons such as miscommunications between the patients and clinicians, deficiency in the collaboration of Health Information Technologies, inadequate healthcare systems for diagnosis etc. For reducing these entanglements, the machine learning models can be implemented for analyzing various cause and effects of different diseases. Machine Learning is the study of statistical models and algorithms which provides the ability to the computing systems to learn and improve automatically without any explicit programming. Current technologies of Artificial Intelligence use several methods for diagnosing the disease such as chatbots, oncology, pathology, rare diseases etc. This work indulges in diagnosing the diseases and side effects by analyzing the patient's reports from different sources. These collected data are called Electronic Medical data. EMR can be of structured, semi- structured and unstructured formats. The intention of this survey is to know how Electronic Medical Records (EMR) and machine learning algorithm helps in identifying the diseases and side effects from the diseases.**

***Keywords*— Healthcare Analysis, Disease Diagnosis, Disease Identification, Patient Report Analysis.**

1. INTRODUCTION

Health care is the important field of study which improves quality of the human life as well as all living beings. The process of identifying, diagnosing, preventing and treatment for diseases using machine learning techniques helps the physicians and healthcare industry to improve the health services. Several surveys in the recent time shows that the modern health care industry based on machine learning algorithms improves the quality and reduces the mortality rate, cost and complications in a considerable amount. During 2009, the US government has created a Health Information Technology for Economic and Clinical Health Act (HITECH) which includes an incentive scheme which is around 27 billion dollar for using Electronic Medical Records. The advances in IT industry encompass the ability of collecting health care data that exists in various forms. Data seems to be an integral part of the healthcare field. A report by Google comments on Big Data that the existing health care data has the potential of 300 billion dollars because of the advancements in the technologies that enables sensing and acquisition of data. The healthcare institutions or organizations and hospitals are collecting patient's health care data [1]. The advanced analytical techniques needs to be developed for better understanding and knowledge gaining of the health care data as it may transform the existing data to a meaningful information. Data analysis forms a critical component of these emerging computing technologies. The solutions that are observed from the analysis is then applied to the available health care data which has the potential for transforming the health care from reactive state to proactive state. It is predicted that for several years the health care analysis will grow more and more. Typically, the underlying patterns of several diseases can be observed and understood by analyzing the health data. This allows the physicians to build a

personalized patient profile and that can support the physician for computing accurate diagnosis for the individual patients whom likely to suffer with disease. Healthcare data are most valuable data that can be derived from variety of sources such as sensors, clinical notes, images, text from biomedical literatures or might be from traditional electronic records. These different types of data that are collected from different sources seems to be heterogeneous in nature which needs several challenging processes for analyzing the data. Various techniques are needed for analyzing the different forms of data. Due to the heterogeneity nature of the data integration seems more challenging [2]. In many of the cases the insights are obtained from diverse data types that cannot be collected from single source. Hence highly potential integrated data analysis methods are needed which seems to be an interdisciplinary of health care. The healthcare field observes more number of advances that are coming from diverse disciplines such as data mining, databases, information retrieval, healthcare practitioners and medical researchers. This booming interdisciplinary nature adds the richness to the healthcare field, it also adds the challenges for making significant advances in the field. Researchers from the computer Science field will not have exposure to domain specific medical concepts likewise practitioners and researchers of medical fields will have limited exposure to the statistical and mathematical concepts that are required for the data analytics [3, 4]. This is felt to be critic situation for creating the coherent body of work in this filed even though the available analysis techniques can process the available data. This diversity results in forming an independent lines of work that is completely in different perspectives.

2. DATA VARIEABILITY

Simply Electronic Medical Record (EMR) is something that provides valuable Healthcare information by analyzing it. EMR data are available in different size and formats. Small sized EMR data can be easily analyzed or understood. But in recent time, the size and growth of the Electronic Medical Records (EMR) are increased tremendously. This may include different forms of data that are needs to be collected and analyzed for extracting the information. The EMR data can be of different forms such as Structured, Unstructured and Semi-Structured. This section explains about the forms of data the EMR are collected.

2.1 STRUCTURED DATA

Structured EMR data are organized or labeled patient records which can be analyzed easily and effectively. EMR data are well organized data that are formatted in a repository such as database. It includes all the data that are stored in SQL database as rows and columns. Usually it has relational keys with which the fields can be easily mapped. Accessibility and searching information is too easy in such type. They are comparatively too simple for storing, retrieving and analyzing, but are strictly defined in terms of field type and field name. Nowadays, these types are seems to be most processed as it is simpler for information processing, but they represent only 10% of all the informatics data which will not be sufficient to extract the information by analyzing the data.

Structured data use a controlled vocabulary rather than narrative text, setting limits on how the data are recorded in the clinic, resulting in consistency of that structured health record data within a health care provider. Structured EMR data also lend themselves to straightforward digital searchability across the EMR for specific information. The EMR structured data source can be from two different sources such as, (1) Human generated EMR data (2) Machine generated EMR data

Human generated EMR data

These data are generated by humans by making an interaction with Health care devices.

Input data: Input data are fed to the machine by the humans. For example, to understand the patient's information and behavior data like their name, date of birth, sex, age, income, medical and surgery details along with date which are non-free and so on will be collected.

Click stream data: This data can be generated from website whenever the link is clicked. This might be analyzed for acquiring the valuable patient information.

Machine generated data

These data are generated automatically without any interaction by the human.

Sensor data: It includes RFID tags, healthcare devices, patient entering the hospital. These can be

used in the inventory control and supply chain management.

Log data: The behavior of the patient applications, healthcare servers, and networks will recorded every now and then while they operate. These behavioral logs are called as log data. This logs will be in huge forms which can be used to predict the security breaches that occur and other changes in the service level agreements.

Point-of-sale data: The product related information can be generated when the bar code of the product is scanned while it is purchased. Ex. Patients buying tablets, tonics and etc.

2.2 SEMI-STRUCTURED DATA

Semi-structured data are structured data formed in an unorganized way. This type of data that will have the properties related to the organization but will not reside in the relational database and can be processed easily. By doing changes in the process, it can be stored in the relational database. Since it does not have the formal structure as relational database or any other form of data tables, it need tags or other forms of markers for separating the semantic elements, hierarchies and the fields that exist within the data. Eg. The data in web such as JSON files, .csv files, XML files, delimited text files. Since this type of data are in unorganized form it is difficult to store, retrieve, and analyze.

2.3 UNSTRUCTURED DATA

The data that is in unorganized format or that do not have data model is said to be unstructured data. Hence it will not fit to the relational database in predefined manner. So for these type of data, the alternative advanced tools, softwares etc. will be used for storing, accessing and managing the data. This type will be most prevalent in IT systems for variety of business intelligence and for different analytics applications for predicting valuable information. Ex. Word, PDF, images, video, audio, Text, web pages, email and other streaming data. Table. 1, represents the difference between structured, semi-structured and unstructured data.

Examples of unstructured data (generated by Humans):

• Text internal of Enterprise: Documents, logs and e-mails that are maintained for the healthcare organization. The information of the patients seems to generate the largest text information.

• Social media data: EMR data generated from Social media Youtube, LinkedIn, Flickr etc.

• Website content: The data that are collected from any website will be in unstructured formats.

3. COLLECTION OF DATA

Data collection is defined as the process of collecting, analyzing and interpreting different types of information related to a particular disease or healthcare needs. Traditional patient records are collected from sources like personal survey, hand written prescription and hardcopy of the patient's record from local hospital. Prior to the evolvement of digital data, the healthcare records are of physical form. So the data are collected and managed within the hospital itself. But after the recent advancements in the Internet Technology, the patient records are collected in a digital format. Some of the example of digital data used in the field of medicine are digital scan reports, videos shot on laparoscopic cameras, digital X-ray reports, endoscopy videos, ultrasonic records.

FIELD	STRUCTURED DATA	SEMI-STRUCTURED DATA	UNSTRUCTURED DATA
Technology	Relational Database	XML/RDF	Binary Data
Management of Versions	Versioning can be done over rows, tuples, tables.	Versioning is possible through graph or tuples	Versioning as the whole
Management of Transactions	Transaction is matured and uses Concurrency Techniques	Transaction is based on DBMS and are not matured.	No concurrency and transaction management
Scalability	Low	Medium	High
Flexibility of data	Flexibility is low and is schema dependent	Flexibility is higher than structuded data and lower than	More Flexible and no schema dependancy

		unstructured data	
Robustness	Highly robust	Not very spread as it is new technology	—
Performance of Query	Queries are Structured which allows joining of complex data	anonymous nodes can be Queried	Query over text is alone possible

Table.1. Differences between structured, semi-structure and unstructured data

These medical data are the fast growing data in the digital world. As per the survey conducted by DELL EMC., 2018, the healthcare data growth rate has increased 878 percentage since 2016. It also claims that the total amount of healthcare data will reach 20,000 petabytes by 2020. In addition to that, more healthcare applications and databases are developed every day to work with the healthcare data.

Important source of Electronic Medical Records are electronic health (ehealth) devices and communication supported health devices. Data are collected in a frequent interval from the patients though ehealth devices and stored in the cloud storage. If the data are collected from patients through an electronic devices directly, then such data are called as a Patient Generated Healthcare Data (PGHD). The Cloud Service Provider (CSP) maintains the patient's clinical data like demographics, progress notes, problems, medications, on the cloud storage. Patient's medical records are digitalized and assists in ensuring data is accurate. Electronic Medical Records (EMR) data collection can be classified as quantitative and qualitative data collection. In quantitative data collection, the data are collected in the form of numeric variables (i.e.) it collects the information from the patient as numeric values, such as count, number and percentage. Qualitative data collection method collects the patient data in a non-numeric fashion. This type of data is collected through methods of observation, one to one interview, and online survey. Qualitative data are also known as categorical data. The important ways of collecting the Electronic Medical Records (EMR) are ehealth device, semantic data collection and patient chatbots.

3.1. EHEALTH DEVICES

Ehealth devices are also called as self-monitoring healthcare devices. It uses the sensors and wireless communication design to measure the patient's health and transfer it to the cloud storage. It allows the patient as well as the physicians to measure and monitor the patient's health remotely. Some of the available healthcare monitoring ehealth devices in the market are temperature device, heart beat tracking device, glucometer, oximeters, pulsometer and Blood pressure devices. These IoT based healthcare devices are considered as an important advancement in the field of healthcare management. As the use of cloud computing and wireless technology, the demand for ehealth devices are increasing drastically. It is predicted that in 2020, the ehealth devices will account 80% of wireless devices. The main advantage of these devices are mobility and accessibility of smart phone and tablets.

3.2. DATA EXTRACTION FROM HEALTH CARE WEBSITES

Semantic extraction of healthcare information: Extract information relate to a particular disease, medical facts**,** attributes from a website or unstructured data. The purpose of semantic data extraction in healthcare enables to analysis the unstructured content, electronic prescriptions, medical text documents, emails, digital images patient report. The main objective of semantic analysis is to structure the unstructured data.

Semantic data extraction on websites has two major approach such as, rule matching data collection and machine learning data collection

Rule matching data collection: It collects the information related to a particular word or phases from the websites. Rule based matching algorithm is used on raw medical website to gather the information about a particular disease. They also provide access to the tokens within the document and their relationships.

Machine learning based data collection: It is a statistical analysis of the content, the potential compute-intensive application that can benefit from using Hadoop. This approach derives the relationship from statistical co-occurrence within the website.

3.3. PATIENT CHATBOTS

To deliver quality services to the patients, Medical informatics entities are using recent technologies like artificial intelligence and prediction technologies in the healthcare application. It is impossible for a patient to get advice from physicians at emergency situation. To provide a round the clock medical advice to the patients, healthcare industries are investing a lot to create an automated medical chat bots. Medical chatbots are conversational software available for smart phone applications. It provide a faster service to the patients. They are adequate enough to communicate and gather information from the patients. The collected information are fed to the deep learning algorithms to improve the intelligence of the chatbots. These medical chatbots are the recent trends in healthcare industry. Some of the most popular medical chatbots in the healthcare industry.

4. PREDICTIVE MODEL FOR ANALYSIS

The predictive models that exists in the data analytics provides a valuable score for measurable medical data elements. It also can predict and provide the probability of diseases that can affect the patient in the near future. It may also provide the likelihood of a patient defaulting on a disease based on his or her personal history or characteristics. Statistical models are also available for predicting the type disease and prevention of diseases [5].

The context of applying predictive analytics is quite diverse. The expected outcomes may vary from binary values such as yes/ no or true /false for fake prediction to predict the real numerical values of medical field. Here different classes of predictive techniques have been discussed which will support the reader to understand the various models that are prevalent.

Predictive analytics can be grouped into two major groups such as Regression and Machine Learning techniques.

A. Regression Techniques

Regression analysis is one of the predictive modeling techniques which predict the dependency between the target variables. It is used to predict sales trends, possibility of churns or fraudulent transaction. It focuses on forming the mathematical equation for capturing the interactions between the different target thereby reducing the overall error in the predicted model

Linear regression is one form of regression models used to predict the response variable in linear manner. The parameters can be adjusted or learnt so that the addition of squared residuals is minimized.

Logistic regression assigns probabilities for the possible outcomes. A binary outcome variable can be converted to an unbounded continuous variable from which a regular multivariate model is estimated.

Time series models are used for predicting the future behaviour of the variables when the internal structures such as trends, auto correlation etc is available. They are capable of decomposing the components such as seasonal and trends through which better models can be produced. Few time series models are Moving Average Model, Auto Regressive Model, a combination of the two models is called as Auto Regressive Moving Average and Auto Regressive Integrated Moving Average.

Decision trees is a model which is the collection of defined rules based on variables in the data set, where the rules are defined such that to obtain the best split for differentiating the observations that belongs to different target classes. Rules are explanatory and are preferred by the data analysts.

B. Machine Learning Techniques

Machine learning based models are another form of predictive analytics that are used for applications such as diagnosing medical conditions, fraud detection etc. However, unlike classification or regression trees, this model remains a black-box without considering the relationship between the predictor variables and it sufficiently predicts the dependent variable [6].

Among the various existing machine learning models, neural networks is the model which is inspired from the nervous system of human have gushed out in popularity in recent days as it is capable of learning complex relationships among the predictor variables. For doing classification there exists wide variety of neural network models[7]. The earlier neural network models use only three layers such as the input, hidden and an output layer and the deep neural model gained popularity by using more than one hidden layers. Large number of neurons and the interconnections

between them are capable of modeling non-linear relationships between input and output variables.

Some of the commonly used neural networks that are suitable for prediction tasks are:

1) Multilayer Perceptron
This neural network uses more than one hidden layer of neurons. It is also known as deep feed forward neural networks.

2) Convolutional Neural Networks
This type of neural networks performs convolutions between the input data and desired filter. They are more efficient in learning hierarchal features from the data by extracting the relationships between the neighbours.

3) Recurrent Neural Network
These types of neural network have hidden layer neurons which will have self-connections, for making the neuron to possess memory. These types of networks are suitable for text processing as the interpretation in the text will be dependent on neighboring words or contexts. Thus, these types of neural networks models the interrelationships of words by considering their sequence.

4) Long-Short Term Memory Network
This type of networks are extensions of recurrent neural networks in which each hidden layer neuron will be incorporated with memory cell. They are good in finding the long distance relationships. These types of networks can be applied for analyzing any kind of sequential data.

5. ALGORITHMS

5.1. NATURAL LANGUAGE PROCESSING

In healthcare industry, the clinical information will be in the form of written text which will be in huge forms, such as laboratory reports, physical examination reports, operation notes of patients, discharge related summary etc these are usually in unstructured form and are not comprehensible for the computer based programs as it needs special models for processing the text. Natural Language Processing model provides solution to these issues by identifying the series of keywords that are relevant to disease in the patient notes based on the existing databases thereby enriching the structured data for supporting the clinical decision making[8].

5.2. NAÏVE BAYES APPROACH

Naïve Bayes classifier is a probabilistic method used for categorizing the text, solving the problem of document predictions for finding the category to which it belongs to Naive Bayes classifier considers that one particular feature of the class will be unrelated to other features. Even though the features of a class are independent, all its properties will independently contributes its probability for a certain category. It is one of the efficient probabilistic classification algorithm that are successfully applied for many of the medical related problems.

5.3 DEEP LEARNING

Deep Learning belongs to the machine learning family and are based on the artificial neural network techniques, as it is a neural network with more number of layers. When compared to traditional machine learning algorithms, the more complex non-linear patterns can be learned using the deep learning algorithms in the data. Modules are pipelined and are trainable, it is a scalable approach and the automatic feature extraction from data can be performed. [9, 10]
In the healthcare applications, these type of algorithms handles both the tasks such as Machine Learning and Language Processing. The predominantly used Deep Learning algorithms are convolution neural network, deep belief network, multilayer perception model and recurrent neural network. It remains one of the most effective classification algorithm and are successfully used in many of the health care related problems, such as health care report classification and journal classification etc.,

5.4 CONVOLUTIONAL NEURAL NETWORK

It is developed for handling high dimensional data or the data with more number of traits. As proposed by LeCun, the pixel values that are rectified with normalization of the images will be the inputs. Convolutional networks were inspired by medical processes such that the connectivity pattern that exists between the neurons with separate cortical neurons which responds to the stimuli in the region which is restricted. However, the whole visual field is covered as the receptive fields of various neurons will gets overlapped. The CNN then transfers the weighted pixel values of the image in the convolution layers and sampling is

done in the subsampling layers. The final output will be a recursive function of the input values [11].

5.5 PHENOTYPING ALGORITHMS

Phenotyping algorithms are implemented using the samples of the diseases on the EHR data that are usually collected from the health care units for diagnosing the diseases. The data may be in unstructured form which contains large amount of texts from the physician reports, various diagnostics of diseases, and different vital signs. A Phenotyping algorithm is a different form of special model that are carried through various number of medical data points with specific codes for radiology results, billing and natural language processing where different form of texts are extracted from the physicians. Machine learning algorithms with supported vector machine can be applied for identifying the arthritis in combination of patient's prescription records for improving the accuracy of predictive models of disease. Example, the prevalent condition of the diabetic patients can be suggested by examining the usage of hypoglycemic agents that are collected from the prescription records.

6. CONCLUSION

From the above chapter it is inferred that there is consequential need for the improvement of structured, semi-structured and unstructured health care data for storing, analyzing and interpreting. Though powerful tools are already exist for analysis that might help the analyzers to analyze the data well there is a lack of standardization which continues to impede the complete process. Machine learning, language processing and artificial intelligence have the prospectus for streamlining the way that the unstructured data can be utilized, but we fails to capture the point that the machines are making the critical decision instead of physicians who were the decision makers traditionally. Regardless, of all these the patients should also expect and look forward for improved medical or health outcomes as the technological upheaval improves the way the health data are looked. Thus this chapters elaborates about the different forms of healthcare data with few algorithms and use cases thereby supporting the users to understand the basic concepts of healthcare data analysis.

REFERENCES

1. Jake Luo, Min Wu, Deepika Gopukumar and Yiqing Zhao, "Big Data Application in Biomedical Research and Health Care: A Literature Review", Biomedical Informatics Insights, Vol. 8, pp. 1-10, 2016.
2. Rosenbloom S.T, Denny J.C, Xu H, Lorenzi N, Stead W.W, Johnson K.B "Data from clinical notes: A perspective on the tension between structure and flexible documentation" Journal of the American Medical Informatics Association, Vol. 18, No. 2, pp. 181–186, 2011.
3. Annemarie Jutel, "Classification, Disease and Diagnosis", Perspectives in Biology and Medicine, Vol. 54, No. 2, pp. 189-205, 2011.
4. Shaik Razia, P. Swathi Prathyusha, N. Vamsi Krishna, N. Sathya Sumana, "Review on disease diagnosis using machine learning techniques", International Journal of Pure And Applied Mathematics, vol. 117, no. 1, pp. 79- 85, 2017.
5. M. Chen, Y. Hao, K. Hwang, L. Wang and L. Wang, "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities," in IEEE Access, vol. 5, pp. 8869-8879, 2017.
6. Harleen Kaur, Siri Krishan Wasan, "Empirical Study on Applications of Data Mining Techniques in Healthcare", Journal of Computer Science, Vol. 2, No. 2, pp. 194-200, 2006.
7. Allen Daniel Sunny, Sajal Kulshreshtha, Satyam Singh, Srinabh, Mohan Ba, Sarojadevi H, "Disease Diagnosis System By Exploring Machine Learning Algorithms", International Journal of Innovations in Engineering and Technology, Volume 10 Issue 2, pp. 14- 21, 2018.
8. Muhammad Asif, Hugo F. M. C. M. Martiniano, Astrid M. Vicente, Francisco M. Couto**,** **"**Identifying disease genes using machine learning and gene functional similarities, Assessed through Gene Ontology", PLOS ONE, Vol. 13, pp. 1- 15, 2018.
9. Keerrthega M.C, D. Thenmozhi, "Identifying Disease -Treatment Relations using Machine Learning Approach", Procedia Computer Science, Vol. 87, pp. 306 – 315, 2016.
10. Min Chen, Yixue Hao, Kai Hwang, Lu Wang, And Lin Wang, "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities", IEEE Access, Vol. 5, pp. 889-8879, 2017.
11. Honghan Wu, Giulia Toti, Katherine I Morley, Zina M Ibrahim, Amos Folarin, Richard Jackson,

Ismail Kartoglu, Asha Agrawal, Clive Stringer, Darren Gale, Genevieve Gorrell, Angus Roberts, Matthew Broadbent, Robert Stewart and Richard JB Dobson, "SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research", Journal of the American Medical Informatics Association, Vol. 25, No. 5, pp. 530–537, 2018.