Journal of Medical Systems (2019) 43: 264
https://doi.org/10.1007/s10916-019-1409-z

IMAGE & SIGNAL PROCESSING

Phenotype Algorithm based Big Data Analytics for Cancer Diagnose

K. Sivakumar[1] · N. S. Nithya[1] · O. Revathy[1]

Received: 29 March 2019 / Accepted: 26 June 2019 / Published online: 4 July 2019

Abstract

Nowadays, Cancer diagnosis is one of the major challenging characteristics for treating cancer. The reality of cancer patients rely on the diagnosis of cancer at the early stages (either in stage 1 or stage 2). If the cancer is diagnosed in stage 3 or later stages means the changes of survival of the patient will become more critical. Normally, single patient records will generate a huge amount of data if the data could be manage and analyze means to solve many problems for identifying the patterns it will leads to diagnose the cancer. Recent work several machine learning algorithms are introduced for the classification of cancer. However still the classification accuracy of machine learning algorithms are reduced because of huge number of samples. So the proposed work introduces a new Hadoop Distributed File System (HDFS) is focused in this work. In this paper, the proposed phenotype techniques are used which handle and classifies the raw EHR (Electronic Health Record) and EMR (Electronic Medical Record). It is based on the HDFS and Two-Phase Map Reduce. Phenotype algorithm uses NLP (National Language Processing) tool which will analyze and classify the cancer patient data like gene mapping, age related data, image and ultrasonic frequency processing, identification and analysis of irregularities, disease and personal histories. In this paper, the three factorized model is used which calculates the mean score values. The values are calculated by disease stage, pain status, etc. This paper focuses big data analytics for cancer diagnosis and the simulation results shows the proposed system produces the highest performance.

Keywords Big Data · HDFS · Phenotype Algorithm · NLP tool

Introduction

Early stage of cancer diagnosis is one of the world's most important problems in cancer discovery process. Every normal cell in a human body contains same number of chromosomes and same volume of DNA. But in contrast the cancer cells contain abnormal growth in the chromosome count. The big data analysis helps to identify such cells and remove them to reduce their abnormal growth. Big data is an energetic technology that improves the quality of research and makes better results quickly. The tumor cells undergo various mutations and combinations which was very vast to have a detailed study on them. Researchers take large data sets and identify the changes to study about particular tumors.

Health care providers provide some effective tools to identify the cancer. Early stage of cancer could be identified by the following symptoms that are bowel changes, persistence back pain, unusual coughing, rectal bleeding, urinary changes, and blood in urine. Most of the cancers arise through the skin which is mostly occurring in glands, breast, testicle and glands or lymph nodes. Blood tests are used to diagnose the cancer, using this tests blood cancer could be detected and few of the abnormal cells are founded through this type of test. Nowadays many type of cancers could affects the people and these stomach cancer, brain cancer, breast cancer, pancreatic cancers are some of the painful cancers.

Microscopic evaluation of biopsy samples which diagnose the PCa (prostate cancer) is a one type of skin cancer. American cancer society says nowadays most of the people die of this prostate cancer. Magnetic resonance (MR) or Transrectal ultra sound (TRUS) using this samples of a prostate could be collected. Pathologist readings are used

This article is part of the Topical Collection on *Image & Signal Processing*

✉ K. Sivakumar
sivakumar.karuppan@gmail.com

N. S. Nithya
nithyaphd@gmail.com

O. Revathy
revathy123@gmail.com

[1] Anna University, Chennai, India

to improve the prostate cancer diagnosis accuracy [1, 2]. Traditional biopsy uses fluorescence technique which identifies the cancer. Process of photon emission is named as fluorescence and the auto fluorescence happens in tissues like nerve, muscle, etc.

Laser induced fluorescence (LIF) is a one of the method used in spectroscopic which specifies the reciprocal action of electromagnetic radiation. Molecule structures, detection of species, visualization flows and measures. To detect the cancer, hyper spectral image uses variety of applications. HIS combine multiple endoscopes based on LCTF technology cancer is a one of the heterogeneous disease and the cancer biomarker discovery is a technology to detect the snapshots but it could not make any dynamic and longitudinal changes in cancer landscape. To overcome this Big Data assemble some distinct molecular features at DNA, RNA, and metabolite levels [3].

In big data platform the health care data's are stored supervised, semi-supervised and unsupervised manner. Medical records of Physician notes, lab reports, x-ray reports, and case histories are maintained by big health care data. The radio frequency identification maintains national health register data and expiry dates of medicine and surgical instruments. Big data uses the genetic algorithm and cancer genome atlas (CGA) using these patient records could be analyzed. To identify the cancer patient records the prototype could be created which is implemented on hadoop which will help the doctors in diagnosis and prognosis of the cancer patient. Based on HDFS the prototype could be developed which is useful to collect, clean, analyze, and manage the data's in effective manner and big health care data improves patient outcomes [4]. In this paper, the proposed phenotype techniques are used which handle and classifies the raw EHR (Electronic Health Record) and EMR (Electronic Medical Record). It is based on the HDFS and Two-Phase Map Reduce. Phenotype algorithm uses NLP (National Language Processing) tool which will analyze and classify the cancer patient data like gene mapping, age related data, image and ultrasonic frequency processing, identification and analysis of irregularities, disease and personal histories.

Related Works

In paper [5] Cyber enabled system and IoT (internet of things) techniques are used which uses large amount of data with various structures. Most of the big health care data solutions are built on hadoop eco system and HDFS (distributed file system) but it has many challenging issues like Inefficient data, require large space and power consumption to overcome these drawbacks the proposed method of data aware module and genetic algorithms are used. The data aware module is used for hadoop eco system and genetic algorithm for distributed encoding technique. Using these, distribution of data and cluster analysis could be managed, larger data types are handled and it optimize the query time.

First, collect all the health care data's with the help of genetic algorithm and data aware module after collecting the data it will be transformed into a network graph and found a patterns in that graph. Data aware module which distributes the data into several data blocks. Data aware analysis and graph distribution acts like data storage which running on top of the HDFS. Clustering framework directly interacts with distributed file system and provides updated clusters to HDFS.

In paper [6] Classification technique uses wide variety of applications which is handled by only few techniques. So, the class distribution is more difficult and imbalanced classification problem could take place. The binary classification problem is one type of supervised learning problem which classifies the elements into variety of groups. To overcome this draw back the proposed method of EUS (Evolutionary under Sampling) techniques are used which is designed in a parallel manner. It deals with class imbalance problem to overcome this, map reduce strategy are used. To speedup the undersampling process without any losing accuracy Windowing approach is used. To overcome the large scale problems two stage map reduce approaches are used which performs EUS preprocessing and classify test set.

In paper [7] the transfer of large datasets using HTTP and FTP has the advantage of low cost and high throughput. But on the other side it results in bottleneck which reduces the speed of transmission. The potential transmission can be done by either minimizing the data or expanding the network. The data minimization algorithm is designed to transfer the bio genomic data bases in a secure way. Goal of this algorithm is to reduce the size of the data and to have a secure transmission. This technique involves assigning different codeword's to the same character for different times and they run in CNN (Coded Neural Network).

In paper [8, 9] EUS (evolutionary under processing) which uses GS (global selection) and MS (majority selection) based on that EUS methods are used to remove instances among both classes and identify the minority class. Prototype selection algorithm could not manage all imbalance problems. In evolutionary under processing, the MS (majority selection) mechanism helps to get accurate subset of instances comparing to the GS (global selection) mechanism. The global selection mechanism is to achieve a highest reduction rate. A difference between GM and AUC is geometric mean which evaluates the subset accuracy and area under curve which evaluate the measures.

Manogaran et al. [10] proposed to make use of a Bayesian hidden Markov model (HMM) with Gaussian Mixture (GM) Clustering approach in order to model the DNA copy number change across the genome.

Sun and Reddy [11] introduced the properties and related mining issues on handling with big medical data. Several of those insights come from medical informatics community, which is highly related towards data mining however focus on biomedical specifics. Review of several associated works from data mining location as well as medical informatics venues in the direction of share with the audience's key issues and trends in healthcare analytics research, with various applications ranging from clinical text mining, predictive modeling, survival analysis, patient similarity, genetic data analysis, and public health. The tutorial will comprise several case studies handling with many of the significant healthcare applications.

Rodrigues et al. [12] introduced on visual and machine learning analysis of medical data acquired by means of diverse nanotech-based methods and on algorithms for Big Data infrastructure.

To measure the classification performance the positive class and negative classes are used. Positive prediction is used to predict the true positive and false positive rates. Negative prediction which predicts the true negative and false negative rates.

Positive Prediction:

$$\text{True Positive Rate} : TP_{rate} = \frac{TP}{TP + FN} \quad (1)$$

$$\text{False Positive Rate} : FP_{rate} = \frac{FN}{FP + TN} \quad (2)$$

Negative Prediction:

$$\text{True Negative Rate} : TN_{rate} = \frac{TN}{FP + TN} \quad (3)$$

$$\text{False Negative Rate} : FN_{rate} = \frac{FN}{TP + FN} \quad (4)$$

where TP-True Positive, TN-True Negative, FP-False Positive, and FN –False Negative.

An aim of the classifiers is to minimize the rate of false positive and false negative or a same way to maximize the true positive and true negative rates. Proteogenomic is based on proteomics and genomics. These are the two high throughput technologies to diagnose the cancer. Proteogenomic uses wide variety of applications such as gene annotation, virology and bacteriology, human neurology, and cancer biology. NGS techs which create some problems between analyze storage, transfer and visualization. To overcome this develops high performance solutions to storage, analysis and transmission of these proteogenomics.

Proposed system

Predicting cancer cells in early stage is one of the most important problems in cancer discovery process. The proposed system of cancer diagnosis is shown in fig. 1 which includes Phenotype algorithm for extracting EHR data, HDFS which act as a primary data storage system, using this health care data could be store and retrieved, Two-phase Map Reduce for processing and producing big data sets. The work flow of proposed system is given in Fig. 1. From Fig. 1 shows the HDFS (distributed file system) which maintains all electronic health records and medical records. The phenotype techniques based on two-phase map reduce. Each map and reduce phase have (K, V) pairs as input and output. First the map phase gathers the information's from the file system based on that information map function will run without any outside help. Now the patient details are stored in true patient state which maintains large amount of Raw EHR data's. Raw EHR data's are very difficult to handle so that, the proposed phenotype algorithms are used. It can easily understand, classify and predict Raw HER data's. After that all the classified data's are merged with the help of intermediate (K, V) pairs and send it to the shuffle step. In this, the shuffle step maintains NH and HCP.

Natural Health which maintains:

Pre-cancer Development, it specifies incidence, location, growth, histology, and malignant transformation.
Cancer Incidence, it specifies growth and the symptoms.
Cancer Mortality, it specifies staging and survival.

Health care process which maintains:

Behavior Modification which will specify the human physical activities.
Screening, Based on the cancer incidence it specifies the multi target Stool DNA.
Diagnosis Treatment, Based on the cancer mortality it specifies the Primary care, gastroenterologist.

On the other hand reduce operations could wait until map function finish its operation. Once map function finishes its operation means the reduce phase, this resulting list as its input to perform its operations after completing its process which will return the final response of the system.

Big Healthcare Data

Big data health care industry generates huge amount of data. Normally the big health care data in the form of supervised, semi supervised and un- supervised manner. Each data could be handled by the help of data handler.

Structured, unstructured and semi-structured data are specified as follows.

Fig. 2 shows the big data characteristics are called 4 V's model. They are:

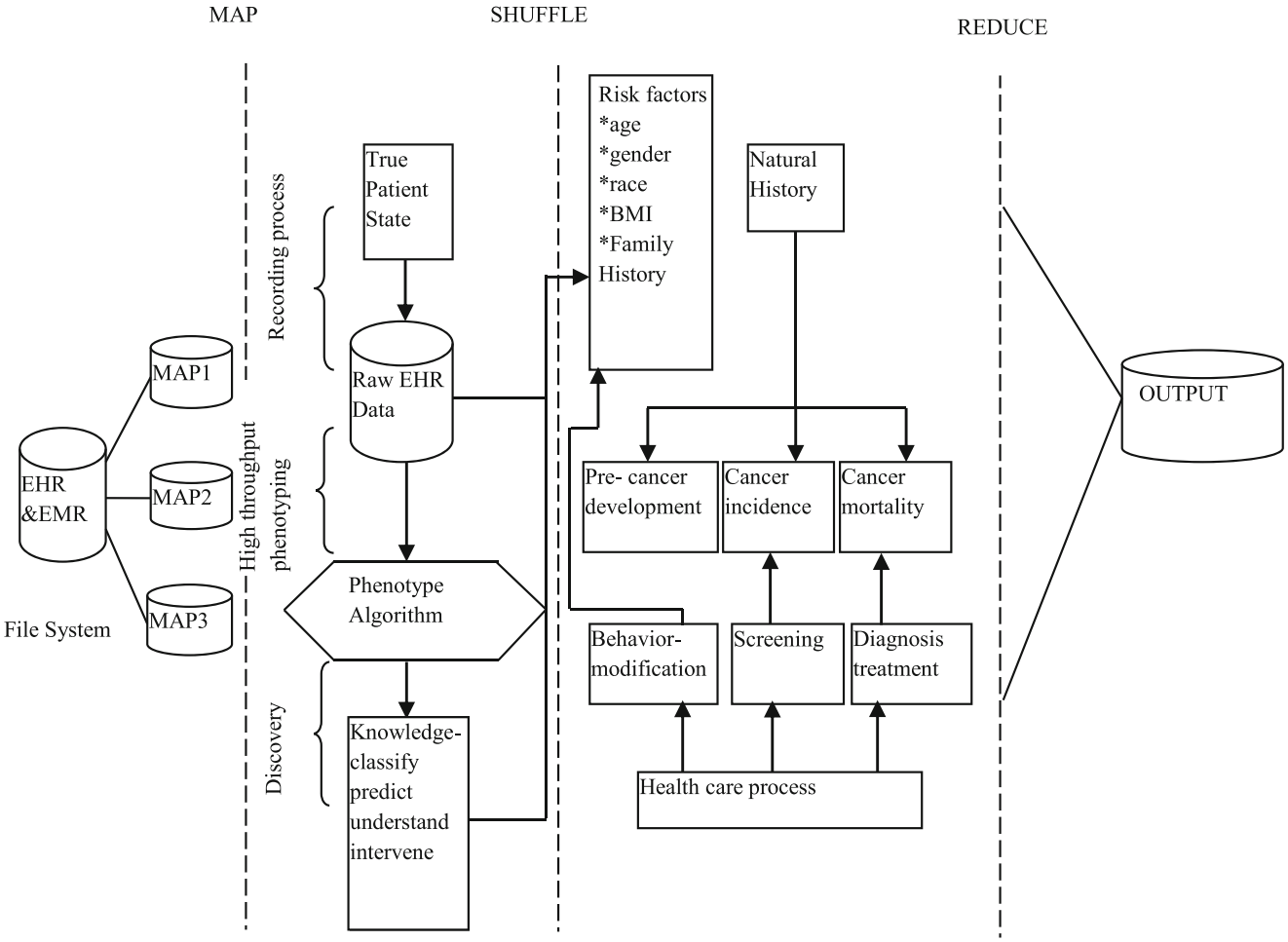


Fig. 1 Flow diagram of proposed big data analytics for Cancer Diagnosis

1. Large volume:

Large volume is one of the most extraordinary characteristic used in big data. It collects data from different sources like sensor or machine to machine data which is fully based on the Hadoop. Main use of this is implies and explosive in the amount of data.

2. Large variety:

Large variety is one of the unique characteristic of big data. In this model data comes in different format like structured, unstructured, and semi-structured. E.g.: text, image, videos or graphics. Unstructured data are collected from internet or mobile devises etc.…

3. Large velocity:

Large velocity it provides online services. It argues with big data to generating or process a data in real time manner. It is fully based on the speed of data processing

4. Large veracity:

It finds the incomplete objects, noisy objects and inaccurate objects. It faces many challenging issues like data mining and information processing and LV is one of the characteristics used in 4 V model.

Phenotype Technique

Phenotype is one of the most recently used techniques in big data health care industry. More than 80% of clinical data are in the form of unsupervised and semi supervised manner. Raw EHR data's are very difficult to handle and take more time to understand. So, the proposed technique of phenotype algorithm is used which classify, predict and understand EHR data's easily. The following diagram specifies how the Raw medical data could be classified using the proposed technique. The methods used to develop the HER phenotype is shown in the fig. 3.

The HER and EMR of patient details are collected in a wide manner. The phenotypes are analyzing the raw EHR and EMR records based upon prevalence of the disease.

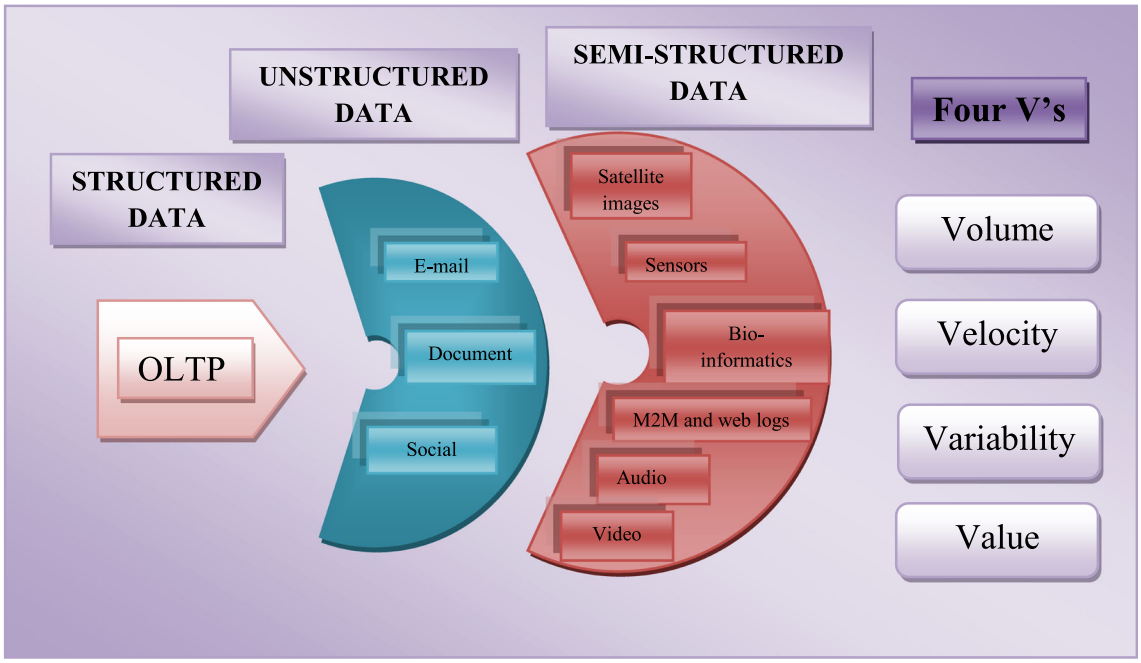


Fig. 2 Flow of big health care data

Initially the patient medical and health records are screened to differentiate data with the help of phenotype technique. Then the patient data's are collected in the sensitive data mart. The following fig. 4 shows how the phenotype algorithm specifies the EHR data could be screened and classified.

Phenotype algorithm is used to identify the Type 1 diabetes melitus (T1DM) and type 2 diabetes melitus (T2DM) from EHR (Electronic Health Record) and EMR (Electronic Medical Record). Dx: it specifies the diabetes and it define records using international classification of diseases and ICD-9 codes. Med: specifies the mediations, physcn: it specifies the physicians, and Rx: specifies the prescriptions [13].

Classification algorithm is used to classify all the data's. Using phenotype algorithm finally the prevalence and type of disease can be diagnosed by screening patient's data. Phenotype uses the following tools. They are:

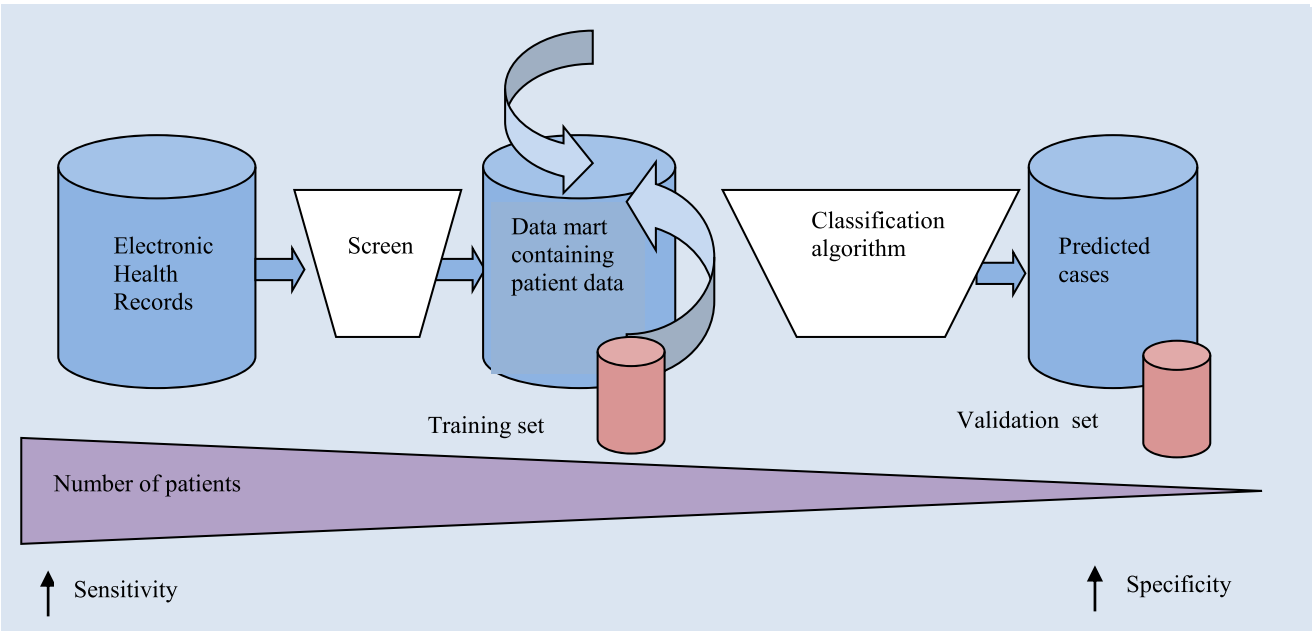


Fig. 3 Overview of methods used to develop HER phenotype technique

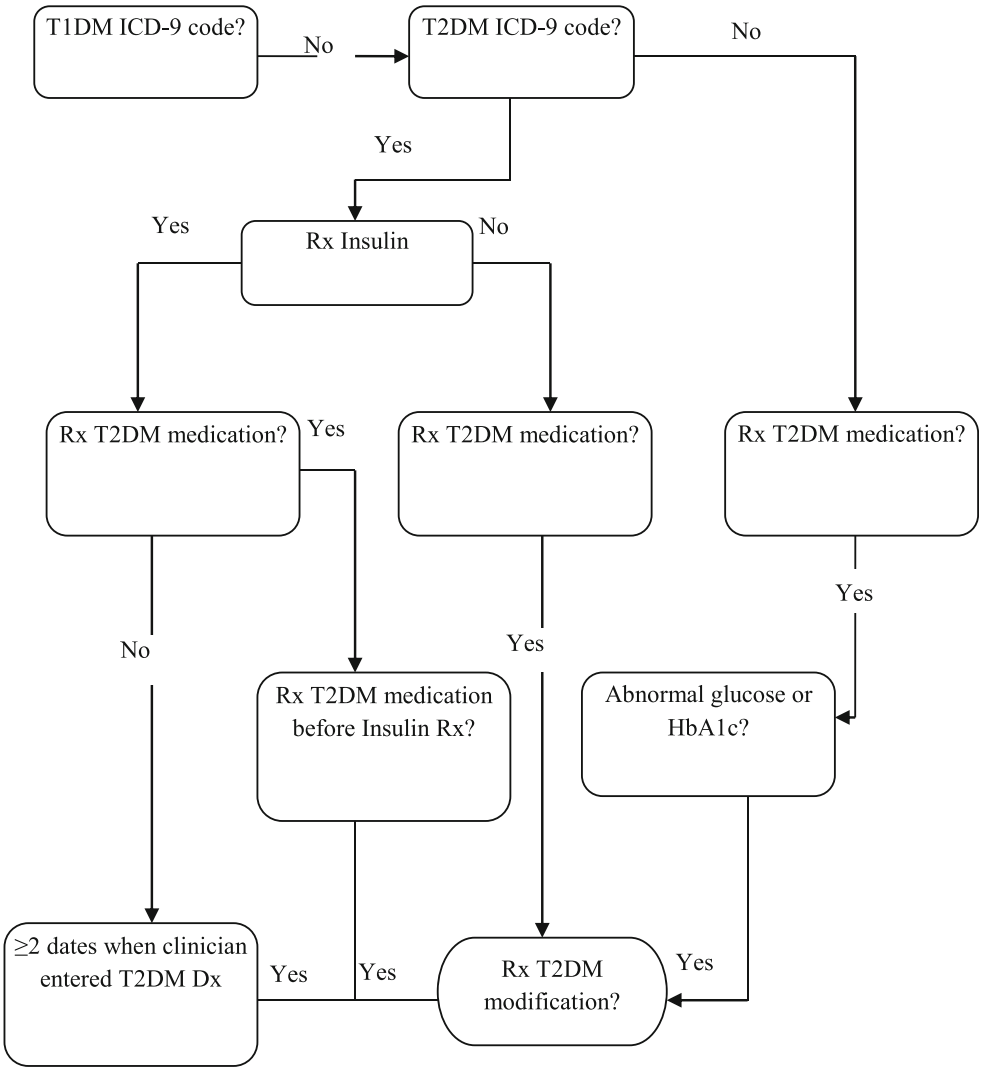


Fig. 4 EHR and EMR data screened and classified using Phenotype algorithm

1. Erc: (Emerge Record Counter)

Emerge record counter which combine DNA biorepositories with EMR systems based on the National human genome research institute (NHGRI). It provides large amount of data's for research planning determination and viability assessment. In this the DNA repositories combined to electronic medical record which includes legal, ethical and social issues that are related to these endeavors [14].

2. PheKB: (Phenotype Knowledgebase)

To identify the characteristics of patient's health care data the Phenotype knowledgebase must validates the electronic algorithms. PheKB provides collaborative environment to develop and validates the electronic phenotype algorithm. PheKB is a work flow management system. The learning center supports creation, and validation of phenotype knowledge.

3. PheWAS: (PheWAS catalog)

PheWAS catalog is based on the phenotype outcomes or measurements which uses the quantitative research technique. It clearly identifies the false positive associations and which is based on the SNP selection process. PheWAS is mainly used to discover the relationships among SNP's and phenotypes [15].

4. SPHINX: (Sequence and Phenotype Integration Exchange)

SPHINX is one of the web based tool using this it will run large amount of web applications. It is used to explore a data for hypothesis generation. Web apps always run inside the web browser. Sequence PHINX it act as searchable database using this it will identifies list of genes, drugs, and pathways in the catalog.

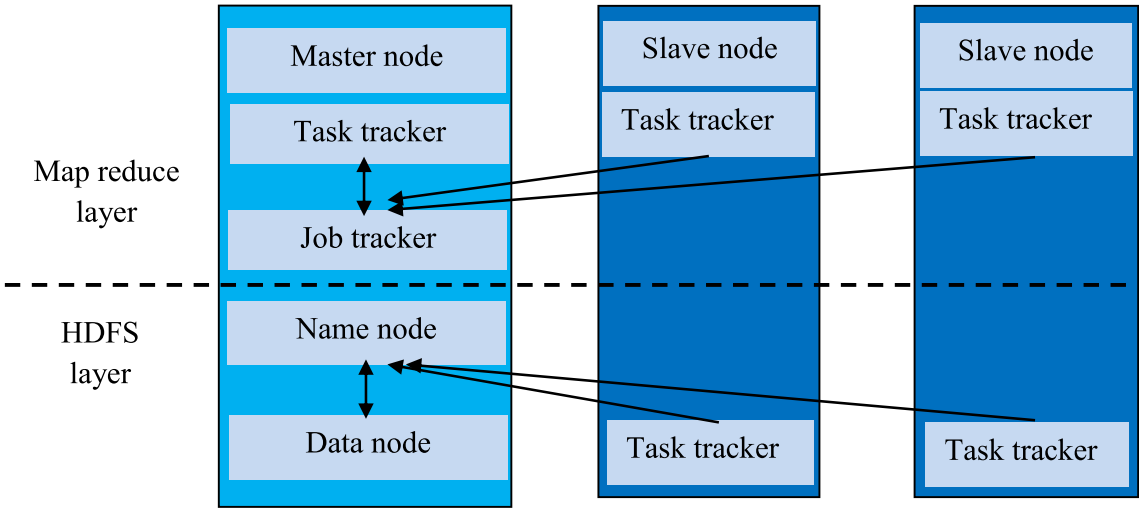


Fig. 5 High Level Architecture of Hadoop Model

HDFS: (Hadoop distributed file system)

Distributed file system is one of the primary data storage systems which support large amount of data sets. It uses name node and data node. Based on these name node and data node architecture distributed file system could be implemented. It provides high performance data access for hadoop clusters. Distributed file system supports rapid transfer of data into several nodes. It closely connected with two phase map reduce. Distributed file system is one of the most important components in hadoop which is run on commodity of hardware. Hadoop and its other component use the fault tolerant storage layer which is provided by HDFS. Hadoop framework maintains the data management process and storage analytic solutions. Map Reduce [16] allows large scale data processing with the help of parallel and distributed algorithm. In map reduce environment two basic stages are applied to improve the fault tolerance, data partition and management.

$$HDFS(RSS) + Map\ Reduce(DC) = Hadoop \quad (5)$$

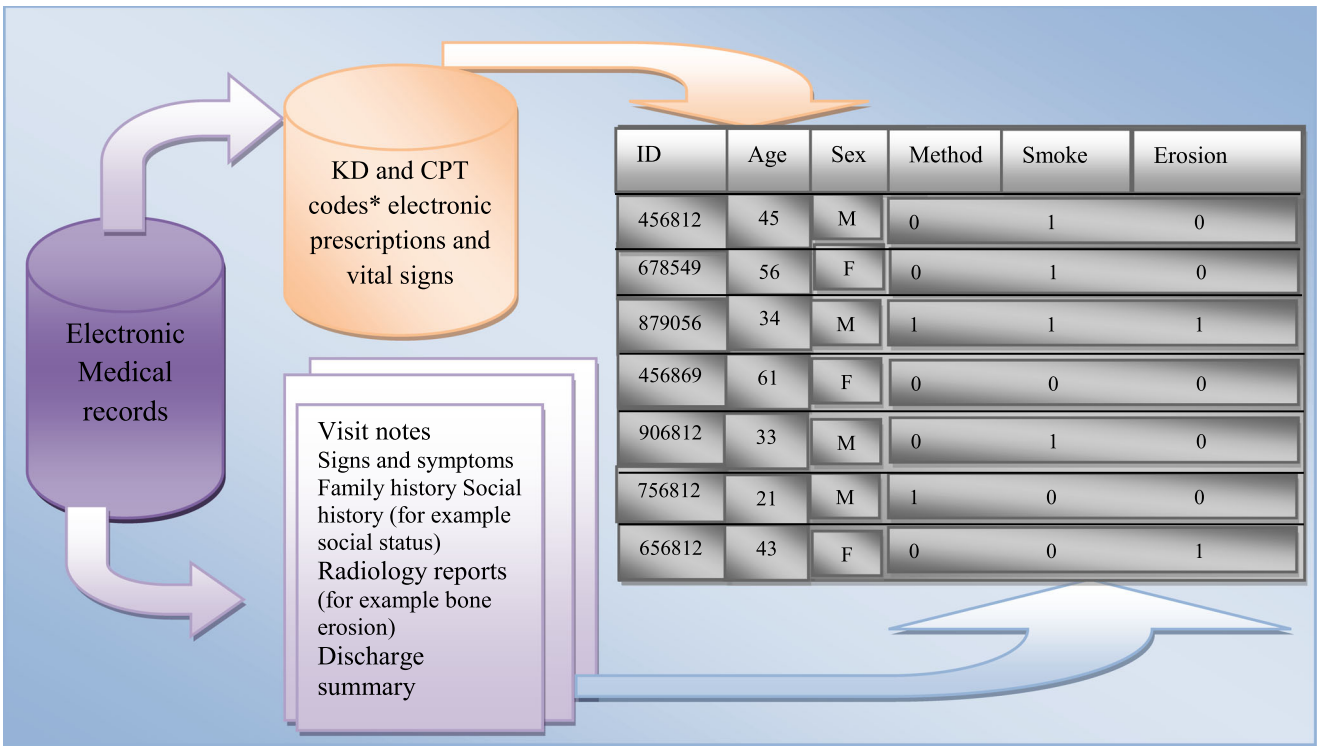


Fig. 6 A general review of two main types of EHR and EMR data, structured and unstructured and how these data can be integrated

Table 1 Using NLP technique big health care data are classified

Relational databases

NLP Tools Required

Diagnosis Codes

Fake ID	Entry-Date	Code
320611	7/9/2015	42.81
432561	5/14/2016	78.99
340611	8/31/2018	72.11
304633	8/6/2018	67.12
240446	8/6/2017	81.34
345687	5/3/2016	112
223698	2/9/2017	13.24
243587	7/8/2015	23.22
224786	7/8/2016	34.67
236478	9/9/2017	12.67
244356	5/6/2017	34.123

Lab Tests

Fake ID	Test	Entry-Date	Value
645789	P01	2/3/1996	13.22
789623	LDL	4/7/1998	21.56
235689	P0C2	6/8/1996	41.67
127896	HDL1	5/8/1999	61.11
579432	HDL	6/8/1998	31.67
347613	PO12	9/5/2014	34.88
654321	P0C2	3/8/2012	11.21
265423	LLD2	3/8/1996	24.35
986421	POC4	6/8/1999	51.11
243783	ALC1	7/8/1998	54.67
316539	P011	6/4/1997	11.21
210087	ACL2	7/9/2011	61.71
453716	P02	8/2/2010	11.42
613857	LLV	5/4/1999	63.44
231697	LC01	8/6/1998	21.37
421957	PO1	8/6/2011	37.41
546921	VLL2	6/4/1998	41.21
441692	P04	8/2/1998	22.11
771336	VL01	6/4/1996	17.74
231231	VLL02	8/8/1999	31.29

Problem Lists

Medications known to be prescribed:
Copra 751 mg ½ tab q am and pm
Deviant 61 mg by mouth daily aspirin 326 mg 2 tablet by mouth daily
Klopidgrel 76 mg tablet 2 tablet by mouth daily

Adverse and Allergic drug reactions:
Sulfa drugs

Significant medical diagnoses:
Aneurysm
Seizure disorder
Heartburn

Known significant medical diagnoses:
Sulfa drugs
Aneurysm
Seizure disorder
Heartburn

Known significant operative and invasive procedures:
2006
Appendectomy
2007
Stents put in
**DATE
[SEP 29 05]

Allergic Drug Reactions:
Sulfa drugs
Aneurysm
Seizure disorder

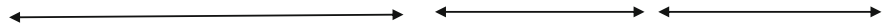
Clinical Notes

Exam: bilateral digital screening MAMMOGRAM WITH CAD
**date[April 18 01]:

COMPARISION:
**DATE[August 05 01]

TECHNIQUE:
Standard CC and MLO views of both breasts were obtained.

FINDINGS: The breast parenchyma is heterogeneously dense. The pattern is extremely complex with post surgical change seen in the righter upper outer quadrant and scattered benign-appearing classification seen bilaterally.
RIGHT: No interval change. No current evidence of malignancy.
LEFT: Possible developing asymmetry superior aspect left breast for which further evaluation by true lateral and spot compression views recommended.
BI-RADS Category 01: Incomplete assessment.
IMPRESSION:
RIGHT: No interval change. No current evidence of malignancy…….



Structured Data

Semi-structured Data

Unstructured Data

Table 2 Phenotype algorithm converts the Raw EHR data

Cancer Type	Clinical Endpoint	Phenotype Algorithm	Benchmark	Improvement (%)	Training data
Brain	Treatment response	MR	Random	N/A	Mixed
Liver	survivability	Decision tree	Statistics	1	Genomic
Liver	Treatment	Genetic algorithm	Statistics	N/A	Clinical
Liver	Survivability	Naïve bayes	Statistics	N/A	Proteomic
Lung	Response	MR	Expert	17	Genomic
Brain	Recurrence	MR	Random	13	Clinical
Bladder	Treatment	Fuzzy logic	Expert	7	Mixed
Bladder	Susceptibility	Genetic algorithm	Statistics	1	Mixed
Bladder	Survivability	ANN	N/A	N/A	Clinical
Bladder	Treatment	ANN	N/A	14	Genomic
Lymphoma	Response	Decision tree	Statistics	2	Proteomic
Lymphoma	Survivability	SVM	Experts	N/A	Clinical
Head/neck	Survivability	SVM	Statistics	14	Clinical
Head/neck	Treatment	ANN	N/A	29	Proteomic
Ocular	Treatment	Fuzzy logic	Expert	0	Genomic
Prostate	Recurrence	Decision tree	Statistics	N/A	Clinical
Skin	Recurrence	Decision tree	Random	29	Mixed
Stomach	recurrence	MR	N/A	1	Mixed
Skin	Recurrence	ANN	N/A	N/A	Mixed
Thyroid	Recurrence	ANN	N/A	1	Clinical
thoracic	Recurrence	MR	Statistics	18	Proteomic
Pleural mesothilioma	Treatment	MR	Statistics	N/A	Genomic
Neck	Treatment	Genetic algorithm	Random	16	Clinical
Neck	Clustering	ANN	Expert	N/A	Clinical

Table 3 Database comprising the individual cancer patient records

Name	Birth	Sex	Zip code	Religion	Disease
Matthau	12/03/1982	Female	21,502	Hindu	Cancer
Mamtha	21/11/1980	Female	18,763	Christian	Cancer
Miry	01/08/1989	Male	32,645	Hindu	Cancer
Shank	08/04/1986	Female	28,372	Christian	Cancer
Salina	12/08/1989	Female	32,144	Hindu	Cancer
John	24/12/1986	Male	77,621	Christian	Cancer
Aisha	19/03/1985	Female	22,139	Muslim	Cancer
Vanity	07/02/1988	Male	44,062	Hindu	Cancer
Lilly	12/08/1983	Female	79,334	Christian	Cancer
Binue	22/11/1989	Female	22,386	Hindu	Cancer
Menu	20/12/1999	Female	27,644	Hindu	Cancer
Abraham	13/02/1992	Male	94,573	Christian	Cancer
Echlin	16/07/1985	Female	22,271	Hindu	Cancer
Vishnu	06/02/1988	Male	16,423	Hindu	Cancer
Ajay	01/04/1997	Male	21,055	Hindu	Cancer
Ayesha	12/02/1995	Male	32,256	Hindu	Cancer
Arvin	21/12/1996	Male	44,032	Hindu	Cancer
Azusa	26/11/1987	Female	55,521	Hindu	Cancer

Hadoop supports Map Reduce and HDFS layers the following fig. 5 which specifies these two layers.

The Map Reduce layer has two nodes that is master node and another one is slave node. The master node will schedule the jobs for slave nodes and monitor that scheduled jobs. In execution time if any task could be failed means the master node will re-execute that failed task. Two phase map reduce which uses key value pairs using this it takes a set of data and converting that data into another set of data. Hadoop is a framework it provides distribution, scheduling and parallelization services. HDFS is one of the storage layers. It has two nodes that are name node and data node. Name node

Table 4 Level of symptom interference with daily function

Functional area	Prevalence (%)	Severity
Enjoyment of life	75.3	5.76
General activity	76.9	6.82
Work	78.3	7.06
Mood	78.5	4.85
Relation with others	49.1	3.45
Walking	59.9	4.78

Table 5 Three factor model for factor matrix

Symptom	Factor Loading		
	Factor1	Factor2	Factor3
Fatigue	0.767		
Sleep disturbance	0.618		
Pain	0.741		
Distress			0.739
Vomiting		−0.977	
Nausea		−0.871	
Sadness			0.728
Eigen value	4.14	3.59	3.42
Variance explained	44.78%	38.67%	36.78%
Total variance explained			66.43%

handles the file system Meta data and save that data into different file blocks. It is designed to be low cost hardware.

Map function is applied in a parallel manner to every pair in the data set

$$\text{Map}\ (\text{K1}, \text{V1}) \rightarrow \text{List}\ (\text{K2}, \text{V2})$$

Reduce function is applied each group in a parallel manner. Finally it produces a multiple values in same domain

$$\text{Reduce}\ (\text{K2}, \text{List}\ (\text{V2})) \rightarrow \text{List}\ (\text{V3})$$

Map Function:
Require: Number of split k

1: Constitute: $T\,R_k$ with the instances of split k.
2: $R\,S_{k=}$ Phenotype ($T\,R_k$)
3: M_k = Build Model ($R\,S_k$)
4: Return M_k

Table 6 Three factors mean scores are calculated by stage of disease, pain status, chemotherapy and psychological status

Variable grouping	Mean		SD	t	p
Disease stage	Factor1 mean score			5.38	0.001
	Stage 3 to 4($n = 43$)	5.68	4.13		
	Stage 0 t0 2(n = 43)	3.75	3.63		
Pain status	With pain($n = 111$)	5.72[a]	3.46	6.89	0.001
	Without pain(n = 43)	3.32[a]	3.13		
Chemotherapy	Factor2 mean score			5.92	0.001
	Yes($n = 78$)	4.47	4.26		
	No($n = 95$)	2.29	3.27		
Depression status	Factor3 mean score			5.38	0.001
	Depressed($n = 79$)	5.68	4.13		
	Not depressed($n = 93$)	3.75	3.63		
Anxiety status	Anxious ($n = 68$)	6.71	3.59	9.35	0.001
	Not anxious($n = 98$)	3.37	3.49		

Reduce Function:

$$M_k, \{\text{Initially M} = \varnothing\}$$

1: $\text{M} = M \cup M_k$
2: return M

Where,
T R, be the training set stored in the HDFS as a single file.
H is the HDFS blocks.
M, is the disjoint subset, Each and every map task (map1,map2,….map n) which develops an associated $T\,R_k$, where $1 < k < m$, with the instance of each chunk, it specifies which T R could be divided.

Simulation Results

Bigdata health care industry generates large amount of data. Normally the big health care data in the form of supervised, semi supervised and unsupervised manner. Each data could be handled by the help of data handler. Initially, Distributed file system store all the EHR and EMR records shown in fig. 6. Structured data are easily identified by the people but unstructured and semi structured data are difficult to handle, understand and classify.

Case 1: Big health care data in the form of structured, semi-structured and unstructured format

All the patient health and medical related records are stored inside the EHR (Electronic health record) and EMR (Electronic medical record). The help of HDFS and Two-phase Map Reduce concept the patient's data could be store and retrieved. HDFS maintains all the HER and EMR. Structured, semi structured and unstructured data are classified separately. Relational databases maintain all the big health care data. NLP tool is a form of artificial intelligence it is used to handle, classify and understand the big healthcare data. National language processing that helps users to understand the machine language into human understandable format. This technique use statistics, semantics and machine learning technique to handle and extract the entities and relationships. NLP inserts part of speech tagging, automatic summarization, entity extraction, relational extractions, natural language understanding and recognition. The details of the NLP with code for diagnosis are shown in the Table 1.

Case 2: EHR and EMR data are classified using phenotype technique

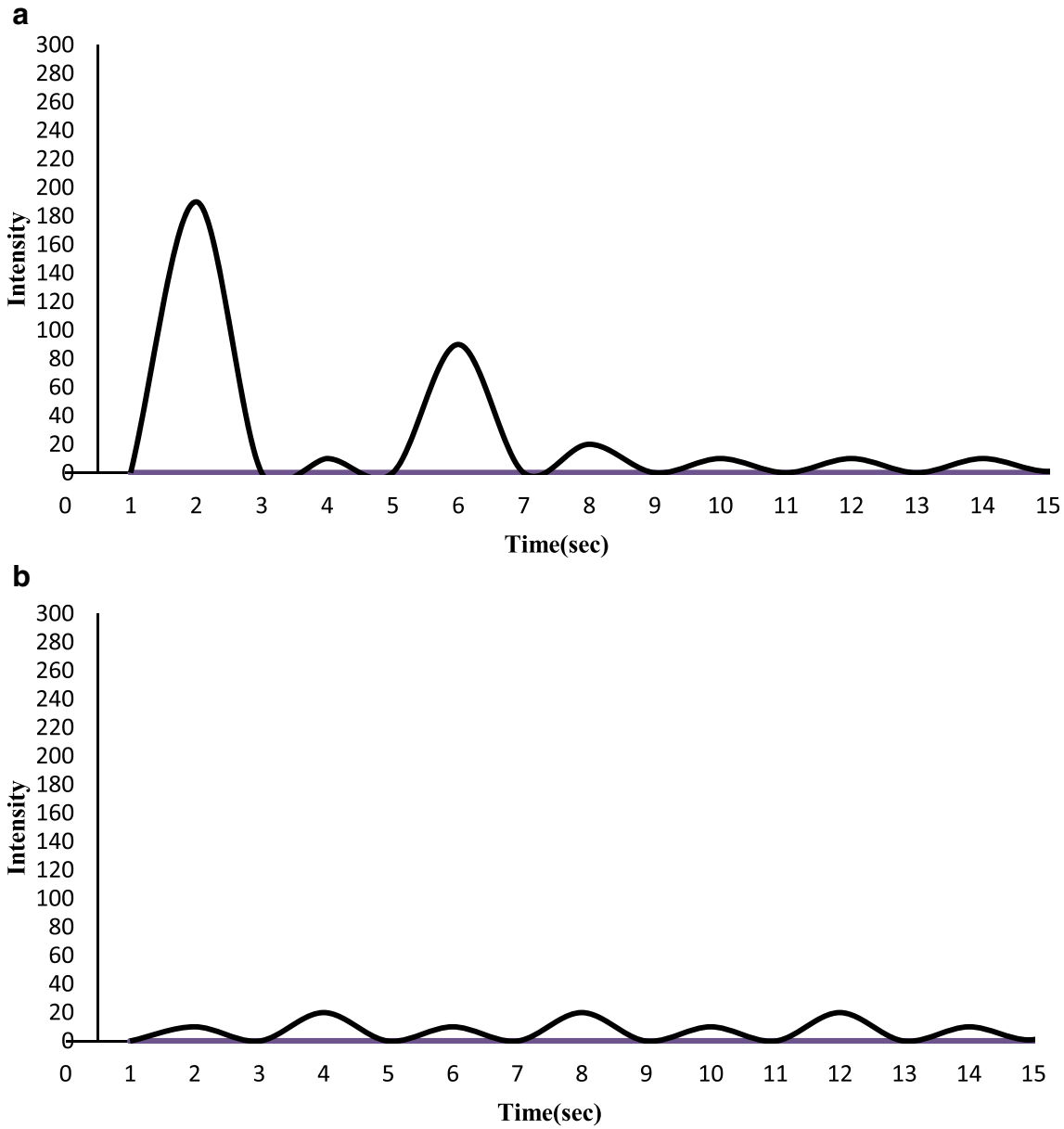


Fig. 7 Comparative big health care data among a healthy person and a cancer patient diagnosis by Phenotype. (a) Cancer patient (b) Normal patient

Raw EHR data are very difficult to handle, classify and understand. Convert these raw data into human understandable format is one of the toughest tasks. To overcome these proposed phenotype algorithm is used. First, the phonotype collects all the patient data into that it will classifies all cancer patient data. In this cancer patient report many of the data's in the form of Raw EHR format. Phenotype algorithm handles these raw data and classifies, predict and understand these into normal form discussed in Table 2.

Phenotype algorithm classifies the cancer patient data's. The above table describes the cancer types, clinical endpoints, algorithm choices, performance and type of training data. Using this detail the individual cancer patient details (see Table 3) are collected with the help of proposed phenotype algorithm.

The above Table 3 represents patient's name, date of birth, zip code, religion and type of disease.

Level of symptoms was calculated by averaging the symptom severity scores over who have the symptom is discussed in Table 4. Symptom interference working was the outlook of daily function with which symptom was most interfered.

(Mean = 6.09, SD = 6.31). Least interfered outlook of daily function was related with others (Mean = 3.73, SD = 4.46)

Case 3: Big data analytics for cancer diagnosis

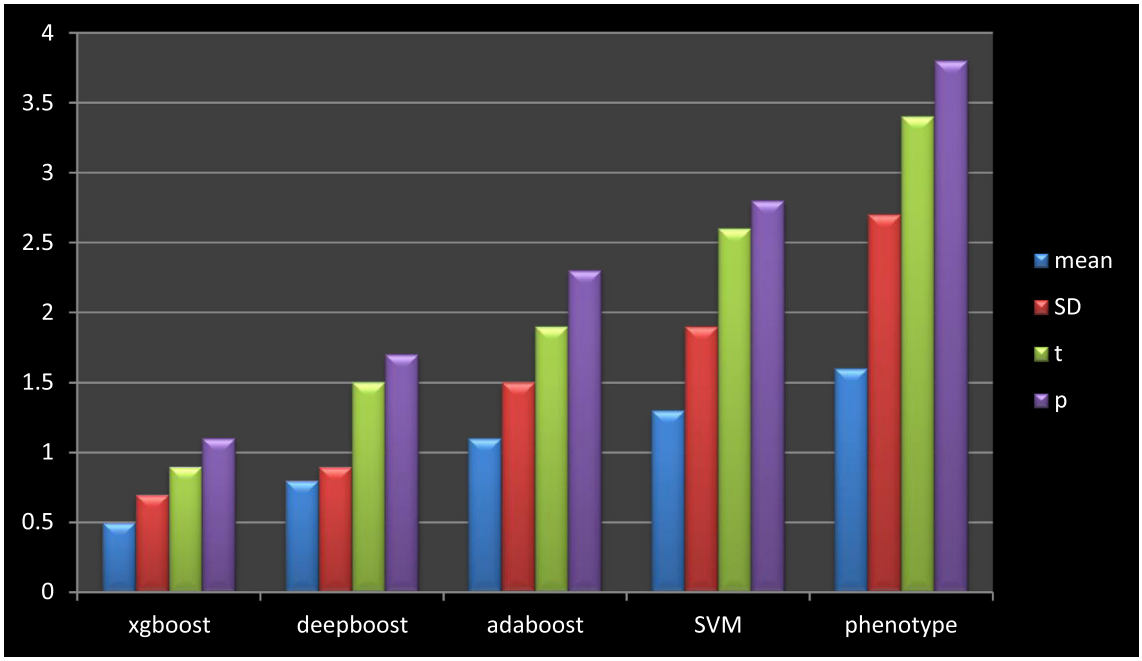


Fig. 8 Three factors mean score performance

The following Table 5 specifies the patient's factor scores, pain status, chemotherapy and psychological status.

In factor1 five items are weighted: fatigue, sleep disturbance, pain, Eigen value, variance explained. In factor2 two items are weighted: vomiting and nausea. In factor3 two items are weighted: distress and sadness. The above three factors specifies the sickness of symptom cluster, gastrointestinal symptom cluster and emotional symptom cluster discussed detail in Table 6. The alpha coefficient of the symptom cluster was 0.89; gastrointestinal symptom cluster was 0.99 and 0.76 for the emotional symptom cluster. The connection between the above three factors are moderate with Pearson r coefficients ranging from 0.54 to 0.67.

Using phenotype algorithm the cancer patient different stages of pain, depression status, anxiety status and mean score factors are calculated (Fig. 7).

The above graph shows the difference between normal patient breath and cancer diagnosed patient breath. Using

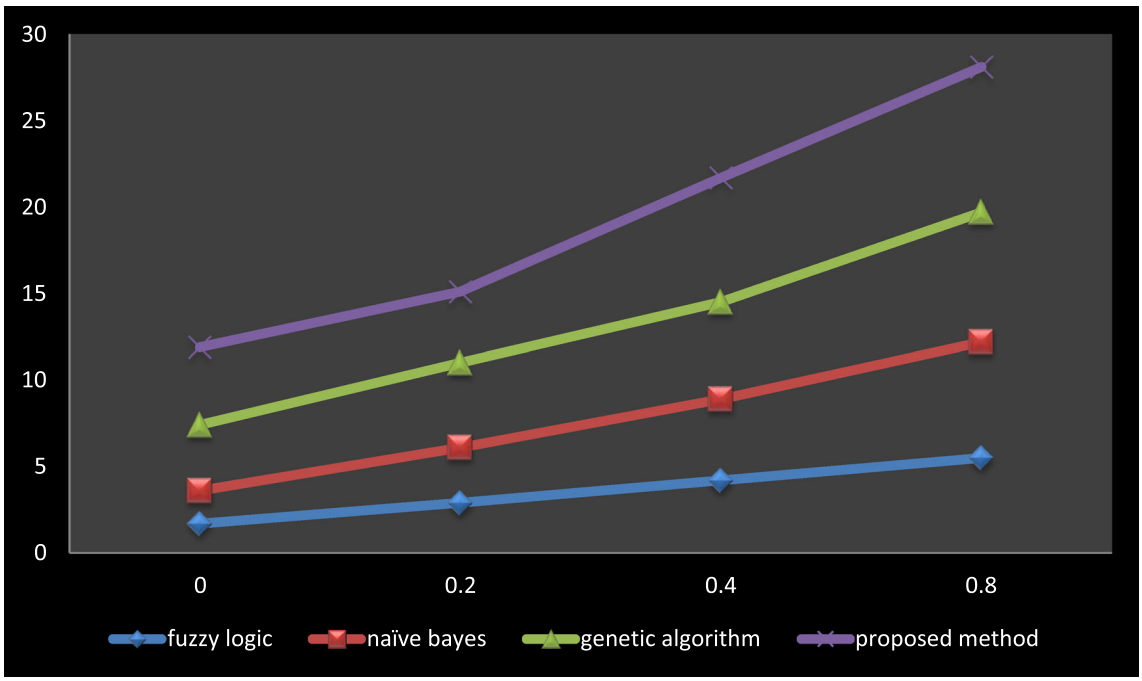


Fig. 9 phenotype performance

proposed phenotype technique finally the cancer could be diagnosed based on the big data platform. First collect all the patient records into that the raw EHR and EMR are classified with the help of phenotype algorithm. Help of TLP tool the phenotype algorithm classifies all the cancer patient medical and health records. The cancer patient details are collect and store into separate database and level of patient symptoms are calculated. Based on the factorized model the patient fatigue, pain, sleep disturbance, Eigen values are calculated. The three factor mean scores are calculated by the disease stage, pain status, chemotherapy and psychological status. Based on these three factorized model and three factors mean score values phenotype technique diagnose the cancer.

Three factorized model mean score values of the proposed system is shown in Fig. 8. The above graph shows the comparison of proposed method with Mean, Standard Deviation, and time.

The performance of phenotype achieves a better performance comparing the all other existing and proposed systems. In this Figs. 8 and 9, the proposed method is compared with the existing techniques such as SVM, xgboost, adaboost, deepboost [1], fuzzy logic, naïve bayes, genetic algorithm [17]. Compared to the other existing techniques, the proposed method produces better performance.

Conclusion

This paper focused on big data analytics for cancer diagnoses based on the proposed phenotype technique. Big data analytics it uses Hadoop which plays a successful role in performing meaningful real time analysis on the large volume of data sets. It able to predict the unexpected situations before it happens. It mainly focuses the accurate prediction of diseases using both structure, semi-structure and unstructured data sets. Also, the proposed work uses HDFS and two phase map reduce. First it collects all the patient data in each mapping process after that phenotype algorithm handles and classifies all the EHR and EMR. The phenotype is further accelerated by adding TLP technique. The accomplishment of this scheme enables both phenotype and TLP tools to be applied on each data sets of three factorized model. Also, this paper analyze three factorized mean score values. The values are calculated by the disease stage, pain status, chemotherapy and psychological status. Finally, compare all the big health care datasets and three factorized mean score values cancer could be diagnosed. The simulation result shows that the proposed system produces higher performance compared to the other existing technique. The scope of the future work is introducing big data algorithms to other health care applications.

Compliance with ethical standards

Conflict of interest The authors have no conflict of interests and the paper has not been submitted to any other Journals.

Research involving human participants and/or animals This article does not contain any studies with human participants or animals performed by any of the authors.

Informed consent It is not required as the dataset is taken online databases.

References

1. Turki, T., An empirical study of machine learning algorithms for cancer identification, In IEEE 15th International Conference on Networking, Sensing and Control (ICNSC) ,pp. 1–5, 2018.
2. Mosquera-Lopez, C., Agaian, S., Velez-Hoyos, A., and Thompson, I., Computer-aided prostate cancer diagnosis from digitized histopathology: a review on texture-based systems. IEEE reviews in biomedical engineering 8:98–113, 2014.
3. Martin, M. E., Wabuyele, M. B., Chen, K., Kasili, P., Panjehpour, M., Phan, M., Overholt, B., Cunningham, G., Wilson, D., DeNovo, R. C., and Vo-Dinh, T., Development of an advanced hyperspectral imaging (HSI) system with applications for cancer detection. Annals of biomedical engineering 34(6):1061–1068, 2006.
4. Korupally, V. R., and Pinnamaneni, S. R., Bigdata analytics for diagnosis and prognosis of cancer using genetic algorithm. International Journal of Computer Science and Information Technologies (IJCSIT) 7(3):1251–1253, 2016.
5. Hajeer, M. H., and Dasgupta, D., Handling big data using a data-aware HDFS and evolutionary clustering technique, IEEE Transactions on Big Data. IEEE Transactions on Big Data 5(2): 134–147, 2017.
6. Triguero, I., Galar, M., Vluymans, S., Cornelis, C., Bustince, H., Herrera, F. and Saeys, Y., Evolutionary undersampling for imbalanced big data classification, In IEEE Congress on Evolutionary Computation (CEC), pp. 715–722, 2015.
7. Aledhari, M., Di Pierro, M., Hefeida, M. and Saeed, F., A deep learning-based data minimization algorithm for fast and secure transfer of big genomic datasets, IEEE Transactions on Big Data, pp.1–13, 2018.
8. García, S., and Herrera, F., Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy. Evolutionary computation 17(3):275–306, 2009.
9. Saeed, F., Big data proteogenomics and high performance computing: Challenges and opportunities, In IEEE Global Conference on Signal and Information Processing (GlobalSIP) , pp. 141–145, 2015.
10. Manogaran, G., Vijayakumar, V., Varatharajan, R., Kumar, P. M., Sundarasekar, R., and Hsu, C. H., Machine learning based big data processing framework for cancer diagnosis using hidden Markov model and GM clustering. Wireless personal communications 102(3):2099–2116, 2018.
11. Sun, J. and Reddy, C.K., Big data analytics for healthcare. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining , pp. 1525–1525, 2013.
12. Rodrigues, Jr., J. F., Paulovich, F. V., de Oliveira, M. C., and de Oliveira, Jr., O. N., On the convergence of nanotechnology and Big Data analysis for computer-aided diagnosis. Nanomedicine 11(8): 959–982, 2016.
13. Mo, H., Thompson, W. K., Rasmussen, L. V., Pacheco, J. A., Jiang, G., Kiefer, R., Zhu, Q., Xu, J., Montague, E., Carrell, D. S., and

Lingren, T., Desiderata for computable representations of electronic health records-driven phenotype algorithms. Journal of the American Medical Informatics Association 22(6):1220–1230, 2015.

14. McCarty, C. A., Chisholm, R. L., Chute, C. G., Kullo, I. J., Jarvik, G. P., Larson, E. B., Li, R., Masys, D. R., Ritchie, M. D., Roden, D. M., and Struewing, J. P., The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC medical genomics 4(1):1–13, 2011.
15. Pendergrass, S. A., Brown-Gentry, K., Dudek, S. M., Torstenson, E. S., Ambite, J. L., Avery, C. L., Buyske, S., Cai, C., Fesinmeyer, M. D., Haiman, C., and Heiss, G., The use of phenome-wide association studies (PheWAS) for exploration of novel genotype-phenotype relationships and pleiotropy discovery. Genetic epidemiology 35(5):410–422, 2011.
16. Milovic, B., Prediction and decision making in health care using data mining, Kuwait chapter of arabian journal of business and management review, vol.33, no.848, pp.1–11, 2012.
17. Cruz, J.A. and Wishart, D.S., Applications of machine learning in cancer prediction and prognosis, Cancer informatics, 2, p.117693510600200030, 2006.