International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 /doi:10.14445/22315381/IJETT-V68I10P209

Volume 68 Issue 10, 52-57, October 2020

Disease Predictive Models for Healthcare by using Data Mining Techniques: State of the Art

Aman[1], Rajender Singh Chhillar[2]

[1]*Research Scholar, Department of Computer Science and Applications, M.D. University, Rohtak, Haryana, India*
[2]*Professor, Department of Computer Science and Applications, M.D. University, Rohtak, Haryana, India*

[1]sei@live.in, [2]chhillar02@gmail.com

Abstract *— Data mining in healthcare has a tremendous ability to explore the hidden pattern in medical datasets. These patterns can be helpful in both Disease diagnosis and prognosis. However, the raw medical data is complex, distributed, and large in size. Due to this, it becomes nearly impossible for the physician to manually process these data all alone. Analyzing this complex data requires lots of effort, time and money. It brings the need for automated Disease Predictive Models which will predict the disease with higher accuracy with lesser efforts. The usage of data mining in Indian healthcare sector has shown excellent potential for growth. This paper offers a summary of Data Mining Techniques, their Applications and current state of India in Healthcare sector in a systematic manner.*

Keywords *— Data Mining, Predictive Modeling, Healthcare, India.*

I. INTRODUCTION

Health is a fundamental human right of a citizen that is relevant to fulfilling the basic needs of human and improving their quality of life. Health is a critical factor that affects the overall level of development growth of a country. Since growth of a country is a result of good health, so investing in the healthcare sector becomes an important task for even a developing country like India. Unfortunately, healthcare sector has been poorly invested in India and that makes it largely neglected. According to Minister of Health and Family Welfare, India spends 1.4% of its GDP (Gross Domestic Product) on health sector which is very less in compare to its neighboring countries like Sri Lanka, Nepal and others [1]. Presently there is one doctor available for 1668 people which is not sufficient for country like India having population of 133.92 crores [2].

There has been an exponential growth in the sector of healthcare over the past two decades owing to big amounts of information produced about patients, medical devices, disease diagnosis, EHRs (electronic health records), and others. According to report, Indian healthcare industry generate zetta-bytes (1024 giga-bytes) of EHRs data every day [3]. This becomes a major challenge for the both physician and data analyst to convert this data into knowledge that can be used and cost-effective for prediction. The data analyst's primary objective is to create various predictive models for a distinct and complex disease such as heart disease, lung disease, diabetes mellitus, liver disease, and many more life-taking diseases.

This paper includes following sections: Recent work in the field of disease prediction has been mentioned systematically in Section 2. In Section 3, Knowledge Discovery in Database have been discussed which tells briefly how relevant knowledge extracted out of complex data. In Section 4, Data Models and Popular methods/algorithms have been discussed statistically and systematically. Finally, in rest Section, Importance and Applications of Data mining in the Indian Healthcare Sector have been discussed.

II. RECENT RESEARCH WORK

Numerous work have been done related to chronic disease prediction using different Data Mining techniques. The Dataset, Techniques, Algorithms used by the authors along with their findings that were carried out in the past few years are discussed below:

Bashir et al. (2019) have suggested for using Logistic Regression SVM (Support Vector Machine) with Minimum Redundancy Maximum Relevance (mRMR) Feature Selection technique for increasing the accuracy of the proposed models. The experiment was carried out on Cleveland Heart disease dataset from the University of California, Irvine (UCI) repository using Rapid Miner tool. mRMR used as a feature selection technique. Decision Tree, Logistic Regression, Logistic Regression SVM, Naïve Bayes (NB) and Random Forest algorithms (RF) are used as a prediction model. 5-fold cross-validation was used for measuring stability in terms of bias & variance. The research concludes that Logistic Regression SVM shows better accuracy (84.15%) than other classifiers [4].

Arif-Ul-Islam and S. H. Ripon (2019) have performed a comparative analysis of boosting algorithms and rules induction algorithms. The experiment carried out on Chronic Kidney Disease (CKD) from UCI repository using the Waikato Environment for Knowledge Analysis (WEKA) 3.8 and Myra Tool. The research concludes that AdaBoost performed better than LogitBoot in boosting algorithm and Ant Miner generate rules more efficiently than J48 decision Tree in rule induction algorithm [5].

Ebenuwa et al. (2019) have proposed variance ranking attribute selection technique and tried to solve issues of

classification of Imbalanced datasets. The experiment uses the British United Provident Association (BUPA) liver disorder dataset, Pima Indians Diabetes dataset, Wisconsin-breast cancer dataset, and Cod-RNA dataset. Proposed Feature selection technique performs significantly better than other Feature Selection Techniques such as Pearson Correlation and Information Gain [6].

Zhou et al. (2019) have proposed unsupervised classification technique, deep—learning Feature Learning (DFL) framework to solve out the issues faced in supervised feature selection technique. Deep learning enables the framework to automatically learn complex representations from the dataset. The experiment carried out by using two datasets. One is real-world pneumonia patient data collected from the National University Hospital (NUH) in Singapore. Another one is Electroencephalogram (EEG) dataset collected from the UCI repository. Pneumonia dataset is first interpolated and then passed to Uniform/Optimized Stacked de-noising Auto-encoder (USDAE/OSDAE). The research found that OSDAE performs better feature selection than PCA, Manual feature section. In order to validate the finding, it then tested with EEG dataset [7].

Al-tashi et al. (2018) have proposed ensemble algorithm that combines Ant Colony Optimization (ACO) and SVM. Discrete Wavelet Transform (DWT) used for filtering out clean and noisy data. ACO was used for feature selection. The proposed algorithm produces a satisfactory result on various datasets like Heart-Statlog, Breast Cancer, Hepatitis, Diabetes, and Liver dataset [8].

Chhabra et al. (2018) have proposed ensemble algorithm (stacking). It uses LDA, KNN as level-0 classifier and RF, Generalised Linear Model (GLM) [level-1] as meta-classifier. The experiment was done on ionosphere dataset collected from the UCI repository and performed using the caret package of R programming language. The research concludes that the ensemble algorithm performs better than SVM with linear kernel and other classifiers [9].

S. Rani and S. Kautish (2018) have proposed a predictive system based on time series data mining for continuous data. The experiment is carried out Pima Indians Diabetes data from the UCI repository and concluded that time series with association mining technique gives better against Artificial Neural Network (ANN) for continuous data [10].

Avci et al. (2018) have performed a performance analysis of NB, K-Star, SVM and J48 classifiers. The experiment carried out on CKD dataset obtained from the UCI repository. Experiment performed on WEKA data mining tool. The research concludes that J48 classifier outperforms with an accuracy of 99.00% [11].

A. Mir and S. N. Dhage (2018) have aimed at building a classifier to predict diabetes disease by using NB, SVM, RF and Simple CART algorithm. The experiment was carried on Pima Indians Diabetes from the UCI repository. Dataset is split into Training and Test dataset in 7:1 ratio. Experiment performed on WEKA 3.8 data mining tool. Research concludes that SVM performed better than another classifier with a maximum accuracy of 79.13% [12].

Singh et al. (2018) have proposed hybrid feature extraction approach for detection of abnormality in ECG. The experiment carried out using Arrhythmia database collected from MIT/BIH. During Pre-processing stage, Dual-Tree Complex Wavelet Transform (DTCWT) performed to filter significant data out of noisy data. Then passed to Linear Discriminate Analysis (LDA) for dimensionality reduction. Various classifier such as SVM, DT, Back Propagation Neural Network (BPNN), Feed Forward Neural Network (FNNN) and KNN are used to test the performance of the hybrid model. BPNN, SVM, and KNN achieved 99.7% of accuracy with the hybrid feature extraction approach [13].

Solanki et al. (2018) have proposed a hybrid approach for classification of brain Magnetic resonance imaging (MRI) images. The author used Gray-level co-occurrence matrix (GLCM) technique used for feature extraction and then these features passed to three classifier SVM, SVM with K nearest neighbor (SVM-KNN), SVM with Radial basis function kernel (SVM-RBF) and finally conclude that SVM-RBF produces maximum accuracy [14].

K. Pahwa and R. Kumar (2017) have proposed hybrid technique for feature selection in predicting heart disease. Experiment analysis carried out on Heart stat-log dataset from the UCI repository. Feature selection is done using Support Vector machine with Recursive Feature Elimination (SVM-RFE) to reduce dataset and to improve performance then feature passed to the random forest and Naive Bayes classifier. Both of the classifiers produce equal and satisfactory result [15].

Chen et al. (2017) have proposed hybrid model for prediction of Type-2 diabetes. Experiment analysis carried out on Pima Indian Diabetes dataset using the WEKA data mining tool. K-mean was used to reduce data size or for feature selection and then passed to J48 Decision Tree classifier. Proposed model provides better accuracy than C5.0, CART, Random Forest, and Naive Bayes [16].

Rajeswari et al. (2017) have proposed a dimension reduction model — Particle Swarm Optimization with Pulse coupled Neural Network (PSO-PCNN). The best attributes from PCO algorithm passed to PCNN for further improvement in dimension reduction. Experiment analysis carried out on Impaired Glucose Tolerance (IGT) dataset for diabetes. The PSO-CNN provide 95% better accuracy when compared to existing dimension algorithms [17].

III. KNOWLEDGE DISCOVERY IN DATABASE

Knowledge Discovery in Database (KDD) is a method of extracting knowledge or significant information from the enormous amount of data. The KDD process includes the following stages as shown in Fig. 1.

Data is stored in a variable. This stored data is complicated, making it hard to understand and needed to be processed. In order to make our Statistical data understandable to Data Mining algorithm(s), it need to be categorized because not all data mining algorithm(s) will

operate on each type of data. Data is generally of two types: Structured Data, and Unstructured Data. Structured Data have pre-defined structure that resides in relational databases. Most of Structured Data set repositories available on UCI (University of California, Irvine) [18], KEEL (Knowledge Extraction based on Evolutionary Learning) [19], Kaggle [20] website and others. Unstructured data doesn't have pre-defined structure that resides in NO-SQL Databases. It can be in Image, and Time Series format. Image dataset includes X-ray, Computed tomography (CT) scan, Positron emission tomography (PET) scan, Magnetic resonance imaging (MRI) scan, and others. Time Series dataset includes Electrooculogram (EOG), Electroencephalography (EEG), Electrocardiography (ECG), and others. Various Types of Data used in previous research work is listed down systematically in TABLE I.

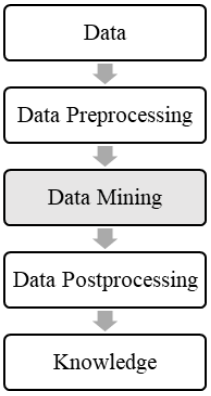


Fig. 1 KDD Process

TABLE I: VARIOUS TYPES OF DATA USED IN THE RECENT WORK

Types of Data	References
Structured Data	[4]–[6], [8], [9], [11], [12], [15]–[17]
Image based Unstructured Data	[14]
Time Series based Unstructured Data	[7], [10], [13]

In *Data Pre-processing stage*, data is selected from various sources like a relational database, JavaScript Object Notation (JSON) files, flat files, and various data repositories. Such data may contain lots of missing values, out of ranges values and noise. Such type of data is called messy data and required to be cleaned (scrubbing). That's why Data Pre-processing sometimes also called by "Data Munging or Data Wrangling". Cleaning of data is a time-consuming process that is why it requires lots of attention by the data analyst. Using encoding and binning methods, data can be transformed into another format. Finally, transformed data is loaded into the destination database by maintaining integrity. Data Pre-processing phase is also known as "Extraction, Transformation, and Loading (ETL)" process.

Data Mining is the process of applying an algorithm(s) on pre-processed data to extract hidden patterns. These algorithm(s) also called Data Mining Engine (DME). Classification algorithms, Clustering algorithms, and Association rules, etc. are an example of DME. Each DME have its own merits and demerits. According to "No Free Lunch (NFL) theory", no single DME can outperform for every single domain [21]. It depends on the nature of the dataset. For some dataset, some algorithm will excel and some other dataset, some other algorithm will perform better. After modeling, evaluation is performed. In evaluation, metrics are used that will help to decide which model fit or support our data properly.

In *Data Post-processing stage*, the extracted pattern is described by using visualization techniques such as Mathematical equations, and Graphs, etc. This is the final stage of KDD that describe knowledge or relevant information.

IV. DATA MINING MODELS, TASKS AND THEIR METHODS

Commonly, there are two types of Data Mining models: 1) Predictive model 2) Descriptive model as shown in Fig. 2. The predictive model often uses supervised learning algorithm(s) to predict unknown or future values of variables of interest. The descriptive model usually uses the unsupervised learning algorithm(s) that group unlabeled data in the form of clusters on the basis of implicit behavior. The predictive models are more commonly used in the healthcare sector.

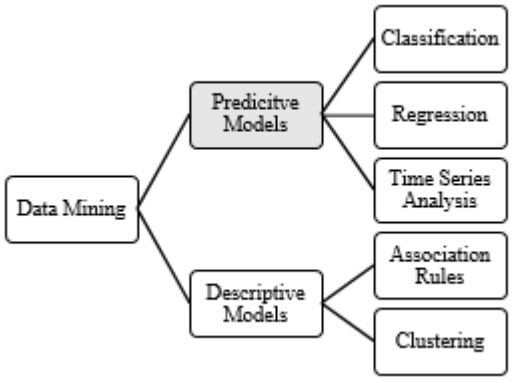


Fig. 2 The different type of models, tasks and their methods in Data Mining

In Simple language, the task is what analyst wants to find out such as Prediction/Forecasting, Segmentation and many more. In descriptive models, clustering and association rules are commonly used. On the other side classification, regression, and test series analysis are used for predictive models.

After defining the Data Mining model and task, the next step would be to use methods (well-established procedures) according to the problem statement. Popular Algorithms/Methods used in the existing work related to the disease prognosis and diagnosis listed down in TABLE II. The methods widely used for classification are Decision tree, K-nearest neighbor, support vector machine, Discriminant

analysis, Markov based, swarm intelligence, Genetic classifiers, Artificial Neural Network, Multivariate Adaptive Regression Splines (MARS), and Association Rule. Vector-based methods are commonly used for clustering.

TABLE II: POPULAR METHODS USED FOR DISEASE PREDICTION IN THE RECENT WORK

Data Mining Methods	References
Decision Tree (J48, CART)	[4]–[6], [9], [11], [12], [16]
KNN	[9], [13], [14]
SVM	[4], [6], [8], [9], [11], [12], [14]–[16]
K-Mean	[16]
K-Star	[11]
Artificial Neural Network	[10]
Naïve Bayes	[4], [11], [12]
Logistic Regression	[6]
Random Forest	[4], [12]
Swarm Intelligence (Ant-miner)	[5], [8], [17]

A. Classification

Classification is a data mining technique that uses labeled data (target class) to classify unlabeled data. The main objective of a classifier to accurately predict the labeled data for each data. For example, a classifier could be used to identify whether the patient is healthy or sick on the basis of their age and gender.

Decision tree is a supervised learning algorithm that builds classifier and regressor in the form of a tree-like structure. Each node represents a "test" on a variable and their corresponding branch represent the result of that "test". Internal-nodes are denoted by rectangle shape and leaf-nodes are denoted by oval shape. Splitting criteria is based on the value of Information Gain (IG), and choose variable as decision node that has the highest IG value. Iterative Dichotomiser 3 (ID3), C4.5/J48, and Classification and Regression Trees (CART) are some variants of decision tree.

K-Nearest Neighbor (KNN) algorithm was designed to find the nearest neighbor to the observed object using distance formulae like Euclidian, Manhattan, etc. The main issue is to decide the value of K. If K is small then noisy data will dominate, If K is large then it becomes computationally expensive.

Support Vector Machines (SVM) algorithm performs classification by finding the hyper-plane. Hyper-plane would have chosen on the basis of margin (from hyper-plane to class). More the Margin, better will be the classification of classes. For classification of simple data, SVM doesn't require kernel trick but for the complex and non-linear problem, SVM requires kernel trick for better classification. Radial Basis Function (RBF), Recursive Feature Elimination (RFE), Polynomial, Hyperbolic Tangent are some popular kernel trick used with SVM.

B. Clustering

Clustering is a data mining technique that categorized unlabeled data on the basis of their similarities. The main issue in building a cluster is to find similar objects. Usually, the clustering technique uses distance formulae such as Euclidian, Minkowski, and Manhatten methods, etc. These Distance method returns a cluster of objects that are similar in behaviour. Smaller the distance, More similar the objects are. Clustering Techniques are mostly used in Pre-processing phase for dimensionality reduction and outlier detection.

C. Association Rules

Association is a data mining that used to discover the relation between variables. It used IF-THEN statements to analyze and predict customer behavior. For example, if a customer buys onion with potatoes and tomatoes then he will buy green chilies with them frequently.

$$buy(onions,potatoes,tomatoes) \Rightarrow buy(green\ chilies)$$

V. HEALTHCARE IN INDIA

Healthcare is now one of the leading sectors in India, both in terms of revenue and jobs. Healthcare covers hospitals, medical equipment, outsourcing, and health insurance. According to IBEF report, The healthcare market could rise by 2022 to INR 8.6 trillion [22]. In Order to boost heath sector, world's largest government scheme, Ayushman Bharat was launched on 23 September, 2018 [23]. Developed countries spend nearly 10 percent of their total spending on healthcare for their citizens, according to the World Bank report. This expenditure helps a country improve its citizens' life expectancy rates. It is therefore suggested that the healthcare sector can significantly enhance a nation's economy. India has failed to achieve healthcare objectives, lagging far behind Bhutan, Sri Lanka, Bangladesh and developed countries in terms of accessibility and quality as shown in Fig. 3. According to the Global Burden of Disease Study (GBD) report, the top three causes of death in India are Heart disease, Lung Disease, and Stroke disease as shown in Fig. 4 [24].

According to research from McKinsey and Company, introducing a Clinical Decision Support System (CDSS) in the healthcare sector will eventually reduce the total expenditure by 12-17% [25]. This can be a win/win situation but due to slower adaption rate of technology, healthcare lags behind in implementing data mining techniques on a broad scale.

VI. DATA MINING APPLICATIONS IN THE HEALTHCARE SECTOR

Healthcare sector generates an enormous amount of complex data. It needed to be processed and analyzed in a cost-effective manner. Data mining techniques aid the Healthcare sector in the decision making the process. Important and well-known Data Mining applications are described below:

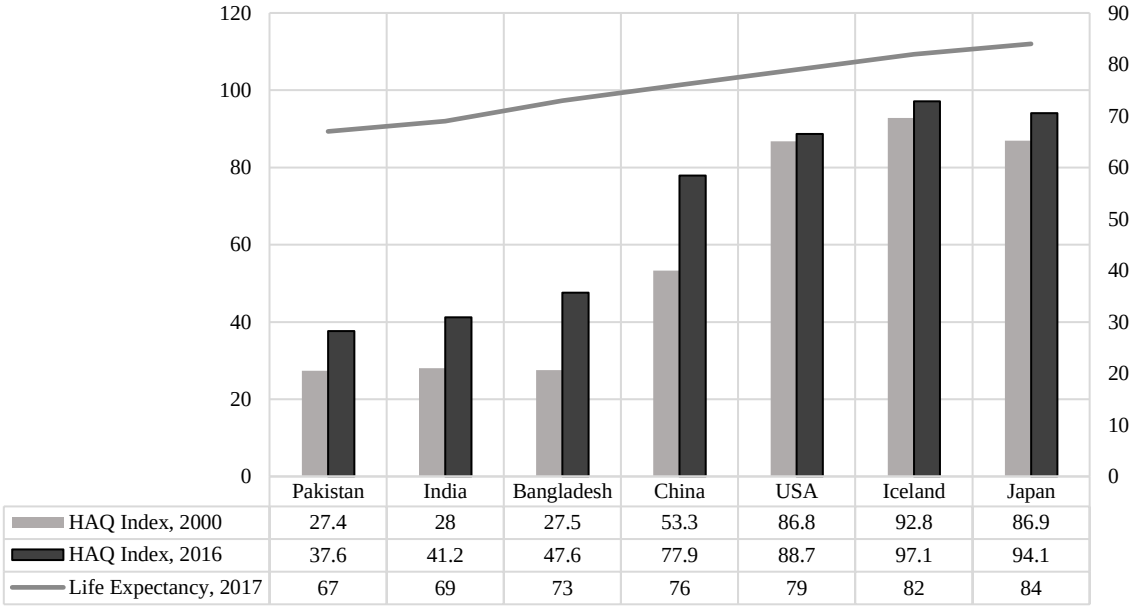
1) *Improving Treatment effectiveness:* More than two-thirds of the deaths are from chronic disease. It's very important to keep track of almost all disease' symptoms by a physician. Data mining helps in building an automated decision support system called CDSS which can be helpful in disease prognosis and diagnosis with more accuracy in lesser

time.

2) *Hospital Management:* It is important to store temporal data of staff members, and patients. It is a very time-consuming process when a patient moves from one hospital to another hospital or in case the patient lost his/her medical reports. The hospital needs to store patient records online using timeline method. It is also beneficial for hospital management to track their staff member — who took how many leaves, their performance, and achievements.

3) *Management of Biological databases:* Biological databases contain a large amount of complex data. For Example, Microarray database contains genomic data. It stores data by experimenting on Deoxyribonucleic Acid (DNA), blood cells, etc.

4) *Pharmaceutical Industry:* In the Pharmacy sector, technology is upgrading day by day up to a satisfactory level. Every day a new technique is developed to tackle dangerous diseases. So they need to update their inventories regarding salts, symptoms of disease.



	Pakistan	India	Bangladesh	China	USA	Iceland	Japan
HAQ Index, 2000	27.4	28	27.5	53.3	86.8	92.8	86.9
HAQ Index, 2016	37.6	41.2	47.6	77.9	88.7	97.1	94.1
Life Expectancy, 2017	67	69	73	76	79	82	84

Fig. 3 Healthcare Access and Quality (HAQ) and Life Expectancy

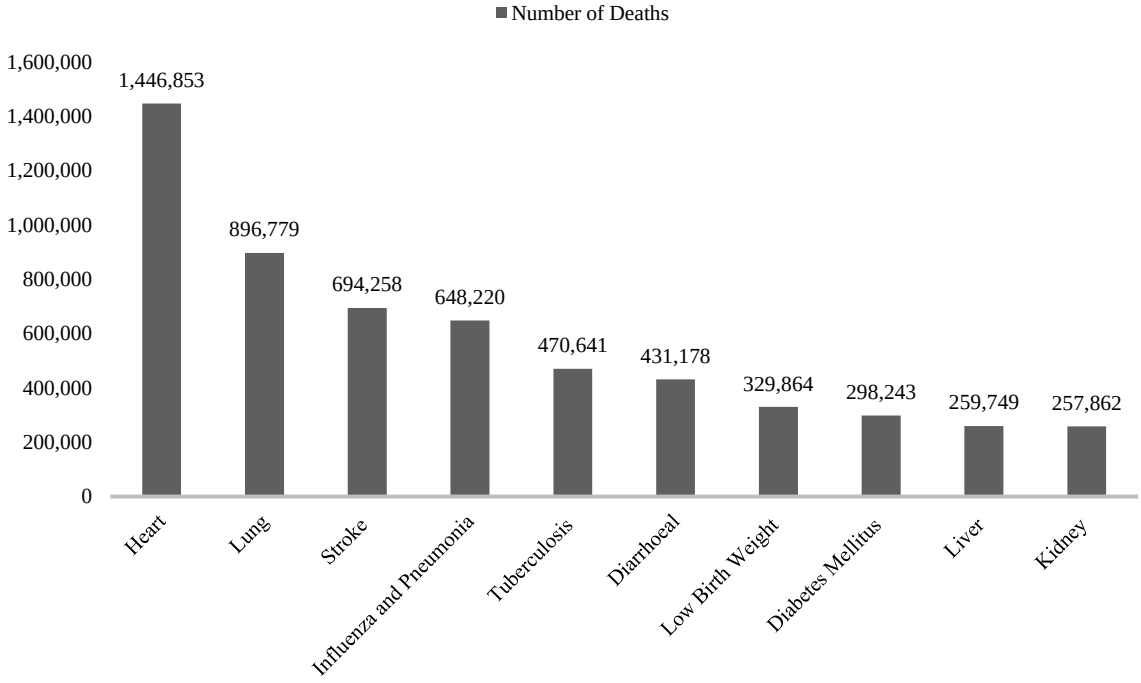


Fig. 4 Top 10 causes of death in India, GBD (2017)

VII. CONCLUSIONS

This paper provides a systematic state-of-the-art of data mining techniques to predict disease. Recent work based on the prediction of disease was taken up and classified in terms of the problems they solved, the types of data they used and the models they used. In addition, the paper also provides a data mining perspective on the state of healthcare research in India. It can be seen that, given its less strict privacy rules, India is lagging behind in research initiatives to make the data accessible. This paper will help the policy makers, researchers, and data analysts to understand the current state of the art of data mining in healthcare sector from Indian Perspective point of view.

REFERENCES

[1] "*India's spending on health sector has grown: Nadda - The Economic Times.*" https://economictimes.indiatimes.com/industry/healthcare/biotech/healthcare/indias-spending-on-health-sector-has-grown-nadda/articleshow/65309487.cms?from=mdr (accessed Dec. 10, 2019).

[2] R. Kumar and R. Pal, "*India achieves WHO recommended doctor population ratio: A call for paradigm shift in public health discourse!,*" J. Fam. Med. Prim. care, vol. 7, no. 5, pp. 841–844, 2018, doi: 10.4103/jfmpc.jfmpc_218_18.

[3] "*Big Data Analytics And Indian Healthcare - Express Healthcare*." https://www.expresshealthcare.in/features/big-data-analytics-and-indian-healthcare/162330/ (accessed Dec. 11, 2019).

[4] S. Bashir, Z. S. Khan, F. Hassan Khan, A. Anjum, and K. Bashir, "*Improving Heart Disease Prediction Using Feature Selection Approaches*," Proc. 2019 16th Int. Bhurban Conf. Appl. Sci. Technol. IBCAST 2019, pp. 619–623, 2019, doi: 10.1109/IBCAST.2019.8667106.

[5] Arif-Ul-Islam and S. H. Ripon, "*Rule Induction and Prediction of Chronic Kidney Disease Using Boosting Classifiers, Ant-Miner and J48 Decision Tree,*" in 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), Feb. 2019, pp. 1–6, doi: 10.1109/ECACE.2019.8679388.

[6] S. H. Ebenuwa and M. H. D. S. Sharif, "*Variance Ranking Attributes Selection Techniques for Binary Classification Problem in Imbalance Data,*" IEEE Access, vol. 7, pp. 24649–24666, 2019, doi: 10.1109/ACCESS.2019.2899578.

[7] C. Zhou, Y. Jia, and M. Motani, "*Optimizing Autoencoders for Learning Deep Representations from Health Data*," IEEE J. Biomed. Heal. Informatics, vol. 23, no. 1, pp. 103–111, 2019, doi: 10.1109/JBHI.2018.2856820.

[8] Q. Al-Tashi, H. Rais, and S. J. Abdulkadir, "*Hybrid Swarm Intelligence Algorithms with Ensemble Machine Learning for Medical Diagnosis*," 2018 4th Int. Conf. Comput. Inf. Sci. Revolutionising Digit. Landsc. Sustain. Smart Soc. ICCOINS 2018 - Proc., pp. 1–6, 2018, doi: 10.1109/ICCOINS.2018.8510615.

[9] G. Chhabra, V. Vashisht, and J. Ranjan, "*A classifier ensemble machine learning approach to improve efficiency for missing value imputation*," 2018 Int. Conf. Comput. Power Commun. Technol. GUCON 2018, no. April, pp. 23–27, 2019, doi: 10.1109/GUCON.2018.8674904.

[10] S. Rani, "*Association Clustering and Time Series Based Data Mining in Continuous Data for Diabetes Prediction*," *2*018 Second Int. Conf. Intell. Comput. Control Syst., no. Iciccs, pp. 1209–1214, 2018.

[11] E. Avci, S. Karakus, O. Ozmen, and D. Avci, "*Performance comparison of some classifiers on Chronic Kidney Disease data*," 6th Int. Symp. Digit. Forensic Secur. ISDFS 2018 - Proceeding, vol. 2018-Janua, pp. 1–4, 2018, doi: 10.1109/ISDFS.2018.8355392.

[12] A. Mir and S. N. Dhage, "*Diabetes Disease Prediction Using Machine Learning on Big Data of Healthcare*," Proc. - 2018 4th Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2018, pp. 1–6, 2019, doi: 10.1109/ICCUBEA.2018.8697439.

[13] R. Singh, N. Rajpal, and R. Mehta, "*Abnormality detection in ECG using hybrid feature extraction approach*," *2*018 First Int. Conf. Secur. Cyber Comput. Commun., pp. 461–466, 2019, doi: 10.1109/icsccc.2018.8703349.

[14] V. Solanki, "*Brain MRI Image Classification using Image Mining Algorithms,"* 2018 Second Int. Conf. Comput. Methodol. Commun., no. Iccmc, pp. 516–519, 2018, doi: 10.1109/ICCMC.2018.8487690.

[15] K. Pahwa, A. C. S. Dangare, S. S. Apte, and I. Study, "*Prediction of Heart Disease Using Hybrid Technique For Selecting Features*," pp. 500–504, 2017.

[16] W. Chen, S. Chen, H. Zhang, and T. Wu, "*A hybrid prediction model for type 2 diabetes using K-means and decision tree,*" Proc. IEEE Int. Conf. Softw. Eng. Serv. Sci. ICSESS, vol. 2017-Novem, no. 61272399, pp. 386–390, 2018, doi: 10.1109/ICSESS.2017.8342938.

[17] S. Rajeswari, M. S. Josephine, and V. Jeyabalaraja, "*Dimension Reduction: A PSO-PCNN Optimization Approach for Attribute Selection in High-Dimensional Medical Database,*" *2*017 IEEE Int. Conf. Power, Control. Signals Instrum. Eng., pp. 2306–2309, 2017.

[18] "*UCI Machine Learning Repository*." https://archive.ics.uci.edu/ml/index.php (accessed Dec. 10, 2019).

[19] "*KEEL: A software tool to assess evolutionary algorithms for Data Mining problems (regression, classification, clustering, pattern mining and so on).*" https://sci2s.ugr.es/keel/description.php (accessed Dec. 10, 2019).

[20] "Datasets | Kaggle." https://www.kaggle.com/datasets (accessed Dec. 10, 2019).

[21] D. H. Wolpert and W. G. Macready, "*No free lunch theorems for search,*" Unknown, 1995, doi: 10.1145/1389095.1389254.

[22] "*Healthcare Industry in India, Indian Healthcare Sector, Services.*" https://www.ibef.org/industry/healthcare-india.aspx (accessed Dec. 11, 2019).

[23] "*Health System for a New India: Building Blocks,*" *Niti Aayog*, 2019. https://niti.gov.in/sites/default/files/2019-11/NitiAayogBook_compressed.pdf (accessed Dec. 11, 2019).

[24] N. Watts *et al.*, "T*he 2018 report of the Lancet Countdown on health and climate change: shaping the health of nations for centuries to come,*" The Lancet. 2018, doi: 10.1016/S0140-6736(18)32594-7.

[25] USF Health, "*Data Mining In Healthcare | USF Health Online*," 2019. https://www.usfhealthonline.com/resources/key-concepts/data-mining-in-healthcare/ (accessed Dec. 12, 2019).