A Review of Data Mining using Bigdata in Health Informatics

D.B.K.Kamesh [1], V. Neelima[2], R. Ramya Priya[2]

[1]Associate Professor, Department of Electronics and Computers, KL University, A.P., INDIA.
[2]Student, Department of Electronics and Computers, KLUniveristy,AP,INDIA.

Abstract- The amount of data produced in health informatics growing large and as a result analysis of this huge amount of data requires a great knowledge which is to be gained. The basic aim of health informatics is to take in real world medical data from all levels of human existence to help improve our understanding of medicine and medical practices. In our paper, we present research using Big data tools and approaches from various sources. Apart from the data gathered in different levels, there are multiple levels of questions addressed.

Index Terms- Big Data, Health informatics, Bio informatics, Neuro informatics, Clinical informatics, Public health informatics, Social media

I. INTRODUCTION

The technology has changed such that the field of health informatics has started to handle the knowledge of big data. By using data mining and big data analytics, diagnosing, treating, helping and healing all patients in health care as become easy. As a result, there is an improvement in healthcare output(HCO). The health care output is the quality of care that health care can provide to end users.

Health informatics is a combination of information science and computer science which deals with various fields like Bio informatics, image informatics, clinical informatics, public health informatics, Translational Bio Informatics (TBI).

Various studies done on Health Informatics uses data from a particular level of human existence. Bio-informatics uses molecular level data, neuro informatics uses tissue level data, clinical informatics uses patient level data, and public level informatics uses population data
TBI uses any of the data from molecular level to entire population. TBI is used to mainly answer clinical level questions. These researches in various subfields are used to improve health care system.

The various subsections in this study are: “Big Data in Health Informatics” which gives an overview on Big Data in Health Informatics, “Levels of Health Informatics data” which discusses the various subfields in Health Informatics, “Using micro level data-Molecules” which describes using data from micro level, “Using Tissue Level data”, which uses data from Tissue Level, “Using Patient level data” which describes about patient data, “Using Population level data- social media”, which describes about population level data, “Translational Bioinformatics”, “Analysis and future work”, which describes about the future work which can be done on Health Informatics using Big Data, “Conclusion”, which gives a brief overview of the study.

II. BIG DATA IN HEALTH INFORMATICS

Big data refers to the tools, processes and procedures which allow an organization to create manipulate and manage very large data sets and storage facilities. Big data enables an opportunity for aggregation and integration leading to cost effective and patient care. Demchenko et al.[1] defines big data by five vs: Volume, Velocity, Variety, Veracity and Value. Volume means the large amounts of data used. Velocity means the speed at which new data is generated. Variety means the level of the complexity of data. Veracity is used to measure the genuineness of the data. The value gives how good the quality of data is.

2.1. Sources and techniques for big data in Health Informatics:

2.1.1. Electronic Health Records(EHR)data:

A. Data:

We have different types of data in Health Informatics like Genomic data which describes the DNA sequences, Clinical data which describes structured EHR, unstructured EHR and medical images, Behavior data which describes social network data, mobility sensor data.

B. Billing data-ICD codes:

ICD stands for International Classification of Diseases. ICD is a hierarchical terminology of diseases, signs, symptoms, and procedure codes maintained by the World Health Organization (WHO). Pros: Universally available. Cons: medium recall and medium precision for characterizing Patients.

C. Billing data – CPT codes:

CPT stands for Current Procedural Terminology created by the American Medical Association. CPT is used for billing purposes for clinical services. Pros: High precision and Cons: Low recall.

D. Lab results:

The standard code for lab is Logical Observation Identifiers Names and Codes (LOINC®). The Challenges for lab results are:

- □ Many lab systems still use local dictionaries to encode labs

- □ Diverse numeric scales on different labs

Often need to map to normal, low or high ranges in order to be useful for analytics – Missing data, not all patients have al labs. The order of a lab test can be predictive, for example, BNP indicates high likelihood of heart failure, Time L.

E. Medication:

Standard code is National Drug Code (NDC) by Food and Drug Administration (FDA), which gives a unique identifier for each drug. Not used universally by EHR systems – To specific, drugs with the same ingredients but different brands have different NDC. RxNorm: a normalized naming system for generic and branded drugs by National Library of Medicine.

Medication data can vary in EHR systems – can be in both structured and unstructured forms. Availability and completeness of medication data vary – Inpatient medication data are complete, but outpatient medication data are not – Medication usually only store prescriptions but we are not sure whether patients actually filed those prescriptions.

F. Clinical notes:

Clinical notes contain rich and diverse source of information.Challenges for handling clinical notes – Ungrammatical, short phrases – Abbreviations, Misspellings Semi-structured information. Copy-paste from other structure source – Lab results, vital signs. Structured template: – SOAP notes: Subjective, Objective, Assessment, Plan.

2.1.2. Analytic Platform:

The various features of analytical platform are Information extraction which describes Structured EHR, Feature extraction and Unstructured EHR ,Feature selection which describes context, patient representation, feature selection, and predictive modeling which describes the classification, regression and patient similarity.

2.1.3. Clinical text mining:

In clinical text mining, description of text mining which is information extraction and clinical text vs biomedical text. The biomedical text means medical literatures. Clinical text is written by clinicians in the clinical settings.

If we consider the data gathered for Health Informatics, big volume refers to the large amounts of data that is records stored for patients. Big velocity is when new data comes in at high speeds. Big variety means the data sets where large amounts of different types of independent attributes of data gathered from different sources and so on... Veracity of data in Health Informatics comes when working with possibly noisy, incomplete data. Such type of data needs to be properly evaluated. Value of data is the main aim because to improve the HCO is the basic aim. Not all studies covered in this field of Health Informatics fit all five of the qualities in [1], all's definition of big data.Data that has value without veracity may need some methods to expand the size of dataset.

In US, using data mining in Health Informatics, can save health care industry upto$450 billion each year. As of 2011, HCO generated 150 exabytes of data (1 exabyte=1000 peta bytes). This information can be stored in EHRs which can store 44+ petabytes of patient data. Only in bio informatics generate many terabytes of data. With these data coming from different places and in many forms, Health informatics (HI) has to find ways dealing with all this data. We can integrate and combine different sources of data even across different sub fields.

III. LEVELS OF HEALTH INFORMATICS DATA

There are various subfields of Health informatics like bio informatics, Neuro informatics, clinical informatics and public health informatics. There is a confusion which sub field a study should fall under because of the line between each subfiled of health informatics which are blurred. So highest level of data is used for deciding subfield membership.

In the research of Health Informatics, there are two levels which are to be considered. One is the level from which the data is collected and the other is the level at which the research question is asked. The question level in a given work may be different from the study corresponding to the data levels. The tissue level data corresponds to human-scale biology question, the patient level data corresponds to clinical questions, population level data corresponds to epidemic scale questions.

3.1. **Bioinformatics**

Bioinformatics focuses on analytical research in order to know how the human body works using molecular level data in addition to developing methods of effectively handling data.

Data in Bioinformatics is certainly Big Volume as the data is continually growing because the technology being able to generate more molecular data per individual. Mc Donald et al.[2] created a software tool Khmer. This tool is used to solve hardware computational problems through software. The tools in this software preprocess Big Volume genomic sequence data by breaking up long sequences into relatively short strings which ca be stored in a bloom filter-based hash table, which helps both the ability and efficiency of Bioinformatics data.

3.2. **Neuro-*informatics:***

Each data instance such as MRI's is quite large and leading to data with Big Volume.The research done on Neuro-informatics concentrates on analysis of brain image data (tissue level). In order to learn how the brain works and find the relation between the information gathered from brain images to medical events, etc., all with a goal of extending the medical knowledge at various levels.

3.3. Clinical Informatics:

The research done on clinical informatics involves making predictions that can help physicians make better, faster, more accurate decisions about their patients through analysis of patient data. Clinical informatics lead to Big Value as their research directly uses patient data. Further efforts can be made to make it more accurate, reliable, efficient. According to Bennett et al.[3] there is about a fifteen year gap between clinical research and the actual clinical care that is in practice. These days decisions are made mostly on general information that has worked before, or based on what work has been done in the past.

3.4. Public Health Informatics-Social media:

The Public Health Informatics applies data mining and analytics to population data inorder to gain medical insight. The data is generally gathered from either traditional means i.e., hospitals or gathered from the population i.e., social media. As a default, the population data has Big Volume, Big Velocity and Big Variety. But the data that is gathered from the population though social media possibly has low Veracity leading to low Value, but the techniques for taking the useful information from social media such as Twitter posts can also have Big Value.

IV. LEVELS OF BIG DATA

4.1. Micro level data-molecules

The data gathered from molecular level often experiences the problem of high dimensionality- where the data has large number of independent attributes. This is because the molecule level of data contains thousands of possible molecules which are represented in datasets as features. Some of the applications of molecular level data are Chemo-Informatics, DNA sequence analysis, high-throughput-screening. By using this molecular level data clinical questions are answered.

4.1.1. Using gene expression data for prediction of clinical outcome

The studies in molecular level data uses gene expression data to answer clinical questions. As per now two researches were done both of which focus on cancer. The first study uses gene expression to categorize leukemia into two different subclasses. The second study uses gene expression to predict relapse among patients in the early stages of cancer. Haferlach et al.[4] formulated to place the patients into 18 different subclasses of either lymphoid leukemia or myeloid. He used 3,334 patients among whom two-thirds were used for training i.e., 2,143 patients and the rest one-third i.e., 1,191 patients for testing. From each patient 54,630 gene probe set samples were taken.

People used an all-pairwise classification designing using the mean of difference between perfect match and mismatch intensities with quantile normalization.

The second part of their study was to get the results for testing pool of patients. [4] achieved a median specificity of 99.8% and a median sensitivity of 95.6% which was used for the classification of 14 subclasses of acute leukemia in which 6 were lymphoid and 8 were myeloid. some authors claim that by using discrepant instances, we can achieve better results. Salazar et al.[5] used a gene expression seeking to predict within a five year period whether or not a patient will relapse back into having cancer. They gathered a total of 394 patients where 188 were for training and 206 were for validation. The training pool was accumulated over a 19 year period from three different institutions from various countries. The validation pool was assembled over 8 years from another institution from another country. The approach used is good strategy and should be employed in future research. The results of these research efforts lead to the idea that microlevel data could give similar results if these procedures were applied to other types of cancers inorder to help doctors diagnose and treat their patients.

4.2. Tissue level data:

The studies in this section use tissue level data to answer human scale biology questions creating a full connectivity map of the brain and predicting clinical outcomes.this level of data incorporates imaging data and gives a number of big data challenges like feature extraction and managing complex images. Those studies which combine imaging data with various other data sources also exemplify the Variety aspect of Big data.

4.2.1. Creating a connectivity map of the brain using brain images:

There is a project name Human Connectome Project (HCP) led by WC Minn HCP Consortuim. The main aim of this project is to map the human brain by making a comprehensive connectivity diagram and to advance the current knowledge of how the brain functions. There are two phases in this project. The first phase is from 2010-2012 in which methods for data acquisition and analysis were improved. The second phase is from 2012-2015 in which methods were applied to 1200 healthy adults who are aged between 22-35.the subjects in this project are varied from twins to non-twins to check the variation. The expected result of this project is that new and extremely important information can be gleaned from mining the HCP data coming out from HCP research.

By this connectivity map, we can know why some people have brain disorders, giving physicians a possibility for easier diagnosis, early detectionof future illness. Till now, 148 participants outcomes are released. Once all the data for the 1200 patients have been generated, there could be similar data created on patients with various ailments and various ages to find the difference between such brains through data mining and analysis.

Annese[6] says that depending on neuroimaging is not correct . there should also include histological methods of studying actual brain tissue and also says that MRI measurements matching up with that of anatomical measurements is a considerable issue interfering with making a comprehensive connectivity model of human brain. MRI’s are large and have high resolution, a higher resolution than histological methods which give results that are neither the same size nor quality and also comments to validate MRI’s with study of actual brain tissue. MRI(Magnetic Resonance Imaging) is a medical imaging technique used in radiology to investigate the anatomy and physiology of the body in body health and disease. MRI scanners use strong magnetic fields and radio waves to form images of body. Histology is. performed by examining cells and tissues by sectioning and staining followed by exam under light microscope.

According to Van Essen et al.[7] , the project looks very promising because new and extremely important information can be gleaned from mining the HCP data coming out from the HCP consortium’s research.

Creating a full connectivity map of the brain could lead to information that could help in determining the reasons why people have certain brain disorders are a level previously unattainable, giving physician a possibility for easier diagnosis, early detection of future illness or maybe even prevention of mental or physical ailments. Once the data for all 1200 patients is generated, there could be similar data created on patients with various ailments and various ages to find difference between such brains through data mining and analysis.

4.2.2. Using MRI data for clinical prediction:

The main aim of this section is to answer clinical level questions which covers two studies. The first study uses both MRI data and a list of clinical features with the aim of finding the correlations between physical ailments to that of many locations of the brain. The second study takes MRI data and determine the amount a patient has Alzheirmer's disease. Yoshida et al. proposed a model combining patients clinical features and MRI image intensities which consists of voxels. A voxel is an element of volume representing a point on a grid in three dimensional space. This method is based on the algorithm of radial basis function sparse partial least squares (RBF-sPLS) giving their method an advantage over similar methods granting the ability to select not only clinical characteristics but also determine effective brain regions. This is to say that by creating sparse, linear combinations of variables which are explanatory , developed approach concurrently performs feature selection and dimensionality reduction. Both of the tasks are problematic for vast amounts of data, but RBF-sPLS manages efficiently. Estella et al.[8] introduced a method with the goal of predicting to what degree a patient has Alzheimer's disease with three levels of classification which are completely healthy,Mild Cognitive Impairment(MCI) and already has Alzheimer's. they gathered almost 240GB of brain image data for 1200 patients stored by Alzheimer's Disease Neuroimaging Initiative(ADNI).

The various steps include spatial normalization, extraction of features, feature selection and patient classification. After the extraction of features, they found two subcategories like morphological and mathematical where 332 and 108 were gathered rsespectively. For feature selection they used a method based on Mutual Information(MI) along with some influence from minimal redundancy maximal relevance criterion. MI assisted in determining the dependence between two given variables.

These studies using MRI data show that they can be useful in answering clinical questions as well as making clinical predictions. More research may be needed which was determined by Yoshida et al. to have correlation with kidney disease, anemia and aging, can be determined as clinically significant. In estella et al. various features extracted from MRIs were shown to have the ability to classify patients into varying degrees of dementia.

4.3. Patient **level** *data:*

This study covers patient level data to answer clinical level questions including prediction of ICU readmission, prediction of patient mortality rate and making clinical predictions using data streams. As per molecular level data, feature selection can help choose the important features, which also helps in making quicker clinical decisions.

4.3.1. **Prediction of ICU readmission and mortality rate**

The main aim of this research is to predict ICU readmission, mortality rate after Icu discharge as well as predicting 5 year life expectancy rate. The study done by Campbell et al. [9] focused on ICU patients that were discharged and expected to both live and not return too early afterwards. There are three research directives that were considered. One of them is death after ICU but before hospital discharge. The second one is readmission to ICU within 48 hours of ICU discharge but before discharge from hospital. The third one is readmission to ICU to any point after ICU discharge but before hospital discharge. An attribute of patient called Apache 2 was considered for prediction of ICU readmission and mortality rate for ICU patients. APACHE2-Acute Physiology and Chronic Health Evaluation 2 score uses physiological variables which are most useful for predicting another non physiological variable and it is increasing age.apart from APACHE 2 there are other scores like SAPS 2 and TISS. In APACHE 2 there are 12 variables considered for prediction. [9] noted that each of the patient potentially fell in more than one of the categories mentioned earlier. But there are three binary models built. The importance these prediction models could have would be help physicians determine which of these patients fall into these three froups. The feature selection method decided was simple logistic regression for all three models to determine which of the 16 attributes had a strong correlation to each prediction. Multiple logistic regression has been chosen for building the prediction models. These methods were tested using the Hosmer and Lemeshow goodness of fit test. According to Bradley [10], the Area Under the ROC curve (AUC) is the best criteria for measuring the classification performance of logistic regression. For determining the quality of the three models, the AUC results for each model are compared to results from APACHE2 score for each prediction. The first model that is predicting death after ICU discharge had an AUC value of 0.74 compared to AUC from APACHE 2 of 0.69. the second model that is for predicting readmission obtained an AUC of 0.67 while APACHE 2 received an AUC of 0.63. the third model of predicting readmission within 48 hours of ICU discharge required an AUC value 0f 0.62 while APACHE2 earned an AUC of 0.59. The three models with the chosen set of features only achieved minimal improvement over APACHE2 for the prediction of ICU readmission and mortality rate for ICU patients. [9] noted that this improvement may be because APACHE2 uses physiological variables. These variables are most useful for predicting ICU readmission and mortality after ICU discharge.all about 23% of patients fall into these three models and should not have been released from ICU. If they were allowed to stay more, they might have been saved.

Ouanes et al. [11] conducted a research aiming prediction whether a patient would die or return to ICU within the first week after ICU discharge. This research was performed on around 3462 patients who were admitted to ICU for minimum 24 hours. AIC(Akaike Information Criterion) is a method which verifies the quality of statistical methods. The model created was subjected to testing inorder to end up with their final set of 6 variables from the original variables of 41. The six variables which were chosen were age, SAPS 2, the need for central venous catheter, SOFA score, discharge at night and SIRS score during ICU stay. These variables were used to make final prediction model using logistic regression. This was used to develop their minimizing ICU readmission(MIR) score. This MIR score is a measurement for determining whether a patient should be discharged from an ICU or not. Through this MIR score, Ouanes et al. was able to achieve good results with good calibration decided by the HL.gof test and an AUC of 0.74 at a 95% confidence interval. The MIR score would have been better if more than one selection method was used.

The studies discussed in this section has the potential to improve clinical discharge procedure, and determines which patients should be released from the ICU and which patients should receive particular treatment. If we look at the research results of Campbell et al. and Ouanes et al. they covered prediction of ICU readmission and death rate after discharge. The top variables are age, APACHE scores, various physiological variables and a few others. According to Ouanes et al. between a day and seven after discharge from readmission rate and death rate go down means that keeping a patient a little longer could be beneficial. The target of this research is on patients that have preventable death as not all death will be preventable.this research is more important for patients with increasing age as the older patient is less likely a harsh treatment would be beneficial.

4.4. Population level data-Social media

This section uses population level data to answer both clinical level questions and epidemic scale questions. Generally health informatics data is gathered from doctors, clinics, hospitals but recently from even internet. Internet data could be from twitter, google, or anywhere else. This form of Big Data brings additional challenges such as text mining and handling noise gind many new breakthrough in the field of medicine. The first study of research is to determine whether message board data can be useful to help patients find information on a given ailment. The second study is testing if using search query data can effectively track an epidemic across a given population. This level of data has high Volume,Velocity, Variety but low Value and Veracity.

4.4.1. Using message board data:

This message board data provides reliable information. This research is used in determining if message board could be useful source of data for helping people find beneficial health information.

Ashish et al. created a platform called Smart Health Informatics Program(SHIP) with the goal of helping patients connect to other patients through internet by means of four websites inspire.com, medhelp.com, and 2 others. They used a pool of 50,000 discussions. The SHIP is a pipeline considering message boards from all four websites. The first step of SHIP is Elementary extraction that will execute some basic text processing for each entry. A unique ID is given to each discussion and post. The next step is Entity extraction to determine which entry has medical significance. Ashish et al. decided to use XAR system by incorporating ontologies from UMLS. After the entries have gone through their process the facts and expressions in each entry are stored in database. For retrieval he extended and open source java based text search engine library. They tested their system on a test case of a patient experiencing severe cough starting Tarceva. Ashish discovered a previously unreported but third most common side effect of the lung cancer chemotherapy drug Tarceva also known as Erlotinib which means onset of severe cough after starting the treatment. But severe cough is also a primary symptom of lung cancer. With a side effect much similar to symptom, doctors complaint that cough as cancer itself. But a team named Abzooba distinguished between drug induced cough and disease induced cough. Big data harvesting helped them separate side effect from symptom. SHIP analyses information about treatments, procedures, side effects, hospitals and physicians from five broad categories personal experience,advice, information, support and outcome. Each user post receives a yes or no(Y,N) designation, a binary way to express whether or not the post contains relevant information.

Rolia et al.[12] proposed a new system to use social health forums in which there are three steps. The first one is determining patients current medical condition from the personal health record. The second step is that system will acertain which other users have a similar condition. The final step is that a metric will be implemented evaluating and ranking forum topics.

4.4.2. Using search query to track epidemics:

The research is done using search query data from two search engines: Google and Baidu for predicting whether such data can be useful to predict the occurrence and movement of Epidemics.

Ginsburg et al. developed a method which can analyse a Big Volume of search queries from google with a goal of tracking Influenza-like-illness(ILI). By taking a historical log of a period of 5 years, they conducted a research using data from Center of Disease Control and Prevention (CDC) and using 50 million of the popular searches. The queries were taken into consideration without any modifications and validation for this data was done in 2008. The CDC splits the US into 9 regions where the study was used to make predictions using regions of separation. This model looks to find the probability that a patient visiting a physician is related to an ILI for a particular region using variable: probability that a given search query is related to an ILI within the same region. A linear model is fit using both the log odds for ILI physician visits and ILI related search queries giving: logit(I(t))=alpha*logit(Q(t))+E, where I(t) is % of ILI physician visits, alpha is a coefficient, Q(t) is the fraction of queries related to ILI at time t, E stands for the error in the formula . Q(t) is determined by an automated technique without any knowledge of influenza. The authors tested each of the 50 million stored queries alone as Q(t) to see which queries fit with the CDC ILI visit % for each region. The top 45 search queries which are sorted by Z-transformed correlation in nine regions, were chosen to belong to Q(t). this examination of top queries showed connection to influenza sysmtoms. Ginsburg et al. was able to obtain good fit when compared to that of reported CDC ILI % scoring a mean correlation of 0.90 in all nine regions. Validation was done on data gathered on 42 points per region and mean correlation of 0.97 was achieved compared to reported DCD ILI. The authors proved thatsearch query can be used to determine an ILI epidemic in a more real time manner. The results may be improved if other techniques were used to obtain an optimal set other than chosen 45. This study have shown that search query can be an useful tool for quickly and accurately detecting the occurrence of an ILI epidemic which could even be extended to tracking an epidemic.

4.4.3. Using twitter post data to track epidemics:

The research result is similar to the previous subsection of attempting to detect and track ILI epidemics, but instead of using search query data the researchers used twitter post data. Twitter posts come with context. This is advantage over search engine. There are many posts related to health care. This Big Volume of people when considered, there is a high probability that there can be useful ILI epidemic information being posted. Apart from this, there may be noisy sensors and through data mining techniques and analysis, useful information can be found.

Signorini et al.[13] has done a research to employ twitter post data across US by searching through particular areas and analyzing data inorder to predicate weekly ILI levels both across and within thes regions. Their focus is on time period when H1N1 epidemic was happening in the US since they gathered a large amount of tweets from October 1,2009-May 20,2010 using Twitter's streaming application programmer's interface(API). The posts were sifted looking for posts containing a preset of key words correlated to H1N1.the tweets if contained any of the following attributes were not considered for analysis: if located outside the US, containing less than five words, from a user with a time zone outside US, not in English, those submitted through API and not containing ASCII characters. The leftover tweets were used to create a dictionary of English words, from which items like hashtags, @user and links are not used. Using this dictionary signori et al. gathered daily and weekly statistics for each word both in dictionary within each of CDC's 10 regions and throughout the US.

The authors generally use weekly statistics to estimate weekly ILI epidemic status by a general class of SVM(Support Vector Regession). This is a type of classifier which attempts to find a minimal-margin separator, which is a hyperplane in space of instances such that one class is on one side of the hyperplane and the other class is on the other side. A kernel function is used to transform the data into a higher dimensional space. [13] used polynomial kernel function. This model disregards any data points that are already within a threshold E of the model prediction and further builds a nonlinear model to minimize the preselected linear error cost function. Each point is a tweet and features each represent dictionary terms which occur more than 10 times per week. This value of each feature is fraction of total tweets within the given week which contain corresponding dictionary word.

To determine if twitter data can indeed detect ILI epidemics by accurate estimation of CDC ILI values which was done on a weekly basis ona national level and a regional level. They trained their method using 1 million of the tweets for national estimation from October 1, 2009-May 20, 2010 throughout US. To determine the accuracy of the model, Leave-one-out-cross-validation was used. Some authors argue that perhaps smaller amount of tweets containing geolocation information could have generated the slightly higher error rating of 0.37% with a standard deviation of 0.26%.

This study could have incorporated more words to include in their tweet searches rather than just 4 they used as well as use methods to determine the most affective set.[13] used the tweets to follow the public concern for ILI epidemics throughout daily and monthly trends of tweets.

*4.5.***Translational Bioinformatics***:*

Some authors say that translational bioinformatics(TBI) is tha way of the future for health informatics. It is subfield that deals with High Volumes of biomedical data and genomic data, in which current research areas include developing new techniques for integrating biological data and clinical data as well as improving clinical methodology by including findings from biological research. According to Chen et al. TBI has the same levels of Health Informatics: Micro Level, Tissue Level, Patient Level and Population Level, and the main goal of TBI Is to answer various questions at clinical level. There is a confusion which is dividing line between what is included as clinical information and overall health informatics.

Butte et al. [14] discussed that several TBI studies featured in JAMIA which conmibe biological data with medical records to achieve medical gains as more data angles are tested. Authors comment that TBI started from a research done by a small group who found how to bridge the gap between computational biology and medicine.

Sarkar et al. discusses that there are three areas of primary research of TBI: determining the molecular level(genotype) impacts on evolution of disease, understanding overall consistency between molecular, phenotype and environmental correlations across different population, learning the impact of therapeutic procedures as can be measured by molecular biomarkers. They believe that TBI is a primary position to possibly determine many of the mysteries of complex diseases or any of the other research with the explosion of both molecular level data and biomedical data.

V. FUTURE WORK

5.1. Molecular level data:

The main challenge is handling Big Volume of data. The future works include developing and testing big volumes of data and make the predictions in a way that is fast, accurate, efficient.

5.2. Tissue level data:

The actual data mining analysis of the connectivity map remains entirely future scope. By extending the work done by [6], a possible discovery of previously unattainable knowledge about brain and how it connects to the health of the human body. This work is for :Creating a connectivity map of the brain using brain images.

The subsection "Using MRI Data for clinical prediction", difficulty is handling Big Volume of MRI data. The challenges include developing of actual data mining and analysis for brain images.

5.3.Patient level data:

It could be beneficial if The data from all levels are used. Only one feature selection technique was used. It will be better if multiple feature selection techniques are used to find which one works the best with medical data. In future anyone must be able to look at patient's medical attributes and make subjective decisions.

5.4.Population level data:

The message board data existing work does have the potential to supply patients with reliable medical information, more real world testing should be implemented. New ways to find the optimal set of keywords/queries to use for predicting the occurrence of ILI epidemic must be done. Also work must be done if research done in one area of work can be translated to another. In twitter data, more work should be done on developing methods to best determine what keywords to use for study as well as texting more text classification methods in order to reduce noisy posts.

5.5.Translational bioinformatics:

The data from all levels of human existence must be considered. Through this way, questions from all levels can be easily answered. Not only combining molecular level data with other levels, by attempting to make connections across as many levels o f data as possible, we get Big Volume, Velocity, Variety, Veracity and Value.

VI. CONCLUSION

We discussed a number of recent studies being done with the most popular subbranches of Health Informatics, using Big Data from all accessable levels of human existence to answer questions throughout all levls. The use of Big Data provides advantages to Health Informatics by allowing for more test cases or more features for research, leading to both quicker validation of studies.

Health care is a data-rich domain. As more and more data is being collected, there will be increasing demand for Big Data Analytics.

Efficiently utilizing data can yield some immediate return in-terms of patient outcome and lowering care costs.

ACKNOWLEDGEMENT

We are greatly indebted to our KLUniversity that has provided a healthy environment to drive us to achieve our ambitions and goals. We would like to express our sincere thanks to our incharges Mrs. Anuradha, Mr. Kameshwara Rao, Mr. D.B.K.Kamesh for the guidance and support that they have provided in completing this paper.

REFERENCES

[1] Demchenko Y, Zhao Z, Grosso P, Wibisono A, de Laat C (2012) Addressing Big Data challenges for Scientific Data Infrastructure In: IEEE 4th International Conference on Cloud Computing Technology and Science (CloudCom 2012).
IEEE Computing Society, based in California, USA, Taipei, Taiwan, pp 614–617

[2] McDonald E, Brown CT (2013) khmer: Working with big data in Bioinformatics. CoRR abs/1303.2223: 1–18

[3] Bennett C, Doub T (2011) Data mining and electronic health records: selecting optimal clinical treatments in practice. CoRR abs/1112: 1668

[4] Haferlach T, Kohlmann A, Wieczorek L, Basso G, Kronnie GT, Béné MC, De Vos J, Hernández JM, Hofmann WK, Mills KI,
Gilkes A, Chiaretti S, Shurtleff SA, Kipps TJ, Rassenti LZ, Yeoh AE, Papenhausen PR, Wm Liu, Williams PM, Fo R (2010) Clinical utility of microarray-based gene expression profiling in the diagnosis and subclassification of leukemia: report from the international microarray innovations in leukemia study group. J Clin Oncol 28(15): 2529–2537.
[http://jco.ascopubs.org/content/28/15/2529.abstract]

[5] Salazar R, Roepman P, Capella G, Moreno V, Simon I, Dreezen C, Lopez-Doriga A, Santos C, Marijnen C, Westerga J,Bruin S, Kerr D, Kuppen P, van de Velde C, Morreau H, Van Velthuysen L, Glas AM, Van't Veer LJ,
Tollenaar R (2011) Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer. J ClinOncol 29:17–24. [http://jco.ascopubs.org/content/29/1/17.abstract]

[6] Annese J (2012) The importance of combining MRI and large-scale digital histology in neuroimaging studies of brain connectivity and disease. Front Neuroinform 6: 13. [http://europepmc.org/abstract/MED/22536182]

[7] Van Essen DC, Smith SM, Barch DM, Behrens TE, Yacoub E, Ugurbil K (2013) The WU-Minn human connectome project: an overview. NeuroImage 80(0): 62–79. [http://www.sciencedirect.com/science/article/pii/
S1053811913005351]. [Mapping the Connectome]

[8] Estella F, Delgado-Marquez BL, Rojas P, Valenzuela O, San Roman B, Rojas I (2012) Advanced system for automously classify brain MRI in neurodegenerative disease In: International Conference on Multimedia Computing and Systems (ICMCS 2012). IEEE, based in New York, USA, Tangiers, Morocco, pp 250–255

[9] Campbell AJ, Cook JA, Adey G, Cuthbertson BH (2008) Predicting death and readmission after intensive care discharge. British J Anaesth 100(5): 656–662. [http://europepmc.org/abstract/MED/18385264]

[10] Bradley AP (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms.Pattern Recognit30(7):1145–1159. [http://www.sciencedirect.com/science/article/pii/S003132039 6001422]

[11] Ouanes I, Schwebel C, Franais A, Bruel C, Philippart F, Vesin A, Soufir L, Adrie C, Garrouste-Orgeas M, Timsit JF, Misset B (2012) A model to predict short-term death or readmission after intensive care unit discharge. J Crit Care 27(4): 422.e1–422.e9. [http://www.sciencedirect.com/science/article/pii/S088394411 1003790]

[12] Rolia J, Yao W, Basu S, Lee WN, Singhal S, Kumar A,
Sabella S (2013) Tell me what i don't know - making the most of social health forums. Tech. Rep: HPL-2013–43. Hewlett Packard Labs [https://www.hpl.hp.com/techreports/2013/
HPL-2013-43.pdf]

[13] Signorini A, Segre AM, Polgreen PM (2011) The use of twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. PLoS ONE 6(5): e19467. doi:10.1371/journal.pone.0019467

[14] Sarkar IN, Butte AJ, Lussier YA, Tarczy-Hornoch P, Ohno-Machado L (2011) Translational bioinformatics: linking knowledge across biological and clinical realms. J Am Med Inform Assoc 18(4): 354–357. [http://jamia.bmj.com/ content/18/4/354.abstract]

AUTHORS

First Author – Neelima Vaddi, K L University
vdneelima3@gmail.com
Second Author – Ramya Priya. R, K L University
rattypriya@gmail.com

Correspondence Author – D.B.K.Kamesh, Associate Professor, K L University. kameshdbk@kluniversity.in