Journal of Science and Technology
ISSN: 2456-5660 Volume 5, Issue 4, July-August 2020

www.jst.org.in *DOI: https://doi.org/10.46243/jst.2020.v5.i4.pp348-360*

Challenges and Trends in Clinical Data Analytics

Shweta S.Kaddi[1], Malini M.Patil[2]

[1](Dept.Of Computer Science and Engineering, JSS Academy of Technical Education,Bengaluru-60/Visveshwaraya Technological University,Belgaum-590018, Karnataka, India)
[2](Dept.Of Information Science and Engineering, JSS Academy of Technical Education, Bengaluru-60/ Visveshwaraya Technological University, Belgaum-590018, India)
[1]Corresponding Author:shwetakaddi@gmail.com

To Cite this Article
Shweta S.Kaddi and Malini M.Patil, "Challenges and Trends in Clinical Data Analytics", Journal of Science and Technology, Vol. 05, Issue 04, July-August 2020, pp348-360

Article Info
Received: 25-04-2020 Revised: 10-07-2020 Accepted: 14-07-2020 Published: 24-07-2020

Abstract:*Today's technological advancements facilitated the researcher in collecting and organizing various forms of healthcare data. Data is an integral part of health care analytics. Drug discovery for clinical data analytics forms an important breakthrough work in terms of computational approaches in health care systems. On the other hand, healthcare analysis provides better value for money. The health care data management is very challenging as 80% of the data is unstructured as it includes handwritten documents, images; computer-generated clinical reports such as MRI, ECG, city scan, etc. The paper aims at providing a summary of work carried out by scientists and researchers who worked in health care domains. More precisely the work focuses on clinical data analysis for the period 2013 to 2019. The organization of the work carried out is specifically with concerned to data sets, Techniques, and Methods used, Tools adopted, Key Findings in clinical data analysis. The overall objective is to identify the current challenges, trends, and gaps in clinical data analysis. The pathway of the work is focused on carrying out on the bibliometric survey and summarization of the key findings in a novel way.*

Key words:*Data mining, Clinical data, Disease diagnosis, Health care data analytics, unstructured data.*

I. INTRODUCTION

Now a day's clinical data analytics (CDA) has become a buzz word in the field of Bio-informatics. The clinical data includes patient personal information, insurance records, doctor's prescriptions, nurse's description, electronic medical records, electronic health records, x-ray images, all types of scanned images, diagnosis records, etc. By analysis of these data, we can get some information that is very useful for some pharmacy companies, hospitals, scientists, doctors, etc. By CDA we can improve the quality of health care. CDA received a lot of attention from researchers in the past decade. CDA is a special domain-specific, initiative as it includes very sensitive issues like privacy, heterogeneity, security, financial, legal issues, etc. Applications of CDA are to forecast the future of many diseases, medicines, gene expressions, patients' health, etc. The purpose of CDA is to mine the useful information from the collected data by using the different steps of knowledge data discovery (KDD).

Clinical data mining means the extraction of useful information, analyze the information, creating a useful medical model for the required purpose from the existing clinical data. CDA can be carried out by using core data mining tasks such as clustering, association, classification, and anomaly detection. Before the application of these tasks, one has to collect the data, the data collection in CDA is a very challenging task. The first challenge is the regulatory action where some countries have the law to not disclose any patient information to the third party (Data Privacy). The second challenge is the technology as the data collected in the health industry will be in the different form some hospitals collect the patient data in the form of papers, whereas some will collect it in the digital form, whereas some data is generated through the scanning machines. The third challenge is the organizational rules; some hospitals have their own rules to make use of the data the analyzer has to follow up those rules. The fourth challenge is the data types are changing very rapidly accordingly to the data handling techniques.

Once the data collection phase is over the next step is the data preprocessing phase. Data preprocessing is one of the important steps in the CDA. Because the collected data is from different resources such as various modern medical devices such as different varieties of scan machines, wearable devices, health monitoring devices, doctor handwritten prescription notes, dietician notes, discharge summary, insurance reports, various types of test reports, etc. The gathered data may consist of some missing data; it may incomplete data, may contain noise in it, and maybe inconsistent. The collected data is not in a structured form. As it is the combination of both structured and unstructured format, one has to prepare it according to their needs by applying different pre-processing techniques such as data cleaning, data integration and transformation, data reduction, data discretization, and concept hierarchy generation.

By clinical data mining, we can predict medical emergencies based on previous experiences and current conditions. Because of many diseases, it's required to predict the future situation. The increased demand for health services is also one of the reasons for the growth of research in CDA. It is also possible to detect fraud with insurance agencies, efficient utilization of resources of the hospital, for good diagnosis and treatment, improved drug reactions, earlier detection of diseases and to improve the customer relationship management using CDA. It is also possible to predict the patient's post-treatment such as at what time he has to readmit or can guess the medicine according to the upcoming situation. CDA helps to answer questions like who, where, and when related to the different areas of the medical field. Also, with the CDA one can take a precautionary measurement, can involve in finding a new type of tools/drugs, increase the precision of drug dosage. The CDA benefits to manage the chronic disease, it provides the care anywhere at any time, makes each day routine life easy like scheduling the appointments with doctors. With the aid of CDA, it is possible to design personalized medicine and to create a clinical decision support system.

II. Data sets for CDA

CDA has witnessed many types of datasets from the last decade as the researcher uses several kinds of datasets according to their requirements from the literature survey, it is found that data sets can be classified as the benchmark and real. Table 1 depicts the benchmark datasets available in the literature survey conducted from the year 2013 to 2019. Few other benchmark data sets are listed in [160] and are based upon clinical and Epidemiological characteristics.

Other than benchmark data sets there are many real-time datasets found in the literature. Some of the datasets identified from the year 2013 to 2019 are described in this section. In papers [39,42,47] the authors have described the real-time hospital and lab data which is a combination of structured and unstructured. In [93,98,101] the authors referred the hospital data for analyzing the health care data using big data, for predicting the sickness, and for analyzing the breast cancer data analytics respectively. In papers [106,107,109] the authors have created a decision support system for predicting the heart failure rate, discussion on early detection and prediction of cancer using the above data, and for improvement of the performance of the machine learning (ML) classifiers used diagnosis of breast cancer. In the article [30] the hospital data is utilized for the comparison of different ML algorithms for disease prediction are used. The authors in [18, 57] used the Electrocardiogram (ECG) as a data set for analyzing the health care big data with a future health condition, In [16, 89] authors used Medical Resonance Images (MRI) for the discussion on deep learning to implement medical 4.0 tool and medical data preparation ML approaches are used. The doctor consultation data were used in [12, 15, 33, 35, 54, 70, and 102] for disease diagnoses and analyses KNN and deep learning methods. The authors mentioned The Electronic Health Record (EHR) in [6, 11, 14, 37, and 102] used to mine health care data using machine learning. In [27, 50,115] the authors used Surveillance, Epidemiology, and End Results (SEER) data set for prediction, classification, and also for performance evolution of different data mining techniques of lung cancer. In [16, 18, and 57] authors specified the Electroencephalography (EEG), Electromyography (EMG) data to apply the big data techniques for prediction of future health. The data were collected from National Claims History, Durable Medical Equipment, outpatient claims for classifying lung cancer with ensemble machine learning techniques in [27]. In [61] data was collected from several institutes for a review on different clinical databases like Cornet (Patient-Centered Clinical Research Network), Open NHS(National Health Services), eICU—Philips, VistA(Veterans Health Information Systems and Technology Architecture) NSQUIP(National Surgical Quality Improvement Project). The author used some of the hospital data such as the Cleveland Clinic Foundation (CCF), Hungarian Institute of Cardiology (HIC).

Beach Medical Center (LBMC) and University Hospital in Switzerland (SUH) in [76] for analyzing Coronary Heart Disease Using Ensemble Machine Learning. The author in [29] used the National Ovarian Cancer Early Detection Program (NOCEDP) and gynecologic oncology clinic at Northwestern University (Chicago, IL, USA) data set to compare performance analysis of different ML algorithms. In [34] the author talked about the National Health and Nutrition Examination Survey (NHANES) data to Predict Diabetes Using Ensemble Perceptron Algorithm. To predict and recognize the disease using a hybrid artificial neural network and decision tree through human blood cells the author used the data collected by Human blood detecting and counting sensor in [38].The data identified by the Korea National Health and Nutrition Examination Survey (KNHANES-VI) for heart disease risk prediction using feature correlation in [44]. In [45] the author specified the baseline data from the Longitudinal Study of Adult Health (ELSA-Brasil) for a Comparison of ML algorithms to build a predictive model for detecting undiagnosed diabetes accuracy study. The author in [64, 85] named StatLog Heart Disease data set to predict heart disease, and in [65] Parkinson's Progression Markers Initiative (PPMI) database was used for prediction of Parkinson's disease. To analyze the performance of different classification techniques to predict diabetes, the author focused The Canadian Primary Care Sentinel Surveillance Network (CPCSSN) database in [78]. In [80] the author used Truven Health Analytics data as claims data. In [90, 97, and 99] the authors used specified Entity-Attribute-Value (EAV) to develop particular medical tool MLBCD, MDL Drug Data Report (MDDR) for drug discovery, the open dataset of Chinese was from China Health and Nutrition Survey (CHNS) to diagnose diabetes using ensemble technique respectively. The authors in [102] specified clinical decision support systems (CDSS) as a data source for health data current perspectives, challenges, and potential solutions**.** In [7] the author used Magnetic Resonance Imaging (MRI) data for detecting brain cancer. For modeling clinical data, the authors associated with [10] used Case Report Format (CRF) data. In [16] the author lighted on Electronic Data Capture (EDC) to model the medical data for deep learning. HL7 virtual medical record (VMR) was used in [103] to create a clinical decision support system. In [4,104] the author quoted on Electronic Medical Record (EMR) data to create a smart disease progress model. The authors in [161] used the national institute of diabetes and digestive and kidney disease data set for detecting early diabetes.

Table 1. Different Data sets used in CDA

Name of the data set	Usability	Contributed by	Referred by
Wisconsin Diagnostic Breast Cancer (WDBC) dataset	To detect breast cancer, Diagnosis of breast cancer, to predict breast cancer	Dr. William H. Wolberg, W. Nick Street, Olvi L. Mangasarian	2019:[166], 2018: [1,49], 2017: [28] 2016: [63,64,75,82], 2015: [152,154] 2014: [153], 2013: [114,151]
University of California Irvine (UCI) heart disease dataset also known as the Cleveland database	For predicting heart diseases, To predict cardiovascular diseases.	Andras Janosi, William Steinbrunn, Matthias Pfisterer, Robert Detrano	2019:[163], 2018: [3,9] 2017:[15,20,28,30,31,36,40,48,53,54,55,157] 2016: [62-64,66,71-73, 76,77,79,81,82,84,86,88,158] 2015: [95,159],2014: [109,112]. 2013: [136]
PIDD (Pima Indian Diabetes Data Set)	For Prediction of Diabetes Mellitus, to identify diabetes,	Peter Turney, National Institute of Diabetes and Digestive and Kidney Diseases	2019:[164][165], 2018:[8] 2017: [15,20,21,32,49,51,52] 2016: [56,61,120], 2015: [94,95] 2014: [138], 2013: [137]
MIMIC II-data harnessed in	To study of patient characteristics, survival rate of patient.	From the ICUs of Beth Israel Deaconess Medical Center from 2001 to 2008	2018: [121], 2017: [119,122,123] 2016: [56,61,120], 2015: [89] 2014: [105,118], 2013: [117]
Open Access Series Imaging Studies (OASIS) is a public data base	For creating decision model of Alzheimer's disease	The Washington University Alzheimer's Disease Research Center, Dr. Randy Buckner at the Howard Hughes Medical Institute (HHMI)	2018: [145] 2017: [17] 2016: [144]
Kent Ridge Biomedical Data Repository	For analyzing cancer related issues.	Online repository	2017: [19], 2016: [142], 2015: [124] 2014: [139,140,141]
Structural Classification of Proteins (SCOP)-	To analyze protein related topics.	Founded in 1994 by Alexey G. Murzin, Steven E. Brenner, Tim J. P. Hubbard, and Cyrus Chothia	2017: [25,126], 2016: [125,147] 2014: [148], 2013: [129]
Breast Cancer Digital Repository (BCDR)- Portuguese Breast Cancer database	For breast cancer related issue analysis	"Hospital São João", University of Porto, Portugal (FMUP)	2018: [131] 2017: [155,156] 2016: [143] 2014: [109,130]
Digital Database for Screening Mammography (DDSM)-	Images for the purpose of breast related issues.	D. Kopans, and R. Moore	2018: [133,135], 2017: [134] 2016: [132], 2014: [109]
FDA-NCI Centre Database	Cancer related database	FDA-NCI	2017: [29,[50]
Lung Nodule Analysis(LUNA16)	Lung base database which is used for analysis of lung related issues.	-	2017: [41,127,128,146]
Kaggles	Contains all type of data	KAGGLE	2017: [41]

Gene expression data	For the purpose of study of genes.	T.R. Golub, et. al	2018: [1], 2017: [14],[16], [22], [37],[46] 2016: [74], 2015: [89], [96]
METABRIC	It is used to classify breast cancer into subcategories.	-	2017: [22]
The Cancer Genome Atlas—TCGA portal	Used for studying the diagnose, treat and prevent cancer disease	-	2019:[162]

III. Techniques and Methods used in CDA

Data analysis mainly includes predictive (regression and classification techniques) and descriptive tasks(association and clustering techniques). Classification is a method of assigning the label for each instance of a data, from which it could predict the particular class. Association rule mining is another technique in data mining; it takes the correlation between the two items to find out the interesting pattern in a large database. The association rule mining is used in CDA to predict several diseases. Clustering is the technique of data mining, cluster means a group of items that are similar to one another, used to group similar items in a dataset. The association between two objects is highest if they belong to the same cluster. One more method in data mining is Outlier detection, in this, if the item is not matched with any of the patterns then it is called anomaly or noise or exception. The summary of all the types of data mining techniques used in CDA is presented in table 2.

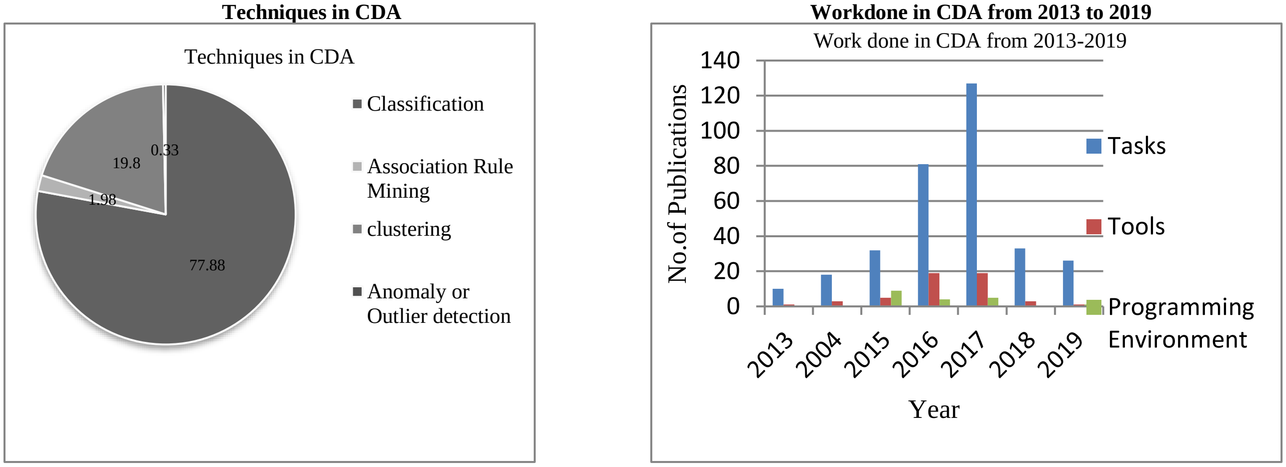
Table 2: Tasks in CDA

Techniques	**Published Papers**
Classification	
Logistic regression(LR)	[1][6][19][22][35][39][45][51][56][65] [91][100][164][165]
Naïve Bayes(NB)	[7][8][9][15][19][30][33][35][37][45][48][49][51][63][65][66][77][82][83][91][95][96][97][98][101][108][114] [163][164][165][166]
Support Vector Machine(SVM)	[1][2][3][4][6][7][8][9][11][15][18][22][26][28][29][30][89][91][95][96][97][98][99][106][109][110][113][31][32][39][42][46][48][50][51][62][64][65][68][71][73][81][82][83][84][87][88][163][164][165][166]
Decision Tree-Based(DTB)	[1][3][4][15][19][21][22][28][29][30][31][31][32][37][38][48][49][50][52][55][56][66][72][75][78][81][85][86][95][97][107][108][95][97][107][108] [113][115][164][165][166]
K-Nearest Neighbors(KNN)	[3][9][28][29][33][37][42][45][47][48][49][54][56][64][77][82][83][96][97][107][109][162] [163][165][166]
Random Forest(RF)	[3] [6] [15][19][22][35][45][47][50][51][53][55][60][65][69][71][72][73][94] [96] [114][161][163][165]
J48	[15][21][30][49][51][66][72][75][77][78][85][86] [98] [108] [163]
C 4.5	[15][19][35][40][52][73][82][88][94][99] [108] [114][115]
Adaboost	[15][34][50][51] [65] [78] [96] [110] [166]
Neural Network(NN)	[19][44] [56][67][68][73] [108] [113][166]
Linear Discriminant Analysis(LDA)	[53] [79][83]
Bayes net(BN)	[15][19] [71] [94]
Genetic algorithm(GA)	[15][19][36] [87]
Radial Basis Function(RBF)	[15] [88]
Multilayer Perception(MLP)	[2][7][9][15][22][29][51][65][68][70][77][88] [94] [98]
Association Rule Mining	
Association analysis: apriori, éclat algorithm, Sequential rule	[10][11][87][112][11] [161]
Distributed data mining algorithm	[102]
Clustering	
K-means	[3] [12][21][22] [87] [161][164]
Estimation Maximization(EM)	[3] [13] [104]
Bayesian method	[1] [11][15][19][43] [65] [71][81] [94] [104]
Artificial Neural Network	[4] [15][22][28][29][30][36][43][45] [66] [84] [97][106] [111][113][115] [161]
Ensemble	[4][8] [27][29][39][50] [76][78]
Bagging	[15] [51] [78][86] [96]
CART	[15][47][52] [69][76] [106] [114]
Convolution Neural Network	[4][7] [14][25][37][41]
Anomaly or Outlier detection	[75]

Few of the other classification techniques found in the literature are presented as follows: The regression-based algorithms are Linear regression/Softmax Regression [2,50,], Statistical learning[74] and Principal Component analysis[81], Functional trees[15], Kstar/ ID3[15], Quadratic Discriminate Analysis [22], Generalized Linear Model[83], Zero R [77]. The neural network-based algorithms are Fuzzy KNN (FKNN) [31][64], Ensemble Perceptron Algorithm (EPA) [34], Probabilistic Neural Network (PNN) [42], Multivariate Adaptive Regression Splines (MARS)/ Tree-Model from Genetic Algorithm (TMGA) [55], Deep learning [11][37]. The survey reviews that the usage of these techniques is found to be is very sparse. Some of the other clustering algorithms are found in the literature survey are as follows: SELDOM (ensemble of Dynamic logic-based Models) [1], PAM(Partition around medoids, K-medoids, ROCK(robust clustering algorithm), DBSCAN(density-based clustering algorithm) [3], Kernel methods, Joint Clustering And Classification (JCC)[6], Trajectory clustering[11], Fuzzy logic[15], Adaptive Neuro-Fuzzy Inference System (ANFIS)[70], Farthest First clustering[75]. The survey analysis tells that the adoptions of these algorithms are very less. The analysis of these techniques is as shown in fig 1. It is found that very sparse literature is found in outlier detection.

IV. Tools and Programming Environments adopted in CDA

Clinical data is a combination of both structured and unstructured; to handle such data it requires special kinds of tools. To analyze the clinical data machine learning provided many tools such as R, TensorFlow, TANAGRA, WEKA, MATLAB, SVMLib, Apache Spark, Stata, Jaspersoft Splunk, etc. From the literature survey, it is found that WEKA is best suited for CDA.In paper [3] author using R tool for concluding the efficient technique for predicting heart diseases using K-means, K-modes, K-modes, PAM, CLARANS, CLARA, FCM, Ward's, ROCK, DBSCAN, OPTICS, EM, SVM, Decision tree, Random Forest and Knn. Authors in [28] discussed a comparison study on performance measurements of various machine learning techniques such as SVM, Decision Tree, Random forest, and KNN using R tool for breast cancer classification. The author in [32] depicted the prediction of diabetes in women using PIMA dataset by applying SVM using the R tool. In [55, 84] authors are using R platform prediction of heart disease using machine learning like SVM, decision tree, and KNN. Authors in [91] discussed the applications of machine learning to forecast postoperative complications. In [102] writer discussed health data analytics using a big data approach in the current perspective, challenges, and solutions also their authors used the R tool to overcome the weakness of the Hadoop. TensorFlow tool is used for building ML model using a deep learning method. This has used in [7, 20] to detect brain cancer and to identify diabetes in respective papers. TANAGRA has used in [9] to improve prediction of the heart attack using a feature selection method.Statistics of the work done in CDA from 2013-2019 is presented in figure 1 and briefed under the following headings.


Techniques in CDA

Workdone in CDA from 2013 to 2019

Fig 1 Techniques in CDA and Workdone in CDA from 2013 to 2019

WEKA is the most popular tool in the healthcare data analysis field. For more different healthcare applications this tool has been using still today because it provides all kinds of tools required for data mining such as data preprocessing, regression, classification, and visualization. Just in one word its collection of algorithms. The authors in [15,77] surveyed to diagnose the diseases using machine learning; they found this tool as the best tool and also in

[30] to make a comparative study of different machine learning algorithms for the prediction of diseases they considered the WEKA as one of the tools also in [16,59] authors talked on a tool called as Medical 4.0 and PredictT-ML tool for preprocessing the medical data which is ready to use by deep learning technique. The authors in [21, 35, 47, 49, 51, 62, 78, 94, 95, 98] discussed on diabetes such as prediction, diagnosis, and performances analysis of different machine learning algorithms for prediction of diabetes, in [48, 72, 88] authors discussed on prediction and diagnosis of heart-related diseases, the discussion on detection of breast cancer also carried out in [71, 75, 82, and 96,109,115]. This tool has been used in [53, 69, and 79] for detecting thyroid disease.WEKA was also utilized for outlier detection in [63, 75]. In the paper [66] authors discussed the liver disorder detection in [68,81] general approaches using machine algorithms for health care industry issues discussed.

MATLAB supports a large set of algorithms and visualization tools, because of these features in health care data mining MATLAB is widely used namely in [15] for diagnosing various diseases, in [70] for prediction of diabetes, in [84] the prediction of heart disease. RapidMiner is having a very rich and powerful graphical user interface that is well suited for the predictive analysis. The authors in [15] used RapidMiner to diagnose various diseases.Caffe is a deep learning framework. In [16] the author used the Caffe tool for designing a Medical 4.0 tool for preparing medical data for machine learning purposes. KEEL (Knowledge Extraction based on Evolutionary Learning) is an open software tool that can be used for many data mining tasks such as regression, classification, etc. In [88] KEEL was used for efficient prediction of heart diseases using machine learning techniques. STATA is a basic command-oriented package, which is most suitable for analysis, graphical visualization, and data management. It is user friendly and has many library functions. It provides a wide range of operations supported by data mining such as regression modeling, cluster analysis, multivariate methods, ANOVA (analysis of variance), etc. The authors in [105] discussed the handling of the clinical data from the perspective of big data their STATA has been used for the analysis of clinical data. Multi-dimensional Text Analytics Tool (MUTATO) is using for text analysis. In [24] MUTATO was used for the study of the supervised machine learning of the online forms for diabetes patients.

Table 3: Tools in CDA

Tool name	**About Tool**	**Referenced by**
R	Open source Developed by Bell lab,By Ross Ihaka and Robert Gentleman in 1993	[3] [28] [32] [55] [84] [91] [102]
Tensor Flow	Open source, developed by Google Brain Team in 2015.	[7][20]
TANAGRA	Free suite, by Ricco Rakotomalala in June 2003	[9]
WEKA	Free software, Waikato Environment for Knowledge Analysis (Weka),developed at University of Waikato, New Zealand at 1993	[15] [16] [21] [30] [35] [47] [48] [49] [51] [53] [59] [62] [63] [66] [68] [69] [71] [72] [75] [77] [78] [79] [81] [82] [88] [94] [95] [96] [98] [109] [115]
MATLAB	matrix laboratory, by Cleve Moler at University of New Mexico, in 1970	[15] [70] [84]
RAPIDMINER	By Ralf Klinkenberg, Ingo Mierswa, and Simon Fischer, in 2001	[15]
Caffe	Open source, Convolution Architecture for Fast Feature Embedding, developed at UC Berkeley, by Yangqing Jia	[16]
KEEL	Open source tool, Knowledge Extraction based on Evolutionary Learning,	[88]
Stata statistical package	In 1985 by StataCorp,	[105]
Multi-dimensional TextAnalytics Tool (MUTATO)	-	[24]
Jupyter Note book	Open tool evolved in 2014	[164]

Table 4: Programming Environment

Programming environment	**About Tool**	**Referenced by**
Apache Singa	In 2014,	[16]
Torch	Open source, by Ronan Collobert, KorayKavukcuoglu, Clement Farabet, in October 2002	[16]
Theano	Theano is a Python library that lets you to define, optimize, and evaluate mathematical expressions, especially ones with multi-dimensional arrays (numpy.ndarray)	[16]
LabKey	Apache license, atFred Hutchinson Cancer Research Center by labkey	[16]
SVMLib	Open source, by Chih-Chung Chang and Chih-Jen Lin	[16] [62]
Apache Mahout	Apache Software Foundation,	[58]
Apache Spark	Open source, By MateiZaharia started at 2009, clustering computing framework.	[58] [59]
MongoDB	Free downloading, used for Classification Discovery, Cluster Discovery, Regression Discovery, Association Discovery, Data Visualizations, Discovery Visualization	[89]

Hadoop	Open source Started at 2007, Doug Cutting and Mike Cafarella,	[89]
Hive	Apache license, its data warehouse	[92]
Vertica	In 2005 by Michael Stonebraker and Andrew Palmer,	[92]
JAQL	Open source, json query language	[92]
AVRO	Open source, by by Doug Cutting,	[92]
IbmSpss 20.0.	By IBM Corporation in 1968 by Norman H. Nie, Dale H. Bent, and C. Hadlai Hull,	[93]
GoMiner		[89]

V. Key Findings in CDA

The major key observations of the literature work carried out in the present work is summarized asfollows.

1**. Data Privacy:** Maintaining data privacy is one of the main challenging issues in CDA. The collected data has to be saved in a way that the information regarding the patient should not be revealed

2. **Heterogeneity of data:** The data sources required for CDA are heterogeneous, involve many different kinds of data which are generated by several resources as discussed in the above topics. It is not so easy to handle such data because of such a reason it requires a special technique. Handling the heterogeneous data is only a big challenging issue.

3. **Technological support:** To handle the heterogeneous data, to maintain privacy in a collected data, for pre-processing a data, etc. The CDA requires special kinds of technical support. But to handle the big data there are already many kinds of tools are available in the market but those are not exactly suitable for CDA.

4. **Societal issues:** By analyzing the clinical data it is possible to predict the future regarding the health-related issues of patients which is very much required in the present lifestyle of people. According to this prediction, one can easily maintain their health and diet. Also, pharmacy companies can also produce the required medicine for a specific area.

5. Scalability of data: In CDA the data plays a very vital role as already we had seen. The reason behind the scalability of data is the data collection techniques differ along with the technology. Due to the usages of different types of instruments for disease diagnosis, for maintaining the patient's records, for various lab tests, etc.

6. Tasks and tools: The main key finding in clinical data analysis is the tasks and tools, as above we had seen that there are many different techniques in data mining but in the field of CDA the classification has been used by many authors, there still scope to make utilization of the clustering, association rule mining and detection of outlier data. Similarly, the WEKA tool has used widely in several papers. But in current days there are many advanced tools emerging in a market it has to check whether those tools are suitable for the CDA or not has to be check.

VI. Conclusion

The exhaustive literature survey finds that in CDA the data privacy plays a very important role as in many countries the hospitals collect the data from the patients and that data should be maintained with high privacy because they have abounded with some law issues like HIPPA Patient Safety and Quality Improvement Act (PSQIA) HITECH Act, (PIPEDA) Personal Information Protection and Electronic Documents Act, Russian Federal Law on Personal Data, etc. So it’s very important to maintain data privacy in CDA. Data preparation is the main task in data analysis. As in CDA, the data is collected by different sources and is unstructured. Data preparation for further data analytics is a major step.So, it’s required to apply some of the pre-processing techniques for the preparation of clinical data. Digitalization in the medical field resulted in storing the voluminous data and can be handled using big data approaches. It is required to apply the concepts and techniques of big data but with some modification due to the nature of the clinical data. As discussed, the clinical data is a special type of data and it requires special types of tools to handle.

VII. Future Scope of the work

The purpose of the bibliometric survey conducted is to find the gaps in CDA. It is observed from the literature survey that though data mining techniques play a prominent role, but the **anomaly detection** method is less attended by the researchers for the survey period. Very sparse literature is found. Future work is more concentrated to explore the possible method in anomaly detection in the clinical data. It is necessary to detect the anomaly in clinical data sets so that the errors in the diagnoses are detected and treated fast.

References

1. RoslavaCuperlovic-Culf,Machine Learning Methods for Analysis of Metabolic Data and Metabolic Pathway Modeling, Metabolites 2018, 8, 4; doi:10.3390/metabo8010004.
2. Abien Fred M. Agarap ,On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset. https://doi.org/10.1145/3184066.3184080.
3. M. Chandralekha_ and N. Shenbagavadivu,Performance Analysis Of Various Machine Learning Techniques To Predic Cardiovascular Disease: An Emprical Study, Applied Mathematics & Information Sciences ,An International Journal, Appl. Math. Inf. Sci. 12, No. 1, 217-226 (2018).
4. Min Chen, Jun Yang, Jiehan Zhou, Yixue Hao, Jing Zhang, and Chan-Hyun Youn ,5G-Smart Diabetes:Toward Personalized Diabetes Diagnosis with Healthcare Big Data Clouds, IEEE Communications Magazine, April 2018.
5. PooyaMobadersanya, SafooraYousefia, Mohamed Amgada, David A. Gutmanb, Jill S. Barnholtz-Sloanc,José E. Velázquez Vega, Daniel J. Brat, and Lee A. D. Cooper,Predicting cancer outcomes from histology and genomics using convolution networks, www.pnas.org/cgi/doi/10.1073/pnas.1717139115.
6. Theodora S. Brisimi, Tingting Xu, Taiyao Wang, Wuyang Dai, William G. Adams , and Ioannis Ch. Paschalidis , Predicting Chronic Disease Hospitalizations from Electronic Health Records:An Interpretable Classification Approach.
7. Aaswad Sawant, Mayur Bhandari, Ravikumar Yadav, Rohan Yele, Sneha Kolhe,Techniques of brain cancer detection from mri using machine learning,International Research Journal of Engineering and Technology, Vol-05, Issue: 01, Jan-2018.
8. MinyechilAlehegn, Rahul Joshi& Dr. PreetiMulay,Analysis and Prediction of Diabetes Mellitus using Machine Learning Algorithm, International Journal of Pure and Applied Mathematics, Volume 118 No. 9 2018, 871-878.
9. Hidayet TAKCI, Improvement of heart attack prediction by the feature selection methods, Turkish Journal of Electrical Engineering & Computer Sciences, Turk J Elec Eng& Comp Sci (2018) 26: 1 – 10, Turk J Elec Eng& Comp Sci (2018) 26: 1 – 10.
10. Gajanan Bhat, Shy Kumar,Data Modeling in Clinical Data Analysis Projects,Data Ware housing and Solutions.
11. Pranjul Yadav, Michael Steinbach, Vipin Kumar,Gyorgy Simon, Mining Electronic health record:A Survey, (EHRs) : A Survey ACM Trans. Embedd. Comput. Syst.Vol 1, Issue 1, 2017.
12. Reza Ghodsi,Shiva Bagheri Marani, Abbas Keramati, Med Crave, Application of K-Means Technique in Data Mining to Cluster Hemodialysis Patients, ,International Robotics &Automation Journal**,** Volume 2 Issue 2 – 2017.
13. Uma K, M. Hanumanthappa ,Data Collection Methods and Data Pre-processing Techniques for Healthcare Data Using Data Mining, International Journal of Scientific & Engineering Research Volume 8, Issue 6, June-2017.
14. Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang and Joel T. Dudley, Deep learning for healthcare: review, opportunites and challenges, OXFORD,Briefings in Bioinformatics, 2017, 1–11, doi: 10.1093/bib/bbx044.
15. Meherwar Fatima, Maruf Pasha,Survey Of Machine Learning Algorithms For Disease Diagnostic,Journal of Intelligent Learning Systems and Applications, 2017, 9, 1-16.
16. Natalia Labuda , Tomasz Lepa , Marek Labuda and Karol Kozak,Medical 4.0: Medical Data Ready for Deep and Machine Learning, Labuda et al., J Bioanal Biomed 2017, 9:6 DOI: 10.4172/1948-593X.1000194, Journal of Bioanalysis & Biomedicine.
17. Aymen Ayaz, Muhammad Zubair Ahmad, Khawar Khurshid, Awais M. Kamboh ,MRI based Automated Diagnosis of Alzheimer's: Fusing 3D Wavelet-Features with Clinical Data, 978-1-5090-2809-2.
18. Prasankumarsahoo, Suvendukumarmohapatra, Shih-linwu, Analyzing healthcare big data with prediction for future health condition, digital object identifier 10.1109/access.2016.2647619, volume 4, 2016.
19. Khalid Raza, Atif N Hasan,A Comprehensive Evaluation of Machine Learning Techniques for Cancer Class Prediction Based on Microarray Data.
20. Sushant Ramesh, Ronnie D. CaytilesandN.Ch.S.NIyengar,A Deep Learning Approach to Identify Diabetes, Advanced Science and Technology Letters Vol.145, pp.44-49 http://dx.doi.org/10.14257/astl.2017.145.09.
21. Wenqian Chen, ShuyuChen,HancuiZhang,TianshuWu,A Hybrid Prediction Model for Type 2 Diabetes Using K-Means and Decision Tree,978-1-5386-0497-7/17.
22. Gamal Elkomy, ElSayedSallam, SherinElgokh ,A Stacked Generalization Method for Disease Progression Prediction, 978-1-5386-4266-5/17 ©2017 IEEE.
23. Rohan Bhardwaj, Ankita R. Nambiar, Debojyoti Dutta A Study of Machine Learning in Healthcare, 2017 IEEE 41st Annual Computer Software and Applications Conference, DOI 10.1109/COMPSAC.2017.164.
24. Jonathan-Raphael Reichert, Klaus Langholz Kristensen, Raghava Rao Mukkamala, Ravi Vatrapu ,A Supervised Machine Learning Study of Online Discussion Forums about Type-2 Diabetes, 2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom).
25. Lin Bai,, Lina Yang ,A Unified Deep Learning Model for Protein Structure Prediction, 978-1-5386-2201-8/17©2017 IEEE.
26. Guanjin Wang, Guangquan Zhang, Kup-Sze Choi, Kin-Man Lam, and JieLu,An output-based knowledge transfer approach and its application in bladder cancer prediction, 978-1-5090-6182-2/17/©2017 IEEE.
27. Savannah L. Bergquist, Gabriel A. Brooks, Nancy L. Keating, Mary Beth Landrum, Sherri Rose, Classifying Lung Cancer Severity with Ensemble Machine Learning in Health Care Claims Data, Proceedings of Machine Learning for Healthcare 2017, JMLR W&C Track Volume 68.
28. Arpita Joshi and Dr. Ashish Mehta, Comparative Analysis of Various Machine Learning Techniques for Diagnosis of Breast Cancer, International Journal on Emerging Technologies ,8(1): 522-526(2017).
29. Sharmistha Bhattacharjee, Yumnam Jayanta Singh, Dipankar Ray,Comparative Performance Analysis of Machine Learning Classifiers on Ovarian Cancer Dataset,2017,Third International Conference on research in Computational Intelligence and Communication Network ,978-1-5386-1931-5/17/©2017 IEEE.
30. Anantvir Singh Romana, A Comparative Study of Different Machine Learning Algorithms for Disease Prediction Romana International Journal of Advanced Research in Computer Science and Software Engineering7(7). DOI: 10.23956/ijarcsse/V7I7/0177, pp. 172-175.
31. M.P.Gopinath ,Comparative Study on Classification Algorithm for Thyroid Data Set, International Journal of Pure and Applied Mathematics Volume 117 No. 7, 2017, 53-63.

32. Aakansha Rathore, Simran Chauhan,Detecting and Predicting Diabetes Using Supervised Learning: An Approach towards Better Healthcare for Women,International Journal of Advanced Research in Computer Science, Volume 8, No. 5, May-June 2017.
33. Deeraj Shetty Kishor RitSohail Shaikh Nikita Patil,Diabetes Disease Prediction Using Data Mining, 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS).
34. Roxana Mirshahvalad ,NastaranAsadiZanjani ,Diabetes Prediction Using Ensemble Perceptron Algorithm, 2017 9th International Conference on Computational Intelligence and Communication Networks, 978-1.
35. P. Suresh Kumar and V. UmatejaswiDiagnosing Diabetes using Data Mining Techniques, International Journal of Scientific and Research Publications, Volume 7, Issue 6, June 2017.
36. KaanUyar, Ahmet İlhan , Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks,9th International Conference on Theory and Application of Soft Computing, Computing with Words and Perception, ICSCCW 2017, 22-23 August 2017, Budapest, Hungary, Procedia Computer Science 120 (2017) 588–593.
37. Min Chen, Yixue Hao, Kai Hwang, Lu Wang, And Lin Wang, Disease Prediction by Machine Learning Over Big Data from Healthcare Communities, Doi:10.1109/Access.2017.2694446,Volume 5, 2017.
38. TharahaS,Rashika K,Hybrid Artificial Neural network and Decision Tree algorithm for Disease Recognition and Prediction in Human Blood Cells,International Conference on Innovations in information Embedded and Communication Systems (ICIIECS)2017.
39. Priya Govindarajan, 1Ravichandran KS, Sundararajan S, Sreeja S,Impact of Modifiable and Non-Modifiable Risk Factors on the Prediction of Stroke Disease, InternationalConference on Trends in Electronics and Informatics ICEI 2017.
40. Felix Tamin,Ni Made SatvikaIswari,Implementation of C4.5 Algorithm to Determine Hospital Readmission Rate of Diabetes Patient,4th International Conference on New Media Studies Yogyakarta, Indonesia, November 08-10, 2017.
41. Wafaa Alajwaa, Mohammad Nassef, Amar Badr, Lung Cancer Detection and Classification with 3D Convolutional Neural Network(3D-CNN), International Journal of Advanced Computer science and Applications, Vol.8, On.8,2017.
42. Raid M. Khalil, Adel Al-Jumaily,Machine Learning Based Prediction of Depression among Type 2 Diabetic Patients, 12th International Conference on Intelligent Systems and Knowledge Engineering, 978-1-5386-1829-5/17 ©2017 IEEE.
43. BerinaAli,LejlaGurbeta, AlmirBadnjevi,Machine Learning Techniques for Classification of Diabetes and Cardiovascular Diseases,6th Mediterranean Conference On Embedded Computing,(Meco), 11-15 June 2017, Bar, Montenegro.
44. Jae Kwon Kim and SanggilKang ,Neural Network-Based Coronary Heart Disease Risk Prediction Using Feature Correlation Analysis,Hindawi Journal of Healthcare Engineering Volume 2017, https://doi.org/10.1155/2017/2780501 Article ID 2780501, 13 pages.
45. André Rodrigues Olivera, ValterRoesler, CiranoIochpe, Maria Inês Schmidt, Álvaro Vigo, Sandhi Maria Barreto,BruceBartholow Duncan, Comparison of machine-learning algorithms to build a predictive model for detecting undiagnosed diabetes –ELSA-Brasil: accuracy study, Sao Paulo Med J. 2017;135(3):234-46.
46. Cai Huang, Roman Mezencev, John F. McDonald, Fredrik Vannberg, Open source machine-learning algorithms for the prediction of optimal cancer drug therapies, https://doi.org/10.1371/journal.pone.0186906 October 26, 2017.
47. P. Suresh Kumar.S,Pranavi, Performance Analysis of Machine Learning Algorithms on Diabetes Dataset using Big Data Analytics, International Conference on Infocom Technologies and Unmanned Systems (ICTUS'2017), 2017, ADET, UAE.
48. Sanjay Kumar Sen,Predicting and Diagnosing of Heart Disease Using Machine Learning Algorithms, International Journal Of Engineering And Computer Science, DOI: 10.18535/ijecs/v6i6.14, Volume 6 Issue 6 June 2017, Page No. 21623-21631.
49. Uswa Ali Zia, Dr. Naeem Khan ,Predicting Diabetes in Medical Datasets Using Machine Learning Techniques International Journal of Scientific & Engineering Research Volume 8, Issue 5, May-2017.
50. Chip M. Lynch, BehnazAbdollahi, Joshua D. Fuqua, Alexandra R. de Carlo,James A. Bartholomai, Rayeanne N. Balgemann, Victor H. van Berkel, Hermann B. Frieboes, Prediction of lung cancer patient survival via supervised machine learning classification techniques,International Journal of Medical Informatics 108 (2017) 1–8.
51. Md. Aminul Islam, Nusrat Jahan, Prediction of Onset Diabetes using Machine Learning Techniques,International Journal of Computer Applications (0975 – 8887) Volume 180 – No.5, December 2017.
52. Gauri D. KalyankarShivananda R. Poojara Nagaraj V. Dharwadkar,Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop, International conference on I-SMAC-2017.
53. Ammulu K. Venugopal T. ,Thyroid Data Prediction using Data Classification Algorithm, IJIRST –International Journal for Innovative Research in Science & Technology| Volume 4 | Issue 2 | July 2017.
54. Shahab Tayeb, MatinPirouz, Johann Sun, Kaylee Hall, Andrew Chang, Jessica Li, Connor Song, Apoorva Chauhan, Michael Ferra,Theresa Sager, Justin Zhan Shahram Latifi,Toward Predicting Medical Conditions Using *k*-Nearest Neighbors, 2017 IEEE International Conference on Big Data (BIGDATA),978-1-5386-2715-0/17©2017 IEEE.
55. Tanvi Sharma, Sahil Verma, Kavita,Prediction of Heart Disease Using Cleveland Dataset: A Machine Learning Approach, International Journal of Recent Research Aspects ISSN:2349-7688, Vol. 4, Issue 3.
56. A.Salcedo-Bernal, M.P.Villamil-Giraldo, A.D.Moreno -Barbosa clincial data analysis: an opportunity to compare machine learning algorithm methods, Procedia Computer Science 100 (2016) 731 – 738.
57. Andreas Holzinger, Machine Learning for Health Informatics, Springer International Publishing AG 2016, pp. 1–24, DOI: 10.1007/978-3-319-50478-0 1.
58. D. P. Acharjya, Kauser Ahmed P, A Survey on Big data Analytics: challenges, open research issues and tools, International Journal of Advanced Computer Science and Applications,Vol. 7, No. 2, 2016.
59. Gang Luo,PreditT-ML:A tool for automating machine learning model building with big clinical data.
60. Sohrab Saeb, Luca Lonini , Arun Jayaraman, David C. Mohr, Konrad P. Kording, Voodoo Machine Learning for Clinical Predictions, http://dx.doi.org/10.1101/059774 doi: bioRxiv preprint first posted online Jun. 19, 2016.
61. Jeff Marshall, Abdullah Chahin and Barret Rush,Review of Clinical Databases, MIT Critical Data, Secondary Analysis of Electronic Health Records, http://www.springer.com/978-3-319-43740-8.
62. Anjali Negi,VarunJaiswal,A First Attempt to Develop a Diabetes Prediction Method Based on Different Global Data sets,2016 Fourth International Conference on Parallel ,Distributed and Grid Computing(PDGC).
63. DelshiHowsalya Devi and Dr. M Indra Devi, A Performance Analysis of the Innovative Methods Employed for Outlier Detection using Data Mining Algorithms with Three Different Applications, Advances in Natural and Applied Sciences. 10(9) Special 2016, Pages: 445-455.

64. MehrbakhshNilashi , Othman bin Ibrahim , Hossein Ahmadi , Leila Shahmoradi ,An analytical method for diseases prediction using machine learningtechniques, Computers and Chemical Engineering 106 (2017) 212–223.
65. Kamal Nayan Reddy Challa , Venkata SasankPagolu , Ganapati Panda , Babita Majhi An Improved Approach for Prediction of Parkinson's Disease using Machine Learning Techniques, International conference on Signal Processing, Communication, Power and Embedded System (SCOPES)-2016.
66. Tapas Ranjan Baitharua, Subhendu Kumar Panib ,Analysis of Data Mining Techniques For Healthcare Decision Support System Using Liver Disorder Dataset, International Conference on Computational Modeling and Security (CMS 2016), Procedia Computer Science 85 (2016) 862 – 870.
67. KG Nandha Kumar1, T Christopher,Analysis of liver and diabetes datasets by using unsupervised two-phase neural network techniques, Biomedical Research 2016; Special Issue:S87-S91.
68. Parisa Naraei ,AbdolrezaAbhari , Alireza Sadeghian ,Application of Multilayer Perceptron Neural Networks and Support Vector Machines in Classification of Healthcare Data,Future Technologies Conference 2016 | San Francisco, United States.
69. Ebru Turanoglu-Bekar, GozdeUlutagay, Suzan Kantarc-Savas, Classification of Thyroid Disease by Using Data Mining Models: A Comparison of Decision Tree Algorithms,Oxford Journal of Intelligent Decision and Data Science 2016 No. 2 13-28, Volume 2016, Issue 2, Article ID ojids-00002,doi:10.5899/2016/ojids-00002.
70. Aparimita Swain, Sachi Nandan Mohanty, Ananta Chandra Das,Comparative Risk Analysis On Prediction Of Diabetes Mellitus Using Machine Learning Approach, International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) – 2016.
71. Dana Bazazeh and RaedShubair, Comparative Study of Machine Learning Algorithms for Breast Cancer Detection and Diagnosis, 978-1-5090-5306-3/16,2016 IEEE.
72. Jaymin Patel, Prof.TejalUpadhyay, Dr. Samir Patel ,Heart Disease Prediction Using Machine learning and Data Mining Technique,Volume 7,Number 1Sept 2015-March 2016,pp.129-137,IJCSC.
73. QiaoPan,Yuanyuan Zhang, Min Zuo, Lan Xiang, Dehua Chen,Improved Ensemble Classification Method of Thyroid Diseas Based on Random Forest, 2016 8th International Conference on Information Technology in Medicine and Education, DOI 10.1109/ITME.2016.86.
74. R. Iniesta, D. Stahl and P. McGuffin,Machine learning, statistical learning and the future of biological research in psychiatry, Psychological Medicine,46,2455–2465.©Cambridge University Press 2016, doi:10.1017/S0033291716001367.
75. R DelshiHowsalya Devi, Dr. M Indra Devi,Outlier Detection Algorithm Combined With Decision Tree Classifier For Early Diagnosis Of Breast Cancer,International Journal of Advanced Engineering Technology.
76. Kathleen H. Miao, Julia H. Miao, and George J. Miao, Diagnosing Coronary Heart Disease Using Ensemble Machine Learning, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 10, 2016.
77. A.Swarupa Rani, S.Jyothi,Performance analysis of Classification Algorithms under Different Datasets,2016 International Conference on Computing for Sustainable Global Development (INDIACom).
78. Sajida Perveena,MuhammadShahbaza, Aziz Guergachib, Karim,Keshavjeec,Performance Analysis of Data Mining Classification Techniques to Predict Diabetes, Symposium on Data Mining Applications,2016, 30 March 2016, Procedia Computer Science 82,115 – 121.
79. G. RasithaBanu ,Predicting Thyroid Disease using Linear Discriminant Analysis (LDA) Data Mining Technique, Communications on Applied Electronics (CAE) –Foundation of Computer Science FCS, New York, USA Volume 4– No12, Jan 2016.
80. David A. Kvancz, Marcus N. Sredzinski, Celynda G. Tadlock, Predictive Analytics: A Case Study in Machine-Learning and Claims Databases, The American Journal Of Pharmacy Benefits, Vol. 8, No. 6.
81. Dhomse Kanchan B, Mr. Mahale Kishor M.,Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis, International Conference on Global Trends in Signal Processing, Information Computing and Communication 978-1-5090-0467-6/16©2016 IEEE.
82. Hiba Asri,HajarMousannif,Hassan Al Moatassime,ThomasNoel,Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis, The 6th International Symposium on Frontiers in Ambient and Mobile Systems (FAMS 2016), Procedia Computer Science 83 (201) 1064 – 1069.
83. MadeehNayerAlgedawy , Detecting Diabetes Mellitus using Machine Learning Ensemble, International Journal of Computer Systems, Volume 03– Issue 12,December, 2016 .
84. S.Radhimeenakshi,Classification and Prediction of heart disease risk using data mining techniques of support vector machine and artificial neural network, 978-9-3805-4421- 2/16 @ 2016 IEEE.
85. R. Vijaya Kumar Reddy , K. Prudvi Raju , M. Jogendra Kumar , CH. Sujatha , P. Ravi Prakash, Prediction of Heart Disease Using Decision Tree Approach, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 3, March 2016.
86. Mudasir Manzoor Kirmani, SyedImmamul Ansarullah ,Prediction of Heart Disease using Decision Tree a Data Mining Technique, IJCSN International Journal of Computer Science and Network, Volume 5, Issue 6, December 2016.
87. Era Singh Kajal , Ms. Nishika ,Prediction of Heart Disease using Data Mining Techniques, International Journal of Advance Research , Ideas and Innovations in Technology,Volume2, Issue3.
88. Purushottama,c , Prof. (Dr.) Kanak Saxenab, Richa Sharma,Efficient Heart Disease Prediction System, Procedia Computer Science 85 (2016)962 – 969,
89. Ashwin Belle, Raghuram Thiagarajan, S. M. Reza Soroushmehr,Fatemeh Navidi, Daniel A. Beard, and KayvanNajarian, Big Data Analytics in Healthcare,Hindawi Publishing Corporation BioMed Research International Volume 2015, Article ID 370194, 16 pages.
90. Gang Luo ,MLBCD a machine learning tool for big clinical data, Luo Health Inf Sci Syst (2015) 3:3,DOI 10.1186/s13755-015-0011-0.
91. Paul Thottakkara, TezcanOzrazgat-Baslanti, Bradley B. Hupf, Parisa Rashidi,PanosPardalos, PetarMomcilovic, Azra BihoracApplication of Machine learning Techniques to high dimensional clinical data to forecast postoperative complications, PLOS ONE | DOI:10.1371/journal.pone.0155705 May 27, 2016.
92. Mohammad Ahmad Alkhatib, Amir Talaei-Khoei, Amir Hossein Ghapanchi ,Analysis of Research in Healthcare Data Analytics, Australasian Conference on Information Systems,Sydney,2015.
93. J.Archenaa and E.A.Mary Anita ,A Survey Of Big Data Analytics in Healthcare and Government, 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15), Procedia Computer Science 50 (2015) 408 – 413.
94. Amit kumarDewangan, Pragati Agrawal, Classification of Diabetes Mellitus Using Machine Learning Techniques,International Journal of Engineering and Applied Sciences (IJEAS), Volume-2, Issue-5, May 2015.

95. Veena Vijayan V, Anjali ,Decision Support Systems for Predicting Diabetes Mellitus –A Review, Proceedings of 2015 Global Conference on Communication Technologies(GCCT 2015).
96. Amit Bhola and Arvind Kumar Tiwari,Machine Learning Based Approaches For Cancer Classification Using Gene Expression Data, Machine Learning and Applications: An International Journal (MLAIJ) Vol.2, No.3/4, December 2015.
97. Antonio Lavecchia, Machine-learning approaches in drug discovery: methods and applications, Drug Discovery Today _ Volume 20, Number 3 _ March 2015.
98. Ms. K Sowjanya , Dr. Ayush Singhal , Ms. Chaitali Choudhary ,MobDBTest: A machine learning based system for predicting diabetes risk using mobile devices, 978-1-4799-8047-5/15@2015 IEEE.
99. Longfei Han, Senlin Luo, Jianmin Yu, Limin Pan, and SongjingChen ,Rule Extraction From Support Vector Machines Using Ensemble Learning Approach: An Application for Diagnosis of Diabetes, IEEE Journal Of Biomedical And Health Informatics, Vol. 19, No. 2, March 2015.
100. Vinaytosh Mishra1, Dr. Cherian Samuel2, Prof. S.K.Sharma, Use Of Machine Learning To Predict The Onset Of Diabetes, International Journal of Recent advances in Mechanical Engineering (IJMECH) Vol.4, No.2, May 2015.
101. Santosh Tirunagari, Norman Poh ,Hajara Abdulrahman‡ , Nawal Nemmour§ and David Windridge, Breast Cancer Data Analytics With Missing Values: A study on Ethnic, Age and Income Groups, arXiv:1503.03680v1[q-bio.QM]12 Mar 2015.
102. Mu-HsingKuo , Tony Sahama , Andre W. Kushniruk and Elizabeth M. Borycki , Daniel K. Grunwell ,Health big data analytics;currentperspectives,challenges and potential solutions, Int. J. Big Data Intelligence, Vol. 1, Nos. 1/2, 2014.
103. Filip Velickovski, Luigi Ceccaroni, Josep Roca, Felip Burgos, Juan B Galdiz, Nuria Marina,MagíLluch-Ariet,Clinical Decision Support Systems (CDSS) for preventive management of COPD patients,Journal of Translational Medicine 2014, 12(Suppl 2):S9.
104. Xiang Wang, David Sontag, Fei Wang,Unsupervised Learning of Disease Progression Models, KDD’14, August 24–27, 2014, New York, NY, USA.ACM 978-1-4503-2956-9/14/08.http://dx.doi.org/10.1145/2623330.2623754.
105. Zhongheng Zhang, Big data and clinical research:Perspective from a clinician, doi: 10.3978/j.issn.2072-1439.2014.12.12.
106. Gabriele Guidi, Maria Chiara Pettenati, Paolo Melillo,ErnestoIadanza,,A Machine Learning System to Improve Heart Failure Patient Assistance, IEEE Journal Of Biomedical And Health Informatics, VOL. 18, NO. 6, November 2014.
107. P.Ramachandran, N.Girija, T.Bhuvaneswar, Early Detection and Prevention of Cancer using Data Mining Techniques, International Journal of Computer Applications (0975 – 8887)Volume 97– No.13, July 2014.
108. Ms.Tejashri N. Giri, S.R.Todamal, Data Mining Approach For Diagnosing Type 2 Diabetes, International Journal Of Science,Engineering And Technology-, Vol 2, Issue8 Nov-Dec 2014, Date Of Publication: Jan 02, 2015.
109. Noel Pérez , Miguel A. Guevara, Augusto Silva, Isabel Ramos and Joana Loureiro, Improving the performance of machine learning classifiers for Breast Cancer diagnosis based on feature selection, Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 209–217 , ACSIS, Vol. 2, DOI: 10.15439/2014F249.
110. Sam Royston,Practical Machine Learning for Diabetes Care, December 16, 2014.
111. Tony Hao Wu, Grantham Kwok-Hung Pang, Enid Wai-Yung Kwong,Predicting Systolic Blood Pressure Using Machine Learning, 978-1-4799-4598-6/14 ©2014 IEEE.
112. Aditya Methaila, Prince Kansal, Himanshu Arya, Pankaj Kumar, Early Heart Disease Prediction Using Data Mining Techniques, Sundarapandianetal.(Eds) : CCSEIT, DMDB, ICBB, MoWiN, AIAP – 2014, pp. 53–59, 2014.DOI : 10.5121/csit.2014.4807.
113. R Shouval1, O Bondi, H Mishan, A Shimoni, R Unger and A Nagler, Application of machine learning algorithms for clinical predictive modelling: a data mining approach in SCT, Bone Marrow Transplantation49, 332–337, 2014 Macmillan Publishers Limited.
114. S. Syed Shajahaan , S. Shanthi , V. ManoChitra, Application of Data Mining Techniques to Model Breast Cancer Data, International Journal of Emerging Technology and Advanced Engineering,Volume 3, Issue 11, November 2013.
115. R.M. Chandrasekar , V. Palaniammal,Performance and Evaluation of Data Mining Techniques in Cancer Diagnosis, IOSR Journal of Computer Engineering (IOSR-JCE)Volume 15, Issue 5 (Nov. - Dec. 2013), PP 39-44.
116. ShomonaGracia Jacob1 , R. Geetha Ramani2, Efficient Classifier for Classification of Prognostic Breast Cancer Data through Data Mining Techniques, Proceedings of the World Congress on Engineering and Computer Science 2012 Vol I WCECS 2012, October 24-26, 2012, San Francisco, USA.
117. Susana M. Vieir, Joao P. Carvalho, Andr´e S. Fialho, S. R. Reti, S. N. Finkelstein, Jo˜ao M.C. Sousa,A decision support system for ICU Readmissions Prevention, 978-1-4799-0348-1/13©2013 IEEE.
118. Joao Paulo Carvalho, SérgioCurto, Fuzzy Preprocessing of Medical Text Annotations of Intensive Care Units Patients,978-1-4799-4562-7/14©2014 IEEE.
119. Diogo Nunes, Paulo Carvalho, Jorge Henriques, Teresa Rocha Multiparametric prediction with application to early detection of cardiovascular events, 978-1-5386-3906-1/17©2017 IEEE.
120. SérgioCurto, Joao P. Carvalho, Cátia Salgado, Susana M. Vieira, João M. C. Sousa,Predicting ICU readmissions based on bedside medical text notes, 978-1-5090-0626-7/16_c 2016 IEEE.
121. Fen Miao, Yun-Peng Cai, Yu-Xiao Zhang, Xiao-Mao Fan, Ye Li, Predictive Modeling of Hospital Mortality for Patients With Heart Failure by Using an Improved Random Survival Forest, 2169-3536, Volume 6, 2018.
122. Janice Pan, Robert Shaffer, Zeina Sinno, Marcus Tyler, and JoydeepGhosh,The Obesity Paradox in ICU Patients, 978-1-5090-2809-2/17©2017 IEEE.
123. Neeraj Paradkar,Shubhajit Roy Chowdhury,Coronary Artery Disease Detection using Photoplethysmography, 978-1-5090-2809-2/17©2017 IEEE.
124. S.Sasikalaa , Dr.S.Appavu alias Balamuruganb and Dr.S.Geetha,A Novel Feature Selection Technique for Improved Survivability Diagnosis of Breast Cancer, 2nd International Symposium on Big Data and Cloud Computing (ISBCC’15), Procedia Computer Science 50 (2015) 16 – 23.
125. LiyingYang,Zhimin Liu, XiguoYuan,Jianhua Wei, and Junying Zhang, Random Subspace Aggregation for Cancer Prediction with Gene Expression Profiles, Hindawi Publishing Corporation BioMed Research International Volume 2016, Article ID 4596326, 10 pages http://dx.doi.org/10.1155/2016/4596326.
126. GunavathiChellamuthu, PremalathaKandasamy , Sivasubramanian Kanagaraj,Biomarker Selection from Gene Expression Data for Tumour Categorization Using Bat Algorithm, International Journal of Intelligent Engineering and Systems, Vol.10, No.3, 2017, DOI: 10.22266/ijies2017.0630.45.

127. Kingsley Kuan, Mathieu Ravaut, Gaurav Manek, Huiling Chen, Jie Lin, Babar Nazir, Cen Chen, Tse Chiang Howe, Zeng Zeng, Vijay Chandrasekhar,Deep Learning for Lung Cancer Detection: Tackling the Kaggle Data Science Bowl 2017 Challenge, arXiv:1705.09435v1 [cs.CV] 26 May 2017.
128. Albert Chon, Niranjan Balachandar, Peter Lu, Deep Convolutional Neural Networks for Lung Cancer Detection.
129. Cesar Suarez Ortega, Jose M. Franco Valiente, Manuel Rubio del Solar, Guillermo Daz Herrero, Raul Ramos Pollan, Miguel A. Guevara Lopez, Naimy Gonzalez de Posada, Daniel C. Moura, Pedro Cunha , Isabel Ramos , and Joana Loureiro,Improving the breast cancer diagnosis using digital repositories, IWBBIO 2013. Proceedings Granada, 18-20 March, 2013.
130. Noel Pérez , Miguel A. Guevara , Augusto Silva , Isabel Ramos and Joana Loureiro,Improving the performance of machine learning classifiers for Breast Cancer diagnosis based on feature selection, Proceedings of the 2014 Federated Conference on Computer Science and Information Systems pp. 209–217.
131. Jose Manuel Ortiz-Rodriguez, Carlos Guerrero-Mendez, Maria del Rosario Martinez-Blanco, Salvador Castro-Tapia, Mireya Moreno-Lucio, Ramon Jaramillo-Martinez, Luis Octavio Solis-Sanchez, Margarita de la Luz Martinez-Fierro, Idalia Garza-Veloz, Jose Cruz Moreira Galvan and Jorge Alberto Barrios Garcia,Breast Cancer Detection by Means of Artificial Neural Networks,INTECH, http://dx.doi.org/10.5772/intechopen.71256.
132. DarvinYi , Rebecca Lynn Sawyer , David Cohn III , Jared Dunnmon , Carson Lam , Xuerong Xiao , Daniel Rubin ,Optimizing and Visualizing Deep Learning for Benign/Malignant Classification in Breast Tumors, 29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.
133. immyWu ,Diondra Peck , Scott Hsieh , Vandana Dialani, MD , Constance D. Lehman, MD , Bolei Zhou , Vasilis Syrgkanis , Lester Mackey , and Genevieve Patterson ,Expert identification of visual primitives used by CNNs during mammogram classification, SPIE.
134. S.Sasikala , of the CAD systems by incorporating new technologi ,Fusion of Two View Binary Patterns to Improve the Performance of Breast Cancer Diagnosis, International Conference on Communication and Signal Processing, April 6-8, 2017, India, 978-1-5090-3800-8/17©2017 IEEE.
135. Basma A. Mohamed and Nancy M. Salem ,Automatic Classification of Masses from Digital Mammograms, 35th National Radio Science Conference 2018, Misr International University (MIU), Cairo, Egypt,© 2018 IEEE.
136. Karim Baati, Tarek M. Hamdani and Adel M. Alimi ,Diagnosis of Lymphatic Diseases Using A Na¨ıve Bayes Style Possibilistic Classifier,IEEE International Conference on Systems, Man, and Cybernetics, 978-1-4799-0652-9/13© 2013.
137. NirmalaDevi.M, Appavu alias Balamurugan.S, Swathi U.V,An amalgam KNN to predict Diabetes Mellitus,IEEE International Conference on Emerging Trends in Computing, Communication and Nanotechnology (ICECCN 2013), 978-1-4673-5036-5/13© 2013 IEEE.
138. Sankaranarayanan.S, Dr PramanandaPerumal.T ,Diabetic prognosis through Data Mining Methods and Techniques, 2014 International Conference on Intelligent Computing Applications, 978-1-4799-3966-4/14© 2014 IEEE.
139. Aarti Bhalla,Microarray Gene-expression Data Classification using Less Gene Expressions by Combining Feature Selection Methods and Classifiers, I.J. Information Engineering and Electronic Business, 2013, 5, 42-48.
140. Dr. M. Thangamani, Ms.J. Malar vizhi , Ms. M. Indhumathi,Intelligent Cancer Classification And Prediction In Micro Array Gene Analysis,International Journal of Electrical and Electronics Engineers,IJEEE, Vol. No.6, Issue No. 02, July-Dec.,2014.
141. Harshad Kokate, Neha Nair , Kalyani Shete, Trupti Thakur, Sujit Ahirrao, Detection of Colon Cancer by Classification of Genes and Feature Selection using Microarray Data, International Conference of Advance Research and Innovation (ICARI-2014).
142. Dr. E.S. Samundeeswari ,C.Sathya,Statistical Measures and Genetic Algorithm for Gene Selection in Breast Cancer, International Journal of Innovative Research in Science, Engineering and Technology, Vol. 5, Issue 3, March 2016.
143. Joana Diz 1 &Goreti Marreiros2 & Alberto Freitas,Applying Data Mining Techniques to Improve Breast Cancer Diagnosis, J Med Syst, 40:203 DOI 10.1007/s10916-016-0561-y, Springer Science+Business Media New York 2016.
144. Ayşe Demirhan,Classification of Structural MRI for Detecting Alzheimer’s Disease, International Journal of Intelligent Systems and Applications in Engineering, IJISAE, 2016, 4(Special Issue), 195–198.
145. V. B. Surya Prasath,Deep learning based computer-aided diagnosis for neuroimaging data: focused review and future potential, Neuroimmunology and Neuroinflammation, Neuroimmunol Neuroinflammation 2018;5:1.
146. Kajal Kiran Gulhare,S.P.Shukla,L.K.Sharma,DeepNeural Network Classification method to Alzheimer‟s Disease Detection, International Journals of Advanced Research in Computer Science and Software Engg. (Volume-7,Issue-6).
147. Leyi Wei and Quan Zou,Recent Progress in Machine Learning-Based Methods for Protein Fold Recognition,International Journal of Molecular Sciences, Int. J. Mol. Sci. 2016, 17, 2118; doi:10.3390/ijms17122118.
148. Mayuri Patel,Multi-class protein Structure Prediction Using Machine Learning Techniques, International Journal of Research in Advent Technology, Vol.2, No.12, December2014.
149. Kui Liu, Guixia Kang, , Ningbo Zhang, D Beibei Hou,Breast Cancer Classification Based on Fully-Connected Layer First Convolutional Neural Networks, 2169-3536 2018 IEEE, VOLUME 6, 2018.
150. JouhyunJeon ,SatraNim , Joan Teyra , Alessandro Datti, Jeffrey L Wrana , Sachdev S Sidhu, Jason Moffat and Philip M Kim,A systematic approach to identify novel cancer drug targets using machine learning, inhibitor design and high-throughput screening, Genome Medicine 2014, 6:57.
151. XiufengYang，HuiPeng ,Mingrui Shi ,SVM with Multiple Kernels based on Manifold Learning for Breast Cancer Diagnosis, Proceeding of the IEEE International Conference on Information and Automation Yinchuan, China, August 2013, 978-1-4799-1334-3/13©2013 IEEE.
152. Dishant Mittal, Dev Gaurav and Sanjiban Sekhar Roy ,An Effective Hybridized Classifier for Breast Cancer Diagnosis, 2015 IEEE International Conference on Advanced Intelligent Mechatronics (AIM), 2015. Busan, Korea.
153. T.Sridevi , A.Murugan,An Intelligent Classifier for Breast Cancer Diagnosis based on K-Means Clustering and Rough Set, International Journal of Computer Applications (0975 – 8887) Volume 85 – No 11, January 2014.
154. GokhanZorluoglu, Mustafa Agaoglu,Diagnosis of Breast Cancer Using Ensemble of Data Mining Classification Methods, International Journal of Bioinformatics and Biomedical Engineering Vol. 1, No. 3, 2015, pp. 318-322.
155. MaleikaHeenaye- MamodeKhan,Automated Breast Cancer Diagnosis Using Artificial Neural Network (ANN), 2017 3rd Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS).
156. Rabha O.Abd-elsalam, Yasser F.Hassan, Mohamed W.Saleh,New Deep Kernel Learning based Models for Image Classification, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No.7, 2017.

157. Kemal Akyol, Yasemin Gültepe, A Study on Liver Disease Diagnosis based on Assessing the Importance of Attributes, I.J. Intelligent Systems and Applications, 2017, 11, 1-9
158. Dilip Kumar Choubey,Sanchita Paul , GA_MLP NN: A Hybrid Intelligent System for Diabetes Disease Diagnosis, I.J.Intelligent Systems and Applications, 2016, 1, 49-59.
159. Seyyid Ahmed Medjahed, TamazouztAitSaadi, Abdelkader Benyettou,Urinary System Diseases Diagnosis Using Machine Learning Techniques, I.J. Intelligent Systems and Applications, 2015, 05, 1-7.
160. https://neugrid4you.eu/datasets-details.
161. Talha Mahboob Alama,Muhammad Atif Iqbala, Yasir Alia, Abdul Wahabb, Safdar Ijazb, Talha Imtiaz Baigb, Ayaz Hussainc, Muhammad AwaisMalikb, Muhammad Mehdi Razab, Salman Ibrarb, ZunishAbbasd A model for early prediction of diabetes, Informatics in Medicine Unlocked, 16 () 100204.
162. IliyanMihaylov, Maria Nisheva, and DimitarVassilevApplication of Machine Learning Models for Survival Prognosis in Breast Cancer Studies, 18th International Conference on Artificial Intelligence: Methodology, Systems, Applications, AIMSA 2018, Varna, Bulgaria, 12–14 September 2018.
163. MohamedAlloghani,AhmedAljaaf,AbirHussain,TharBaker ,JamilaMustafina,DhiyaAl-Jumeily and Mohammed Khalaf, Implementationofmachinelearningalgorithmstocreatediabeticpatient re-admissionprofiles,BMCMedicalInformaticsandDecisionMaking2019,19(Suppl9):253 https://doi.org/10.1186/s12911-019-0990-x.
164. Sai PoojithaNimmagadda, Sagar Yeruva, Rakesh Siempu,Improved Diabetes Prediction Model for Predicting Type-II Diabetes,International Journal of Innovative Technology and Exploring Engg. (IJITEE) ,Vol.-8 Issue-12, October, 2019.
165. Aishwarya Jakka, Vakula Rani J, Performance Evaluation of Machine Learning Models for Diabetes Prediction, International Journal of Innovative Technology and Exploring Engineering (IJITEE)Volume-8 Issue-11, September 2019.
166. Priyanka Israni, Breast Cancer Diagnosis (BCD) Model Using Machine Learning, International Journal of Innovative Technology and Exploring Engineering (IJITEE),Vol0ume-8 Issue-10, August 2019.