International Journal on Emerging Technologies 11(3): 554-560(2020)

ISSN No. (Print): 0975-8364
ISSN No. (Online): 2249-3255

A Survey on Big Data Analytical Tools & Techniques in Healthcare Sector

Sarita Mishra[1], Manjusha Pandey[2], Siddharth Swarup Rautaray[2] and Mahendra Kumar Gourisaria[3]
[1]M.Tech. Scholar, School of Computer Engineering, KIIT (Deemed to be University) (Odisha), India.
[2]Associate Professor, School of Computer Engineering, KIIT (Deemed to be University) (Odisha), India.
[3]Assistant Professor, School of Computer Engineering, KIIT (Deemed to be University) (Odisha), India.

(Corresponding author: Sarita Mishra)
(Received 04 March 2020, Revised 27 April 2020, Accepted 29 April 2020)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: Massive amounts of data in different forms (structured, semi-structured, unstructured) belonging to various applications of healthcare needs to be handled for efficient and informed decision making in such healthcare applications. Such enormous amounts of data enthralled by some specific features are called 'Big data'. These features are volume, velocity, variety, variability, veracity, volatility, visualization, value, and vagueness. This big data needs to be properly stored and preprocessed before being analyzed for targeted results. Due to the huge volume of different variety of medical data arriving at a high velocity to the information repository centres of healthcare units, it becomes very difficult for the concerned data scientists to analyze this data, thus increasing its significance. Because of this significance, traditional data handling mechanisms often fall back in completely exploiting the enormous amount of available data in depth; hence the development of more efficient algorithms, tools, techniques for analysis of medical big data is called for. This paper provides an idea about the areas of the healthcare sector which use big data analysis. It also specifies various features extraction techniques for the selection of relevant features for detecting a disease ignoring the rest. And more importantly, this paper has provided knowledge about the commonly used machine learning algorithms (various classification, prediction, clustering algorithms) such as decision trees, SVM, neural networks, linear regression, and a few more for analysis of the selected relevant features. Also, this paper enumerates some application domains of big data analysis in the healthcare sector have been discussed. Lastly, it provides the advantages of some of the machine learning algorithms that are commonly used.

Keywords: healthcare, big data, feature engineering techniques, machine learning algorithms, healthcare application domains.

I. INTRODUCTION

The phrase "*big data*" has come into widespread use among data scientists all over the world for decades. It plays a vital role in every sector of research, starting from schools and universities to big industries, which generate large amounts of data and also need to analyze this data for various purposes. Health care communities have also become advanced in achieving the desired accuracy in predicting and curing diseases using big data analytics. The importance of big data analysis lies in the provisioning of proper medication and treatment at the proper time. When the patients show symptoms of diseases that are distinctly visible, existing healthcare systems can treat them. However, early detection of acute diseases helps to treat the patient, preventing the disease to reach a fatal state. Otherwise, it may sometimes lead to chronic disorders, and sometimes the death of the patients may occur. Medical data is sensitive in nature and is available in different formats that are structured, semi-structured, or even unstructured. In order to handle this heterogeneous data, designing a distributed data handling platform becomes unavoidable. However, there are certain challenges faced while building such a platform. Firstly, the collection of an enormous amount of data from different sources in different formats at a rapid rate is a challenging task. Secondly, storage is the main problem for such massive, heterogeneous datasets as they need to be transformed into a format compatible with the storage system, without any degradation in performance. The last challenge is related to data scientists whose task is to analyze the big data in real-time or nearly real-time and come up with patterns that aid in developing visualization, optimization, and prediction algorithms. These challenges require new machine learning algorithms as the traditional data analysis systems are not capable of handling heterogeneous data in real-time.

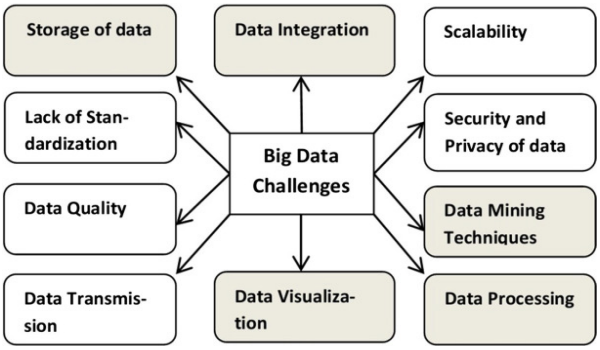

Fig. 1. Challenges in handling big data.

However, the major challenge on which this survey is based on is data mining techniques, i.e., identifying the hidden patterns in huge data sets available and drawing conclusions based on them.

Identifying these hidden patterns from huge datasets can be easily achieved using Artificial Intelligence algorithms. Also, it would be a remarkable feat to be able to automate decision-making with the help of machine learning algorithms like decision trees, neural networks, etc. However, without the usage of proper software and hardware, it becomes very difficult to handle this vast ocean of information and extract the desired knowledge from it. Hence, the requirement to develop better techniques to analyze this data arises, thus making the healthcare services provided to the patients more rapid and effective.

The healthcare system comprises of health professionals (doctors and other medical staff), healthcare facilities (hospitals and properly equipped test centers), and a financial organization supporting them. The health professionals belong to different specializations such as medicine, physiotherapy, dentistry, psychology, etc. Healthcare services are provided to the patients based on the criticality of their condition. Professionals are consulted at first as a part of primary care, acute care is provided by even more skilled doctors or physicians constituting the secondary level, more acute treatment involving advanced technologies that is the tertiary level care, is provided when suggested by former two and quarternary care is provided in case of rarely occurring diseases or surgeries. At all aforementioned healthcare levels, it is the responsibility of the healthcare professionals to generate and handle all information related to the patient involving the patient's past medical records, current check-up results as well as current test results. Earlier the patient's medical history and test results were commonly handwritten or typed and stored in form of hard-coded documents. However, with the rapid increase in the volume of health data generated, the storage and maintenance of the paper file system became very difficult thus paving way for digitization of data.

The first advantage of digitization of healthcare big data is the improved access privileges to all the relevant information about a patient by the healthcare professionals. This information contains basic information like name, age, gender as well as health-related information such as past diseases, undergoing check-ups, results of previous and current medical tests, etc. This ease in access provided to healthcare professionals led to prompt and early recognition and treatment of medical conditions thus making the healthcare system more efficient. Digitization of healthcare data has significantly reduced ambiguities caused by illegible handwriting or typing mistakes thus eliminating the repeated examination, redundant medication for a particular patient. This has led to a remarkable improvement in the healthcare services provided to the patients.

The remaining part of the paper is arranged as mentioned herewith. Section II discusses the work done by several researchers and the accuracy level achieved by them. Section III consists of the various healthcare activities which involve big data analysis and the categorization of machine learning algorithms being used. Sections IV and V provide knowledge about the commonly used feature extraction algorithms and data analysis algorithms based on machine learning respectively. Section VI mentions the several tools and techniques used by researchers in their works. And lastly, preceding the conclusion, section VII provides an idea about various application domains in healthcare sectors where big data analytics can be used.

II. STATE OF THE ART

Here are a few works by researchers who have used several machine learning algorithms to detect the presence of heart disease in order to provide better and cost-effective healthcare services to the patients. Purushottam *et al.*, (2016) used an open-source JAVA programming apparatus known as KEEL (Knowledge Extraction based on Evolutionary Learning) to implement a developmental process for data mining issues [1]. He used the Decision Tree algorithm to generate the classification rules for heart disease classification. The generated rules are accepted or rejected based on 3 parameters: *Confidence, MinItemSets, and Threshold.* The performance accuracy achieved in this experiment is 86.7%. Princy and Thomas (2016) used the k-nearest neighbors (KNN) classification method and Iterative Dichotomiser 3 (ID3) algorithm for prediction of the class to which a test case belongs [2]. The ID3 algorithm generates a decision tree for the data classified using the KNN algorithm. In the decision tree generated, each subnode holds the already classified data for each class. Using this tree, the non-classified data points are checked and the presence or likelihood of the disease in the patient is estimated. The performance accuracy of this method was 80.6%. Saboji and Ramesh (2017) has proposed the prediction of heart disease with a small number of attributes using Random forest classification which gives immense opportunity to healthcare analysts to deploy this solution on the ever-changing, scalable big data landscape for insightful decision making. The accuracy obtained here is 98% [3].

Mane and Tejaswini (2017) has proposed a heart disease prediction technique that is smarter than all others using Improved K-Means for clustering followed by the ID3 algorithm of decision tree for classification which led to a prediction of heart disease patients with 96.73% accuracy. Most of the previous researches on heart disease prediction detect the presence of heart disease for patients above 50 years of age [4]. Karthick and Priyadharshini (2018) have presented a methodology for prediction of the same for people within 50 years of age. In their work, they have used Principal Component Analysis (PCA) for feature selection and Naive Bayes classifier for classification purposes [5]. Latha and Jeeva (2019) [6], in their paper, have investigated ensemble classification for improving the accuracy of weak algorithms. The classification algorithms ensemble includes Bayes Network, Naive Bayes, Random Forest, C4.5, Multilayer perceptron, and PART. The results of this experiment have shown that ensemble methods like bagging, boosting, stacking, majority vote, etc have led to an improvement in accuracy by 7% in heart disease prediction. Further improvement in performance was achieved with a feature selection implementation.

Ed_daoudy and Maalmi (2019) have proposed a machine learning algorithm for real-time prediction of

heart-related issues based on the Apache Spark platform [7].

The system uses 2 important partitions, namely, stream data storage and preprocessing followed by data visualization. The first part uses the Spark streaming feature of apache spark and applies machine learning algorithms on health datasets to determine the presence of heart disease. The second part uses Apache Cassandra for storage purposes. The prediction model for detection and prediction of heart-related issues is constructed using the random forest algorithm and MLlib. Mohan *et al.,* (2019) have introduced the concept of Hybrid Random Forest with the Linear Model (HRFLM) to predict heart disease [8]. In HRFLM three association rules of mining are used for feature selection: Apriori, Predictive, and Tertius. HRFLM has used Artificial Neural Network (ANN) with back-propagation along with 13 attributes/features as input. The platform used in this research is R Studio and the classification algorithm used is a hybrid of Decision Tree, Language model, SVM, Random Forest, Naive Bayes, Neural Network, and K-Nearest Neighbors. The accuracy obtained is 88.47%. Alarsan and Younes (2019) proposed an ECG classification approach based on several ECG features implemented using MLlibs and Scala language on the Apache Spark framework. The feature extraction technique used is a DWT (Discrete Wavelet Transform) function [9]. The result of feature extraction, that is, the selected features were categorized into categories: Summits, Temporal, and Morphological. After feature extraction, the classification was done using 2 ways: Gradient Boosted Trees (GBT) with MaxDepth and MaxIter, Random Forest with MaxDepth, and NumTrees.

III. CLASSIFICATION

There are several sections in the healthcare domain where big data analysis plays a vital role. These sectors are represented in the following figure.

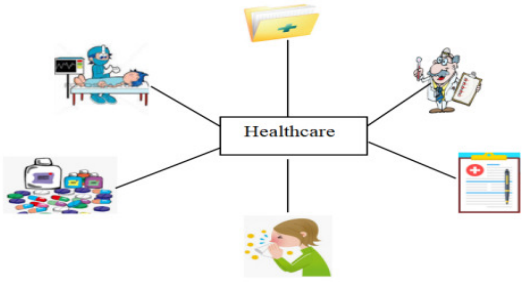

Fig. 2. Healthcare Activities.

As shown in the above figure, the first section of healthcare includes the collection of patient's basic details such as name, age, gender, etc, any allergies and the health issues which he has faced previously. In the second phase the diagnosis of various factors or symptoms of the patient to detect the presence of any disease.

In the third phase, results collected from the previous 2 stages are combined, organized, and analyzed by the physician. The fourth and fifth stages provide the patient with appropriate preventive measures to prevent any disease which may affect the patient or proper medication, based on the analysis performed. Lastly, based on the analysis performed, surgeries or operations are performed if a requirement arises.

The most critical task among all others is the analysis of health data of the patient. This data is heterogeneous in nature i.e., comes from different sources and includes both structured and unstructured data of various types like text, image, sound, etc. Hence, in order to deal with such big data, various machine learning algorithms are developed.

The diagram below shows the various commonly used machine learning algorithms.

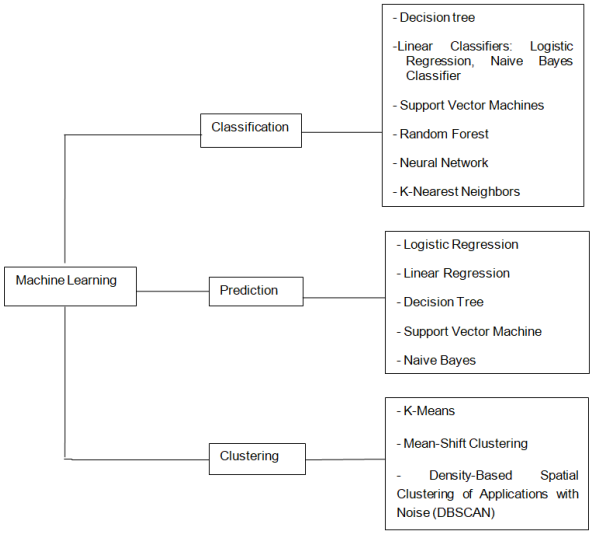

Fig. 3. Classifications of Machine Learning Algorithms.

IV. FEATURE SELECTION

Nowadays data scientists belonging to almost all domains need to work with enormous data sets which not only contains a large number of observations but also hundreds or even thousands of features to be dealt with. Such datasets having the number of features almost similar to the number of observations are said to be suffering from Overfitting. In order to overcome the overfitting condition, the number of features needs to be reduced by using various dimensionality reduction techniques available. These dimensionality reduction techniques are also called Feature Extraction methods.

Feature Extraction refers to a reduction in the number of features in the dataset by summarizing or merging the data of two or more features into a single feature, eliminating the original feature. This reduces the dimension of the dataset without any loss of information. There exists another procedure to reduce the number of features in a healthcare dataset, commonly called Feature Selection. Feature selection ranks existing features based on their importance and discards less important ones.

Various methods for feature extraction are as follows. Principal Component Analysis (PCA) is one of the most commonly used linear feature extraction techniques based on unsupervised learning. In this technique, the entire dataset is protruded on a set of perpendicular axes, and each axis is ranked based on the relevance of the feature it represents. Post this ranking of axes, PCA tries to maximize the variance and minimize the reconstruction error by observing the distances pairwise. Following this method for all features, PCA attempts to

construct an ensemble of only those features which can represent the entire dataset without loss of generality. As it is an unsupervised learning algorithm, it may sometimes lead to misclassification of data. Independent Component Analysis (ICA) is also a dimensionality reduction method. It accepts a certain number of independent components as input and classifies them into a relevant feature or noise. If both linear and non-linear dependencies between 2 features become zero, then they are said to be independent. This method of linear dimensionality reduction is often used in medical applications such as EEG and fMRI analysis to distinguish between relevant signals and noise. Linear Discriminant Analysis (LDA) is a supervised learning dimensionality reduction algorithm. LDA feature extraction method can be applied only on the gaussian distribution of data and may prove to be futile when applied on non-gaussian data. It aims to minimize the spread of data within a class and enlarge or maximize the distance between means of each class. LDA is a good classification choice when data is projected in a lower-dimension space. Another feature extraction technique is ***Autoencoder.*** Autoencoders refer to an ensemble of Machine Learning algorithms that can be used for feature extraction in a dataset. Autoencoders differ from other dimensionality reduction methods due to the fact that Autoencoders apply non-linear transformations to protrude data from a high dimension space to a lower dimension space. The construction of an Autoencoder comprises of 2 vital units: Encoder: It accepts the input data and compresses it, in order to discard any noise present in the data set. The output generated by this stage is referred to as 'bottleneck' or 'latent-space'. Decoder**:** It accepts as its input the output of the encoder unit and attempts to regenerate the original input provided to the encoder using the encoded latent space. If all the input features are found to be independent of each other, then it becomes very difficult for the Autoencoder to encode and decode the input data into a lower-dimensional space.

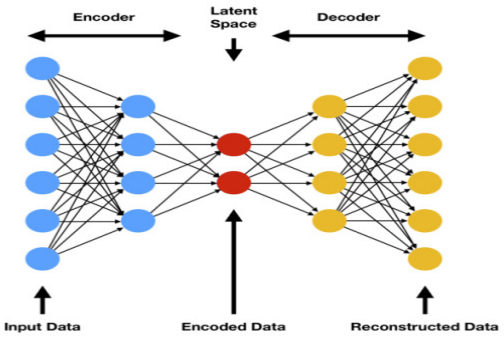

Fig. 4. Autoencoders.

In the healthcare domain, in order to detect if a person is a victim of a disease or not, several features or characteristics of the person needs to be taken into consideration. This information includes basic details of the person like age, gender, etc. and some disease-specific attributes. Let us consider heart disease for instance. There are 14 features selected for analysis after applying the feature extraction technique on 76 attributes. The selected features are mentioned in below Table 1.

Table 1: Heart Disease Prediction Features.

Feature	Definition
Age	A patient's age
Sex	A patient's sex
Cp	Chest pain experienced
Trestbps	The patient's resting blood pressure
Chol	Patient's cholesterol measurement in mg/dl
Fbs	The patient's fasting blood sugar (>120mg/dl)
Restecg	Resting electrocardiographic measurement
Thalach	Patient's maximum heart rate achieved
Exang	Exercise-induced angina
Oldpeak	ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot)
Slope	The slope of the peak exercise ST segment
Ca	Number of major vessels (range 0-3)
Thal	A blood disorder called thalassemia
Target	Heart disease detection result

V. ALGORITHMS

Till today, many researchers have conducted several experiments for disease prediction using several machine learning algorithms. Some of the commonly used classification algorithms are as follows.

Decision Tree There are two steps involved in this approach. They are the construction of a decision tree and applying the data set on the tree to classify it. There exist several decision tree algorithms that are commonly used. These commonly used algorithms include CART, ID3, C4.5, J48, etc. The overall concept is to build a tree that strikes a balance between flexibility and accuracy [3, 13, 15].

The Logistic Regression is also known as the sigmoid function is a machine learning classifier that is used when the feature to be predicted is binary. It helps in the easy representation of data in graphs. It also provides high accuracy. In this algorithm, the input data is first trained with the expected output and the best and approximate coefficients, to construct the logistic regression equation, are calculated [10]. Another classification technique is the Naive Bayes classifier. It uses the Bayes theorem to classify a tuple into a particular class or label. The Bayes theorem is given by,

P(Label|Features) = P(Label)*P(Features|Label) / P(Features)

Where P(label) gives the prior probability of a label. P(features|label) is the prior probability that a given feature set is classified as a particular label. P(features) is the prior probability that a specific feature set has occurred.

Support Vector Machine (SVM) is one of the supervised data-classification algorithms that can categorize a new data element to one of the existing labeled categories. SVM is also a binary classifier.SVM plots the trained data points onto a plane and draws a line, known as a hyperplane between the two categories. After this, when a new data point is introduced for testing, it is said to belong to the category based on which side of the hyperplane it falls. SVM has been lucratively implemented in many applications from several domains such as image recognition, medical diagnosis, and text analytics. Random Forest [6] as the name suggests, constitutes a forest of trees, i.e., several decision trees are constructed from the training data points. The prediction capability or accuracy is estimated for each decision tree using the Gini index and the best solution is selected by majority voting or averaging the result. ***Neural Network*** [14] comprises of many neurons, organized in the form of layers, as its basic unit. Each neuron has a non-linear function and a weight associated with it. The input, Xi, is provided to the neurons layer, a function is applied to it and the result multiplied with the weight is passed on to all neurons in the next layer. The outputs provided by the nodes in the last layer of the network are combined to generate the overall outcome. It is then compared with the expected result and error, if any, is estimated. In order to minimize the error, the weights associated with the neurons are changed and the output is again computed. This process is continued until the output becomes error-free.

K-*Nearest Neighbors* (KNN) [8, 9] is yet another famous classification algorithm used because of its simplicity and accuracy. KNN algorithm takes all available cases in the dataset as its input and determines the class labels of new cases based on a similarity measure by measuring the distance between the already classified data point and the new data point. In KNN, K denotes the number of nearest neighbors of a data point. This number of neighbors plays a crucial role in the efficient working of this algorithm. If the class label of the feature which is to be determined is binary, then K is an odd number. For K=1, the algorithm called the nearest neighbor algorithm. This is the simplest case. When k>1, the following distance measures can be used. These measures are *Euclidean distance* given by √(∑(xi - yi)^2), *Manhattan distance* which is given by √(∑(xi - yi)) and *Minkowski distance* which is given by [∑(|xi - yi|)^q]^1/q. The *K-Means* [4, 11, 12] algorithm takes two inputs: a set of data points that are to be clustered and a parameter K denoting the number of clusters to be formed. Based on the input provided, K centroids are fixed. The distance between the data points and centroid is estimated for each data point and it is clustered with the centroid to which it is closest. Another algorithm commonly used is *Linear Regression.* There are two kinds of variables in a linear regression model: predictor X that is used to determine the class label of the output feature and output variable (Y) is the variable whose value we want to predict. To estimate Y using Linear Regression, we use the equation: $Y_e = \alpha + \beta X$ where Y_e is the estimated or predicted value of Y based on our linear equation. The objective of the linear regression method is to estimate the significant values of the parameters α and β such as to minimize the difference between the predicted value and the estimated value.

Other commonly used decision tree algorithms are Iterative Dichotomiser 3 (ID3): The ID3 algorithm takes up the entire dataset as its root node [15]. During each iteration, it loops through every new feature of the dataset and computes their information gain. Information gain, in simple terms, may be defined as the amount of relevant information or knowledge a particular feature encapsulates. It then selects only those attributes whose information gain value is the highest. This extracted feature set is then sub-divided on the selected features with the highest information gain to produce subsets of the data. The algorithm continues to recurse on each subset, considering only those attributes which were not previously selected. The stopping conditions of the recursion are

***Case 1*:** Each feature in the dataset belongs to a single class label. In such a case, the node is turned into a leaf node.

Case 2: There will be no more attributes to be selected for further sub-division, but there will be no feature belonging to the same class.

Case 3: There are no examples in the subset for further splitting. In the algorithm, the decision tree is modeled based on every non-terminal node that represents the selected attribute for splitting, and terminal nodes that represent the class label of the final subset.

C4.5 [15] also constructs decision trees from the training dataset, based on information gain, however in this algorithm the information gain needs to be normalized. The training data is a set of previously classified data points. Each data point refers to a p-dimensional vector, where the vector constitutes the values of all p features for that data point, including the class in which it falls. For each node of the tree, this algorithm selects that feature of the dataset which can split its set of unique values into subsets most effectively such that each subset is classified into one class or the other. The splitting criterion used in this algorithm is the normalized information gain. The feature which shows the highest normalized information gain is selected for sub-division. C4.5 continues to iterate in the same way on the partitioned subsets.

VI. TOOLS AND TECHNIQUES

The several platforms, tools, and techniques used for the implementation of machine learning algorithms are: ***RStudio*** is an open-source, integrated platform for development using the R programming language. RStudio comprises of a console and a syntax-highlighting editor. It also includes various tools for workspace management. RStudio can run on a desktop or a browser connected to the RStudio Server.

Another tool used is ***Apache Spark*** which is a unified cluster computing technology based on Hadoop MapReduce, for faster computations. The credit of faster computation of applications by Spark can be given to its in-memory cluster computing feature. Apache Spark is capable of handling a huge amount of work including batch processing, streaming, query processing, etc. all in a single platform.

Table 2: Tools and Techniques used in previous works.

References	Year	Task	Domain oriented	Algorithms Used	Tool Used	Data set/Source
[1]	2016	HDP	Yes	All-Possible MV Algorithm	KEEL (JAVA based open source apparatus	Cleveland Heart disease dataset
[2]	2016	HDP	Yes	KNN, ID3		Cleveland Heart disease dataset
[3]	2017	HDP	Yes	Random Forest	Apache Spark	Cleveland Heart disease dataset
[4]	2017	HDP	Yes	Improved K-Means, ID3	Hadoop MapReduce	Cleveland Heart disease dataset
[5]	2018	HDP	Yes	Random Forest, Naive Bayes	RStudio	Cleveland Heart disease dataset
[6]	2019	HDP	Yes	Ensemble method		Cleveland Heart disease dataset
[7]	2019	HDP	Yes	Random Forest	Apache Spark (MLlib, Cassandra)	Cleveland Heart disease dataset
[8]	2019	HDP	Yes	Decision tree using hybrid Association rules	RStudio	Cleveland Heart disease dataset
[9]	2019	HDP	Yes	GBT or Random Forest	Apache Spark (MLlib, Scala)	MIT BIH Arrhythmia database

VII. APPLICATION DOMAIN

Healthcare Data Solutions Big data helps in storing large amounts of data of all formats without any loss of information in a secure manner. This organized storage of data enables the healthcare service providers to access data with ease as and when required and take the right decision at the right time. *Tracking Patient Vitals* in operation theaters and ICUs is yet another advantage of big data analytics. The various devices with sensors plugged into the human body, continuously measure and keep records of characteristics such as blood pressure, heartbeat and respiratory rate, etc. *Anti-Cancer Therapy Using Big data* Cancer, being one of the biggest challenges to a long and healthy life all over the globe, is very difficult to treat. However, applying predictive analytics, previous health conditions, and behavioral patterns can be analyzed and chances of an individual likely to suffer from cancer can be predicted much earlier and appropriate measures can be taken before it is too late. Data accumulated and assimilated from a patient's medical history has also proved to be essential and helpful in identifying proper treatments causing minimum side effects on the patients.

Some other application domains are *Fraud Prevention and Detection* Using big data techniques to handle health-related information has eliminated the security threats which may arise due to human errors or the use of faulty instruments or any other third party who may have access to the data. ***Improved Hospital Administration*** Big data can conveniently maintain the workflow of multiple regulatory activities. It assists the healthcare service providers in handling huge amounts of data having different formats from multiple, diversified sources and provides insights on sketching down plans for upcoming troublesome situations like epidemics and natural calamities.

VIII. DISCUSSION

The purpose of this paper is to instil the importance of big data analytics among readers. It has presented a brief introduction to various machine learning algorithms that are used in previous researches by many data scientists to predict if a person is suffering or likely to suffer from a disease. The prediction made is based upon the identified relevant features using a suitable feature extraction method.

Among the various classification and clustering algorithms mentioned in section V, some of the algorithms which are more efficient than others are decision tree algorithms like C4.5 and ID3, logistic regression, SVM, KNN, and Naive Bayes. The advantages of using decision tree algorithms [15] are ease of implementation and flexibility i.e., allows adding new data. Additionally, C4.5 allows not only discrete but also continuous features and also handles missing data. Another efficient and quick algorithm is the logistic regression algorithm. It works most efficiently when only relevant features are included in the dataset. It does not require any additional computational resources thus making it easy to implement.

Support Vector Machine (SVM) is yet another algorithm that has proved to be highly efficient in handling the heterogeneity of healthcare data as it is capable of managing data in different formats like structured, unstructured, semi-structured, images, text, trees, etc. Works well with even unstructured and semi-structured data like text, images, and trees. Also, it is capable of handling high-dimensional data, and chances of overfitting are reduced.

Another classifier that can be used in the prediction of disease is the K-Nearest Neighbors. It requires only 2 parameters, the value of K and the distance measure. Also, it does not need any prior training before being applied to new data; hence the addition of new data at any time will not degrade its accuracy. Naive Bayes [15] classifier is yet another supervised classifier that outperforms most of the other algorithms when the dataset consists of independent predictors.

IX. CONCLUSION

This paper has presented a traditional survey on big data and it's usage in various sectors of the healthcare domain. It has presented a brief introduction to various commonly used linear and non-linear feature engineering methods such as principal component analysis, auto-encoders, etc which is required to identify

features or attributes which are relevant to the purpose of analysis and thus reduces the dimensionality of the data set making the analysis easier. Also, this paper has discussed various machine learning algorithms which include classification algorithms such as Decision trees (ID3, C4.5), linear classifiers, Support Vector Machines, etc, Prediction algorithms such as Naive Bayes, K-Nearest Neighbors, Random Forest, etc and Clustering algorithms such as K-Means, DBSCAN, etc. These algorithms are used in the healthcare domain for prediction or classification of potential diseases that may affect a person. This paper has also summarized some of the previous research works in the healthcare domain along with their accuracy in predicting the risk level of heart disease in a person applying various machine learning algorithms on features like age, sex, cholesterol, blood sugar, etc, and different tools used for implementing the same.

Lastly, it has also specified some other application domains of big data analytics in the healthcare sector. This survey encourages further research works in the healthcare sector in various domains including the aforementioned domains. These future researches can be done by combining various machine learning algorithms mentioned in this paper to achieve better results.

REFERENCES

[1]. Purushottam, Saxena, K., & Sharma, R. (2016). Effective Heart Disease Prediction System. *Procedia Computer 85*, 962-969.

[2]. Princy, T., & Thomas, J. (2016). Human Heart Disease Prediction System using Data Mining Techniques. *IEEE, 2016 International Conference on Circuit, Power, and Computing Technologies, Nagercoil*, 1-5.

[3]. Saboji, R. G., & Ramesh, P. K. (2017). A Scalable Solution for Heart Disease Prediction using Classification Mining Technique. *IEEE, International Conference on Energy, Communication, Data Analytics, and Soft Computing (ICECDS)*, 1780-1785.

[4]. Mane, U., & Tejaswini, Ms. (2017). Smart Heart Disease Prediction System using Improved K-Means and ID3 on Big Data. *IEEE*, *International Conference on Data Management, Analytics, and Innovation (ICDMAI)*, 239-245.

[5]. Karthick, D., & Priyadharshini, B. (2018). Predicting the chances of occurrence of Cardio Vascular Disease in People using Classification techniques within fifty years of age. *IEEE, 2nd International Conference on Inventive Systems and Control (ICISC)*, 182-1186.

[6]. Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked*, *16*, 100203.

[7]. Ed_daoudy, A. & Maalmi, K. (2019). Real-time Machine Learning for early detection of heart disease using the big data approach. *IEEE*, *International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS)*, Fez, 1-5.

[8]. Mohan, S. K., Thirumalai, C., & Srivastava, G. (2019). Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. *IEEE Access*, vol. 7: 81542-81554.

[9]. Alarsan, F. I., & Younes, M. (2019). Analysis and classification of heart diseases using heartbeat features and machine learning algorithms. *Journal of Big Data*, *6*(1), 1-15.

[10]. Mathew, T. E. (2019). A Logistic Regression with Recursive Feature Elimination Model for Breast Cancer Diagnosis. *International Journal on Emerging Technologies, 10*(3), 55-63.

[11]. Yadav, R., & Sharma, A. (2012).Advanced Methods to Improve Performance of K-Means Algorithm: A Review Clustering. *Global Journal of Computer Science and Technology, 12*(9), 46-52.

[12]. Chadha, A., & Kumar, S. (2014). An Improved K-Means Clustering Algorithm: A Step Forward for Removal of Dependency on K. *International Conference on Reliability, Optimization, and Information Technology –ICROIT,* 6-8.

[13]. Thariq Hussan, M. I., & Saini, H. S. (2019). Predictive Mining Model for Transactional Data Pattern using Probabilistic Based Decision Tree Model. *International Journal of Emerging Technologies*, *10*(3), 38–44.

[14]. Al-milli, N. (2013). Backpropagation neural network for prediction of heart disease. *Journal of Theoretical and Applied Information Technology*, *56*(1), 131–135.

[15]. Gawali, M., Shirwalker, N., & Kalshetty, A. (2018). Heart Disease Prediction System Using Data Mining Techniques. *International Journal of Pure and Applied Mathematics*, 499-506.

How to cite this article: Mishra, S., Pandey, M., Rautaray, S. S. and Gourisaria, M. K. (2020). A Survey on Big Data Analytical Tools & Techniques in Healthcare Sector. *International Journal on Emerging Technologies*, *11*(3): 554–560.