International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 9, No.6 (Nov-2020)
http://dx.doi.org/10.12785/ijcds/0906017

Taxonomy on Healthcare System Based on Machine Learning Approaches: Tuberculosis Disease Diagnosis

Priyanka Karmani[1], Aftab Ahmed Chandio[1], Vivekanand Karmani[1], Javed Ali Soomro[2], Imtiaz Ali Korejo[1] and Muhammad Saleem Chandio[1]

[1] *Institute of Mathematics and Computer Science, University of Sindh, Jamshoro 70680, Pakistan*
[2] *Centre for Physical Education, Health and Sports Science, University of Sindh, Jamshoro 70680, Pakistan*

Received 19 Jul. 2020, Revised 30 Aug. 2020, Accepted 19 Oct. 2020, Published 1 Nov. 2020

Abstract: This study enlightens the impact of Machine Learning algorithms and practices in the context of Healthcare Informatics. In the domain of Healthcare Informatics (HI), Machine Learning (ML) procedures have been classified into four classes named as ML-HI types, ML-HI approaches, ML-HI paradigms and ML-HI algorithms. In this study, we provide an overview of the state-of-the-art, the research challenges, and the forthcoming directions, specifically driven to the diagnosis of Tuberculosis (TB) disease. Moreover, we introduce our proposed framework for TB diagnosis disease based on ML. We emphasized the strengths and weaknesses of the studied methods facilitate to the aid analysis community to pick the suitable technique to use within the Healthcare Informatics domain.

Keywords: Machine Learning; Healthcare Informatics; Tuberculosis.

1. INTRODUCTION

Health is one of the most significant traits in humans' life. Without sound health, it is hard to survive for an individual in this world. The international body World Health Organization (W.H.O.) has defined the term Health as "a condition of complete physical, mental and social well-being" [1]. Healthcare Informatics (HI) acts a vital role in the improvement of healthcare sectors. The U.S. National Library of Medicine has defined Healthcare Informatics as "an interdisciplinary study of the plan, advancement, acceptance, and implementation of modern technological tools and techniques in the provision, supervision, and formation of health care amenities" [2]. Healthcare Informatics aims to improve healthcare through any amalgamation of greater quality, greater efficiency (i.e. high availability and low cost), and novel prospects. The tools of Healthcare Informatics include computers, medical guiding principle, proper clinical terminologies, and information and communication systems [3-4]. For the management of the patient's health (either an individual or a group therapy), Healthcare Informatics makes use of computational intelligence [5]. Overall, Healthcare Informatics is targeted to improve the overall effectiveness of patient's care delivery by ensuring that the data generated is of high quality [6].

With the rapid technological advancement in every aspect of a life, health measures are also being switched from manual to digitized form. Machine Learning (ML) is an emerging field in computer science, and Healthcare Informatics is among the utmost challenges [7-8]. Machine Learning is defined as "a field of study which provides computers the capability to learn deprived of being explicitly programmed" [9]. Machine Learning has been evolved from the study of pattern recognition and computational learning theory in Artificial Intelligence. The ultimate objective of Machine Learning is to develop such algorithms, which are capable of learning and improving over time and can be used for predictions. Now-a-days, Machine Learning is vastly being utilized in the domain of healthcare. Machine Learning targets to provide medical decision support for patient management [10]. However, Machine Learning assures the enhancement in the development of Healthcare Informatics, and the enhancement in our capability to tailor care to the specific physiology of an individual, the perception of such ML-based systems into real hospitals is in its embryonic stage [10]. As our society is infected with enormous dreadful diseases, our ultimate goal is to provide solutions to reduce the impact of these diseases and save human lives to a possible height. According to W.H.O., Tuberculosis (TB) is considered one among the

* A preliminary version of this paper [108] was presented at the International Conference on Intelligent Technologies and Applications, INTAP-2018, Springer CCIS (932), Islamia University Bahawalpur, Pakistan, October 23-25, 2018

E-mail: cspriya66@gmail.com; chandio.aftab@usindh.edu.pk

top-ten dreadful syndromes around the globe [11]. TB is an infectious disease, usually caused among the resource-poor communities. Initially, human breathing organ i.e. Lungs are effected by the TB bacteria. However, it also influences other human organs [12].

In this paper, we enlighten the impact of ML algorithms and practices in the context of Healthcare Informatics (HI). We categorize Machine Learning Healthcare Informatics (ML-HI) into: (a) two types such as aML and iML; (b) two approaches such as regular ML and ensemble ML; (c) three learning types including supervised ML, semi-supervised ML and unsupervised ML. Furthermore, we provide a thorough explanation of ten ML algorithms commonly used in HI, which were found during literature review. Such algorithms includes Decision Tree (DT), Support Vector Machines (SVMs), Naive Bayes, Regression (Linear and Logistics), Neural Networks (NNs), k-Nearest Neighbour (kNN), k-Means Clustering, Genetic Algorithms (GAs), Deep Learning, and Ensembles (Bagging and Adaboost). We highlight the strengths and weaknesses of ML-HI state-of-the-art in order to help the ML-HI research community to select the appropriate ML algorithm in order to apply in the healthcare domain. We also provide ML research challenges and future directions in aspect to tuberculosis disease diagnosis. In the last, our proposed framework for TB disease diagnosis is introduced, which is based on ML. In section 2, we discuss Machine Learning in the perception of Healthcare Informatics in a great detail. In section 3, 4, 5 and 6, we describe the ML-HI types, approaches, paradigms, and algorithms, respectively. In section 7, we provide the literature review about Machine Learning algorithms applied in the domain of Healthcare Informatics. While in section 8, we highlight the merits and demerits of the discussed algorithms and discuss about the Tuberculosis disease diagnosis using Machine Learning techniques. In the last, the conclusion is provided in Section 9.

2. MACHINE LEARNING IN HEALTHCARE INFORMATICS (ML-HI)

The health of an individual can be effected by a variety of factors including age, gender, genetic history, family background, birth-time problems, individual behavior, environmental impact, living settings, quality of education, workplace atmosphere, hygiene of food and water, health care amenities, racial factors, agricultural factors, global warming, tension and depression, loneliness, poverty, poor resource communities, financial issues, etc. Moreover, the factors of the general health of the inhabitants can be intellectualized as rainbow-like layers of influence [13]. With the passage of the time, our lives have been revolutionized by modern technology to a great extent. Today, almost in every sector of life, technology has been deeply rooted whether it is military, industries, education, medical and health sector and even homes, etc. Healthcare Informatics (HI) is the vigorously developing modern arena that deals with the medical and health data by integrating computer science and information technology. In hospitals, doctors approach enormous amount of data on patients, however no time and no apparatuses to deal with that data. The solution to this challenge is intelligent medical decision-making systems; which are capable to envision the data and making predictions to cure the patient [14]. Intelligent solutions provide the humans doctors different tools and techniques in order to advance the Healthcare Informatics and helps to treat the patients in a more erudite manner.

At present, Machine Learning (a sub-domain of Artificial Intelligence) is widely being applied in the domain of Healthcare Informatics [15]. Machine Learning was initially intended and used to scrutinize medicinal datasets. Machine Learning makes available numerous crucial apparatuses for intelligent data analysis. Contemporary hospitals are well resourced with monitoring and auxiliary data collection devices, where data is congregated and pooled in huge information systems. Machine Learning is compatible for analyzing health-related data [16]. In the hospitals, medical data about accurate diagnosis are available; only there is a need to input the patient data with accurate diagnostic values into the computer program in order to execute a learning algorithm. The knowledge about the medical diagnosis can be spontaneously derived from the history of the patients dealt in the past. The resulting classifier can be utilized:

- to help the doctor while handling new patients in order to enhance the investigative speed, precision, and consistency [16],
- to prepare understudies or doctors (non-expert) to analyze patients in a symptomatic issue [16].

Machine Learning has been deeply rooted in almost every sector of Healthcare Informatics including disease diagnosis [16], health monitoring and management [17], fraud detection in health insurance [18-19], and many more. Here, we define four different layers of ML for HI including the types of ML being applied in the domain of Healthcare Informatics, two different approaches, learning types and most commonly applied ML algorithms in Healthcare Informatics. Fig. 1 shows the layers of ML-HI.

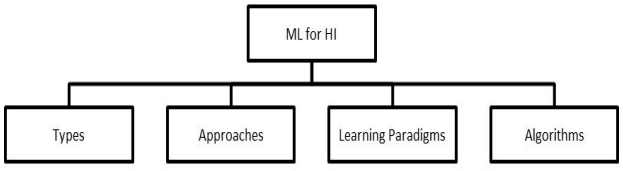

Figure 1. Four layers of ML for HI

A. Machine Learning Healthcare Informatics (ML-HI) Types

We define two types of the Machine Learning in the notion of Healthcare Informatics: iML (interactive Machine Learning) and aML (automatics Machine Learning).

1. iML (interactive Machine Learning)

   iML (interactive Machine Learning) is the type of ML in which human data is involved throughout the learning process. In the domain of Healthcare Informatics, we often deal with small, complex, ambiguous and messy data, where iML comes in the action. iML can be defined as "algorithms which correlate with agents (usually humans) and boost their learning behavior through these correlations" [20-21]. iML (i.e., often known as the human-in-the-loop approach) aids in resolving problem which are computational hard (i.e., NP-Hard problems). In these kinds of problem, human involvement can decrease an exponential search space by making use of heuristics. In this manner, iML reduces the complexity of algorithms. In the healthcare informatics, the doctor-in-the-loop approach is being applied for solving problems including protein folding, k-anonymization of health records, subspace clustering, etc. [20-21]. In iML approach, the updates are hasty, attentive and progressive, which enables the users to intuitively look at the effect of their activities and modify consequent inputs to acquire desirable results. Moreover, these quick correlation cycles enable users with practically no machine-learning mastery can guide machine-learning practices through ease experimentation or centered experimentation with sources of info and yields [22].

2. aML (automated Machine Learning)

   aML (automated Machine Learning) is the type of ML in which a human agent is excluded throughout the learning process, dissimilar to iML. In the domain of healthcare informatics, while dealing with large health data sets i.e. Big Data, aML comes in action. For example, for a nurse-scheduling task, a tool is obligatory which automatically selects the required parameters in order to progress healthcare delivery. Thus, aML can be defined as "algorithms which don't correlate with agents (usually humans) and are completely self-automated" [23]. Dealing with the massive data often involves the huge number of users, enormous complex programming frameworks, and huge-scale diverse computing and storage. The development of these tools includes various dispersed design choices. In this case, automation is desirable because dealing with big data is beyond human ability. For example, in the Healthcare Informatics, scheming of tools for clinical analysis of patient's data set for better healthcare delivery, implicates numerous tunable configuration parameters. These parameters are frequently indicated and hard-coded into the product by the developers [23]. aML is also known as a human-out-of-the-loop approach. aML is directed towards selecting the best available algorithm and the relevant parameters in order to solve a task for the given data set. aML aims to enhance the quality of the product as well as human efficiency [23]. The various platforms have been developed to tackle through the different stages of aML. Examples may include Auto-WEKA [24], Auto-sklearn [25-26], TPOT [27-28], ATM [29], etc.

B. Machine Learning Healthcare Informatics (ML-HI) Approaches

We define two different approaches of machine learning applied in the healthcare informatics. These are: regular ML and ensemble ML.

1. Regular ML

   Regular ML is an approach which simply refers to the machine learning algorithms applied in a particular domain. Machine learning algorithms including support vector machine, neural network, decision Tree, etc. all falls in regular ML approach. Regular ML approach is widely being used in the domain of healthcare informatics, for example, diagnosing the disease using regular ML algorithms [30].

2. Ensemble ML

   Ensemble ML is another approach that is widely being applied in the domain of healthcare informatics, for instance, disease diagnosis using ensembles and the results are far better than the regular ML approach. Ensemble ML approach combines several learning algorithms in order to acquire better prognostic performance than the single learning algorithm [31]. Commonly used ensembles methods may include Bootstrap aggregating (Bagging), and Boosting [31]. In the field of Proteomics, Neuroscience and medical diagnosis, ensemble classifiers have been efficaciously applied. For example, detection of Neurocognitive disorder including Alzheimer or Myotonic dystrophy by manipulating MRI datasets [32-34].

C. Machine Learning Healthcare Informatics (ML-HI) Learning Paradigms

We discuss three different types of learning paradigms commonly being used in the field of healthcare informatics. These are: Supervised ML, Unsupervised ML, and Semi-supervised ML, as shown in Fig. 2.

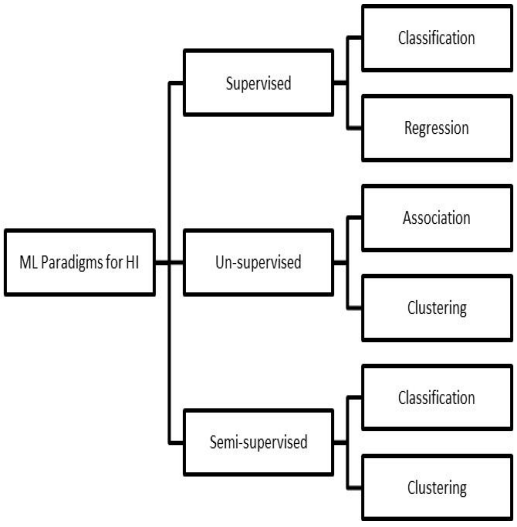

Figure 2. Learning paradigms of ML for HI

1. Supervised Machine Learning

Machine learning algorithms use a variety of data set to manipulate according to the system or scenario being considered. However, each instance in the data set is signified by identical set of features including binary, continuous, and categorical. When these instances are provided with acknowledged tags, the learning is known as supervised Machine Learning (SML) [35]. SML basically defines a function (from the labeled training data) which relates the input to the corresponding output based on training examples, where each training example consists of a pair of input/output values [35-36]. Common algorithms fall into this category include K-nearest neighbor (KNN), Decision Tree, Support Vector Machines, etc. SML is further categorized into;

- Classification – is a SML technique where the output variable is given values that are equal and opposite. It splits the training data into labeled target classes. It foretells the target class for each data points. For instance: in disease diagnosis, the patient data can be divided into two categories, i.e., Healthy (H) and Infected (I). According to the nature of patient's disease, the patient will either fall in any one category. Classification techniques are of two types namely binary classification and multiclass classification [37].
- Regression – is another SML technique which is applied to identify the function that exhibits the association between the variables (i.e. one dependent variable and one or more independent variables). Regression can be further classified into two types namely Linear Regression and non-linear (or logistics) regression [37].

SML is extensively being utilized in Healthcare Informatics due to its dominant characteristics. Parmar et al. investigated 12 different SML algorithms along with 14 different feature-selection procedures in order for radiomic based survival forecast. About 440 radiomic features were mined from pre-cure 464 computed tomography imageries of patients infected with lung cancer [38]. SML permits to remove undesirable results by boosting the appropriate results relevant to the target variables. However, SML techniques are time-consuming and required technical expertise.

2. Un-Supervised Machine Learning

Unsupervised Machine Learning (UML) refers to the learning without any supervision. It makes use of data which is not verified. It comprehends the hidden patterns from a dataset deprived of reference to labeled results. It is used to learn the core architecture of the data [39]. UML has a unique characteristic that the outcomes in this approach are not limited. Additionally, UML can be divided into the following types;

- Clustering – is an UML technique, which divides the data set into smaller groups, known as clusters. The division is made on the basis of similarities among the data elements. Each data element within a cluster has same characteristics, while different to other clusters. It reveals hidden or unknown patterns by applying unguided search and signifies a data concept as an outcome. Clustering can be of various types i.e. partitioned clustering, hierarchal clustering, density-based clustering, etc. [40]
- Association – is another UML technique, which is utilized to discover the recurrent patterns and stimulating associations with in a dataset in the data warehouse by applying a set of rules. Association is basically rule learning methodology because it learns via set of relational rules. In the health informatics, association is applied to discover the associations among syndromes, fitness condition and indications. [41].

UML is a significant approach in the perspective of Healthcare Informatics. It is applied in the estimation of illnesses since it has no predetermined conditions due to the unlabeled data. It has been applied for the prediction of medication effects and type-2 diabetes detection. However, its use is restricted due to the diverse outcomes, assorted data, logical biases, and haphazard inaccuracies [42].

3. Semi-Supervised Machine Learning

It is an intermediary between supervised and unsupervised machine learning. Since supervised machine learning entails intricate data and algorithms which deduce the outcomes after their comparison, making it an expensive approach to apply. On the other hand, unsupervised Machine Learning is low-cost since it

deals with unlabelled data. However, the outcomes can't be validated in this approach due to the unlabeled data. To overcome this lacking, semi-supervised machine learning has been introduced. In this type, an algorithm learns by making use of both labeled as well as unlabeled data. However, the labeled data is relatively in small ratio. This type is significant to apply when there is no sufficient labeled data available [43]. Further, it combines the flavour of both supervised and unsupervised machine learning techniques i.e. Classification and Clustering. This type is broadly applied in the healthcare informatics. Wang et al. applied semi-supervised machine learning to mine the diagnosis and examine the results from unstructured text in electronic health records [44].

D. Machine Learning Healthcare Informatics (ML-HI) Algorithms

From literature review, we found the following common algorithms of Machine Learning used for Healthcare Informatics, as shown in the Fig. 3.

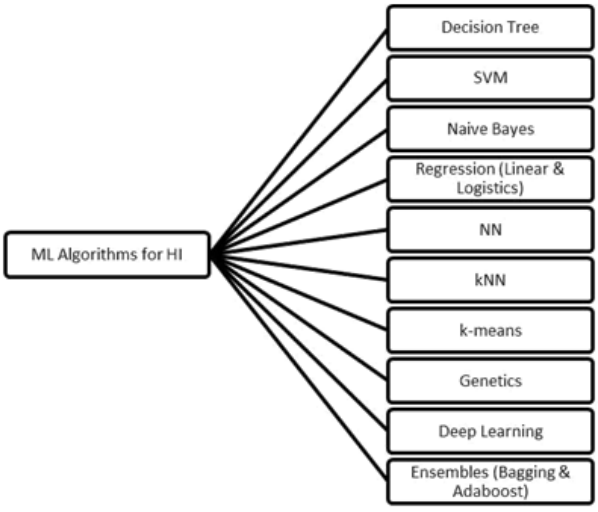

Figure 3. Common Algorithms of ML used for HI

1. Decision Tree (DT)

Decision Tree (DT) is a most commonly applied Machine Learning algorithm. It is defined as a top-down hierarchical structure which consists of three different types of nodes i.e. a root node (top-most node), internal nodes (non-leaf nodes), and terminal nodes (leaf nodes). DT structure resembles to the traditional Binary Tree. In DT, each inner nodule performs a test on an attribute, each branch indicated the result of the test, and each terminal nodule grasps a class label. This is how the algorithm makes a decision [45-46]. This algorithm falls under the umbrella of supervised Machine Learning. It is frequently used to solve classification and regression problems. It deals with different types of data including categorical as well as numerical data. It can manipulate multidimensional data [46]. Their interpretation of learned data in hierarchal shape (Tree-like structure) is instinctive and very easy to adapt by human being. These algorithms are very fast and have good precision rate. DT algorithms are successfully being applied in many fields including health informatics, industries, commercial analysis, astrophysics, etc. There are many popular DT algorithms C4.5, ID3, CART, CHAID, J48 and MARS [46-47]. DT algorithms are widely being applied in the Healthcare Informatics [109]. Tayefi et al. applied DT algorithm for the prediction of coronary heart disease [48]. Abdar et al. applied Boosted C5.0 and CHAID DT algorithms for early detection of liver disease [49]. Shouman et al. made use of J4.8 DT algorithm for the heart disease diagnosis [50].

2. Support Vector Machines (SVMs)

SVMs were initially introduced by Vapnik et al. [51-52]. SVMs fall into the category of Supervised Machine Learning and it is used for classification, regression and even other tasks. SVMs are provided a labeled training data set, the algorithm intends to create a hyper plane (i.e. a separator line) which splits the data set into the pre-defined classes in a manner relevant with the training samples. And any misclassification attained while the training phase of the algorithm, would be reduced by this separation as it defines the decision margin [53]. SVMs create the hyper plane by making use of margins and support vectors. Margin is defined as the distance between hyper plane and contiguous data points, where the data points are termed as support vectors. SVMs algorithms are implemented by using a mathematical function called Kernel. It takes the data as input and transmutes it into the desired output. SVMs kernel includes linear, non-linear, polynomial, sigmoid and radial basis function. Gamma is a regulating consideration of SVMs kernels [54-55]. SVMs are widely being applied in the domain of Healthcare Informatics. Polat et al. utilized least square SVM along with the combination of generalized discriminant analysis in order to diagnose diabetes disease [56]. Magnin et al. applied SVM algorithm to differentiating patients effected with Alzheimer's disease (AD) from aging controls [57]. Huang et al. utilized hybrid SVM approach to build a predictive model for the diagnosis of breast cancer [58].

3. Naïve Bayes

Naive Bayes is an effective and efficient Machine Learning algorithm which is methodically based on Bayesian theorem. It relies on conditional independence, which means that the occurrence of a specific trait value in a class is not linked to the occurrence of values of other traits [59]. Naïve Bayes is frequently being used in the Healthcare Informatics. Kazmierska et al. studied Naïve Bayes algorithm in the assessment of patients' risk of cancer deterioration or improvement after radiotherapy

Pattekari et al. introduced a forecast system based on Naïve Bayes for heart disease diagnosis [61]. Bhuvaneswari et al. highlighted the use of Naïve Bayes approach in the medical care and acknowledged it as best decision support system [62].

4. Regression (Linear & Logistics)

In the domain of Healthcare Informatics, the two most commonly applied regression models are Linear Regression and Logistic Regression. Linear Regression defines a relation among variables (i.e. one dependent and one/more independent variable). It has two types namely Simple Linear Regression (i.e. only one independent variable); and Multiple Linear Regression (i.e. more than one independent variable) [63]. Linear function is the basis for the formation of Linear Regression. Linear Regression discovers a route, estimates perpendicular distances of the data points from the route (or line) and reduce totality of square of perpendicular distance. Both the dependent as well as independent variables are already known in Linear Regression, the only task is to generate a line that show a relationship among those variables [63]. Logistics Regression (or non-Linear Regression) is another approach which takes binary dependent variables. Binomial (i.e. only two possible outcomes) and Multinomial (i.e. more than two outcomes) are its two types. It can deal with categorical data [64-65]. Regression models are extensively being used in the Healthcare Informatics arena. Thirumalai et al. applied Linear Regression approach for the decision making in Breast Cancer Type I Skin disease [66]. Saleheen et al. carried out a study on coronary heart disease using linear and Logistic Regression [67].

5. k-Nearest Neighbor (kNN)

kNN is one of the most easiest Machine Learning algorithm. kNN classifies the instances based on the nearby neighbors in the feature space. More specifically, it learns the unrevealed data point by means of the already identified data points i.e. nearest neighbor, and then classifies the data points conferring to the polling scheme. The k in kNN denotes the figure of nearby neighbors that the algorithm will consume to make the prediction. It is known as instance-based (or lazy) learning [68-69]. In the domain of Healthcare Informatics, kNN is commonly applied. Chen et al. used fuzzy kNN based approach for the diagnosis of Parkinson's syndrome [70]. Jabbar et al. applied a combinational approach of kNN and genetic algorithm for the classification of heart syndromes [71].

6. k-Means Clustering

k-Means is one of most prominent clustering algorithm. It falls under the category of partition-based classification [72]. k-Means algorithm use to update the cluster centroids, which is signified by means of center of data points, via computational iterations. The iterations keep continuing till certain norms for convergence is encountered [73-75]. k-Means algorithm is broadly being utilized in the arena of Healthcare Informatics due to its iterative process, compute-rigorous calculations and accumulating native outcomes in a similar setting [75]. Zheng et al. used a hybrid approach of k-Means and SVM for the feature extraction and diagnosis of breast cancer [76]. Escudero et al. applied k-Means method for the classification of data features of Alzheimer syndrome into pathological and non-pathological categories [77]. Balasubramanian et al. used clustering method to examine the influence of ground water on human health [78].

7. Genetic Algorithms (GAs)

Genetic Algorithms (GAs) were developed by John Holland in 1970. The GAs are very much similar to the natural herbal evolution. The basic theme is to utilize the power of evolution in order to solve the optimization problems. GAs are basically heuristic search techniques primarily based at the evolutionary thoughts of herbal selection and genetics. They signify an intellectual exploitation of a random search which is used to solve optimization problems. GAs are based on the Charles Darwinian theory of evolution or most commonly known as the survival of the fittest. The one who is healthier will survive the most. Initialization, selection, crossover, mutation are the steps involved in the genetic algorithm process [79-80, 110]. In the Healthcare Informatics, GAs have been deeply rooted. Guo et al. applied Genetic Algorithm for the optimum placement of sensors in order to monitor the health [81]. Shah et al. applied Genetic Algorithm in combination with other data mining techniques for cancer-gene search [82]. Yan et al. proposed GA-based system to choose the critical medical features vital to the heart diseases diagnosis [83].

8. Neural Networks (NNs)

Neural Networks (NNs) are recently arisen computational modeling approaches, which are widely being accepted in many fields in order to solve multifaceted real-world complications. NNs are also called as Artificial Neural Networks (ANNs), since it working mechanism is based on artificial neurons. It is mathematical depiction of human neural structural design, replicating human learning process and generalization aptitudes. A Neural Network is made up of a series of artificial neurons (nodes), organized in a layered structure. Each node in one layer is associated to each node in next layer via a weighted link. Total number of layers and number of nodes in each layer varies according to the complication of the system being considered [84-86]. Nodes at input layer get the data; transmit them to the nodes at first hidden layer via the weighted associations. Then, the data are processed and the end result is transmitted to the nodes of following layer. Eventually, nodes in the last layer deliver the net result. NNs have been deep-rooted in the arena of

Healthcare Informatics [86]. Abbass et al. proposed an evolutionary ANN approach for the diagnosis of breast cancer [87]. Raith et al. recommended an ANN based model to classify the dental cusps with sufficient accuracy [88]. Bhardwaj et al. applied genetically optimized NN approach to classify breast cancer in benign or malignant tumor [89].

9. Deep Learning

Deep Learning is defined as an emerging class of Machine Learning algorithms, which utilizes a group of numerous layers of non-linear processing units for the purpose of feature mining and alteration. Output attained form the preceding layer acts as an input in the next consequent layer. Deep learning algorithms may learn in any manner i.e. supervised, unsupervised or semi-supervised. It is also known as hierarchal learning, since it takes in various levels of interpretations representing numerous levels of perception. And all these levels form a hierarchal structure [90-93]. Deep Learning comes in different architectures including Deep Neural Network, Deep Auto-encoder, Deep Belief Network, Deep Boltzmann Machine, Recurrent Neural Network, Convolutional Neural Network etc. Deep Learning is extensively being utilized in the domain of Healthcare Informatics [94]. Liu et al. proposed a Deep Learning based early diagnostic system for the Alzheimer disease [95]. Acharya et al. applied Deep Convolutional Neural Network approach in order to automatically detect a regular and MI ECG beats [96].

10. Ensembles (Bagging & Adaboost)

Bagging and Adaboost is the most frequently used Machine Learning ensembles in Healthcare Informatics. Bagging (Bootstrap Aggregating) is a technique which builds various feeble learners for numerous learning datasets developed by re-testing from a given dataset. Bagging aims to reduce the variance and chance of over fitting [97-98]. AdaBoost is the method that alters the likelihood appropriation of learning data with the goal that frail learner emphases on the data to which other frail learners don't change enough. AdaBoost provides good precision rate [97-98]. Chau Tu et al. proposed a bagging approach to classify the cautionary symptoms of heart disease [99]. Morra et al. did a comparative study of Adaboost and SVM for the detection of Alzheimer syndrome by means of automated Hippocampal Segmentation [100].

3. LITERATURE REVIEW

In this section, we address the key contributions of the several researchers in the context of Machine Learning Healthcare Informatics (ML-HI). The highlights of their work are also shown in Table 1.

Ilhan, H. O., & Celik, E. [101] have focused on Malignant Mesothelioma disease wherein they compared different types of Artificial Intelligence methods for effective Malignant Mesothelioma disease classification. Malignant Mesothelioma disease is one amongst the foremost dangerous types of cancer triggered by asbestos mineral. The common symptoms of the disease can be found in human body as constant pain and progressive shortness of breath can lead the human to expire in a short interval. In [101], Support Vector Machine (SVM), Neural Network (NN) and Decision Tree (DT) are selected as Regular ML techniques; while Bagging and Adaboost are selected as Ensemble ML techniques. Authors used 324 data samples having 34 features in the study. K-fold cross-validation technique is implemented to calculate the performance of the algorithms with diverse K values. Moreover, two primary metrics such as accuracy and time complexity are considered to evaluate the technique. The results revealed that linear SVM provided 100% accuracy. Furthermore, DT and linear kernel SVM are simpler algorithms consume less computational time compare to the Bagging DT. On the other hand, DT fails on handling big data problem. In the conclusion, linear SVM is considered as the best algorithm to diagnose Malignant Mesothelioma disease [101].

Yahiaoui, A., Er, O., & Yumusak, N. [102] have focused on Tuberculosis (TB) disease wherein they applied a Machine Learning method for effective Tuberculosis disease diagnosis. TB is an infective syndrome initiated by a bacillus known as Mycobacterium tuberculosis. It causes death when left untreated or improperly treated. Thus, early stage diagnosis and treatment of TB is crucial. At present, because of high classification and identification rates, specialist systems became a vital tool in identification of the syndromes. In [102], a Machine Learning technique named as Support Vector Machine (SVM) was used for early stage diagnosis of TB. A recognition system was developed using patients' attributes attained in the patient lab reports of a local hospital. The system was then tested for its performance and accuracy. The results shown that SVM could be considered as the best diagnosing method for Tuberculosis syndrome with 96.68% classification accuracy rate and low computational time [102].

Er, O., Yumusak, N., & Temurtas, F. [103] provided a comparative study for the effective diagnosis of different chest diseases using diverse machine learning techniques. Chest diseases include chronic obstructive pulmonary, pneumonia, asthma, tuberculosis, and lung cancer. In [103], machine learning techniques included multilayer neural networks, probabilistic neural networks, learning vector quantization, and generalized regression neural networks were used. A dataset of 357 samples with 38 features was used in this study and prepared by using patient's lab reports obtained from Diyarbakir Chest Diseases Hospital (Southeast Turkey). The experimental results shown that neural network structures could be efficaciously used for the diagnosis of chest syndrome. PNN is considered as the best method with average

classification accuracy. MLNN with two hidden layers is better approach than MLNN with one hidden layer for chest disease diagnosis [103].

Er, O., Yumusak, N., & Temurtas, F. [104] used the artificial immune system for the diagnosis of chest diseases including chronic obstructive pulmonary, pneumonia, asthma, tuberculosis, and lung cancer. A dataset consisting of 357 infected samples having 38 features were used in this study and prepared from a chest syndromes hospital's databank using patient's lab reports in Turkey. 93.84% classification accuracy was obtained with artificial immune system and thus it is acknowledged that artificial immune system could be successfully used for the chest diseases diagnosis [104].

Weng, C. H., Huang, T. C. K., & Han, R. P. [105] applied different artificial neural network classifiers for the diagnosis of syndromes including breast cancer, liver patients, vertebral column, and heart disease. The key purpose of this research was to examine the performance of diverse classifiers, including individual classifiers involved in an ensemble classifier and solo classifiers. Different evaluation benchmarks were used to examine the performance of these classifiers with real-world datasets. Statistical testing was also used to assess the implication of the difference in performance among the three classifiers. The statistical testing results show that an ensemble classifier carries out better than an individual classifier within an ensemble. On the other hand, the solo classifier doesn't perform worse than the ensemble classifier assembled with the same size training dataset [105].

TABLE 1. THE CONTEXT OF MACHINE LEARNING HEALTHCARE INFORMATICS (ML-HI)

Ref:	Disease	Dataset	Algorithms	Result
[101]	Malignant Mesothelioma (i.e., lungs cancer caused by asbestos mineral)	n=324 (samples) f= 34 (features)	Regular ML: SVM, DT, NN Ensemble ML: Bagging with DT, Adaboost with DT	Linear SVM provided 100% accuracy and less computational time
[102]	Tuberculosis (TB) (i.e., caused by a bacillus, mycobacterium)	n=150 (samples) f= 38 (features) (i.e., n2=50 infected samples and n3=100 healthy samples)	Support Vector Machines (SVMs) with linear kernel	SVM is considered as the best method with 96.68% success rate and low running time
[103]	Chest diseases: TB, COPD, Pneumonia, Asthma, Lung cancer	n=257 (samples) f=38 (features) (i.e., n1=50 for TB, n2=71 for COPD, n3=60 for Pneumonia, n4=44 for Asthma, n5=32 for Lung Cancer)	Artificial Neural Networks: MNN, PNN, LVQNN, GRNN or Bayesian networks, RBFNN	TB: MLNN with LM provides 90%, COPD: PNN provides 88.73%, MLNN with LM provides 88.73%, ASTHMA: PNN provides 90.91%, MLNN with LM provides 90.91% Pneumonia: MLNN with LM provides 91.67% Lung Cancer: PNN provides 93.75%.
[104]	Chest diseases: TB, COPD, Pneumonia, Asthma, Lung cancer	n=357 (infected samples) f=38 (features) nh=38 (healthy samples) c=6 (classes) (i.e., n1=50 for TB, n2=71 for COPD, n3=60 for Pneumonia, n4=44 for Asthma, n5=32 for Lung Cancer, and n6=100 for normal)	Artificial Immune System (AIS)	TB: MLNN with LM and AIS provides 90%, COPD: AIS provides 92.96%, ASTHMA: PNN, MLNN with LM, MLNN with BP, AIS provide 90.91%, Pneumonia: AIS provides 93.33%, Lung Cancer: PNN, MLNN with LM, MLNN with BP, AIS provide 93.75%.
[105]	Breast Cancer, Liver patients, Vertebral column, Heart disease	Wisconsin Diagnostic Breast Cancer (WDBC) dataset: f=10 (features), i=569 (375 are benign and 212 are malignant) Indian Liver Patient Dataset (ILPD) dataset: f=11 (features), n1=416 (liver patients), n2=167 (non-liver patients), i.e., 441 male and 142 female. Vertebral Column Dataset (VCDS): f=6 (features), n1=100 (normal patients), n2=210 (abnormal patients) Heart Disease Data Set (HDDS): f=14 (features) n1=188 (normal patients), n2=106 (abnormal patients)	Artificial Neural Network classifiers: Multiple type classifier, Ensemble classifier (EC), Single type classifier, Individual classifier (IC), Solo classifier (SC)	Statistical test showed that: EC performs better than the IC with a mean difference of 0.0119 (p=0.005) and regardless of the type of disease. EC does not always perform better than the SC. Cost-effect analysis show that: EC and SC are in a tie based on the four datasets.

4. DISCUSSION AND CHALLENGES IN TB DISEASE DIAGNOSIS

This section discusses the major strengths and weaknesses of the above-described algorithms. From the aforementioned sections, we can conclude that the decision tree (DT) algorithms are easy to infer which reduce the uncertainty of complex decisions and assigns exact values to the results of several actions. However, DT algorithms can deal with numerical as well as categorical dataset but their performance varies according to nature of the dataset. Hence, it is an unstable classifier. On the other hand, support vector machines (SVMs) provide better performance in aspect to the accuracy though it is computationally cost-effective. Conversely, Naïve Bayes algorithm can scale with the dataset and are easy to implement. Due to its conditional dependency, it often becomes a naïve classifier, i.e., the outcome are inappropriate. Linear regression is visibly understandable and easy to explain. It can be standardized to neglect over-fitting. However, it carries out poor performance while dealing with non-linear associations. Logistic regression also throws away over-fitting and offers good probabilistic analysis. Yet, they are not good enough to deal with interactions that are more complicated. Neural networks (NNs) certainly identify the associations between variables (dependent and independent) and manipulate noisy dataset. Over-fitting, time-consumption, and local minima are the major drawbacks. K-Nearest neighbor (kNN) acts as a fast algorithm during training phase and easy to apply. It has also some issues including slow testing, need large memory and sensitive to noisy dataset. k-Means is an efficient clustering algorithm. The major drawbacks include knowing the amount of clusters in advance, dealing with categorical dataset, and effect of outlier on the performance. On the contrary, genetic algorithms (GAs) are robust, understandable, parallelized, stochastic, and supports multi-objective optimization. However, these algorithms are time-consuming in terms of computation. Deep learning architectures can be applied to a variety of problems since their out of sight layers decrease the necessity for feature mining. Nevertheless, these algorithms require massive data and are computationally exhaustive to train. Bootstrap aggregating (Bagging) is used to improve the accuracy of other ML algorithms through reduction of variances and over fitting of data. However, this algorithm cannot be

used on its own because it depends on other algorithms. Adaptive boosting (Adaboost) is used in conjunction with other algorithms to improve their performance. Because this algorithm evaluates all the data and classifies it in different values, it is relatively slow and reduces the accuracy of the primary algorithm. However, it improves the classification process by reducing the dimensionality of the output data.

A. Challenges in Tuberculosis disease diagnosis using ML

Tuberculosis (TB) is one of the leading fatal diseases around the globe. TB is defined as, "an infectious syndrome triggered by a microbes (bacilli) called Mycobacterium Tuberculosis" [106]. The tuberculosis bacteria affect the human respiratory tract, more specifically, the lungs. However, it can also affect other organs of the human body. TB is supposed to be one of the foremost syndromes of poverty since the risk factors of TB includes: HIV patients, overpopulation, starvation, vaccinating illegitimate medications, natives/personnel of areas where susceptible individuals gather (such as jails and impoverished asylums), medically deprivation, resource poor societies, high-risk tribal minorities, children in close contact with high-hazard class patients, and poor health-care infrastructure and coordination, smoking, intoxication, diabetes, individuals with weaken immune systems, etc. It has been observed that the symptoms of TB disease (seems to be very common) include long-lasting coughs, phlegm with blood, temperature, night-time sweats, chills, weakness, loss of appetite and loss of weight, therefore these might be negligible for a long time [106]. This can prompt deferrals in looking for care, and results in diffusion of the microscopic organisms to others. In addition, these key factors incorporating the deferral in diagnosis, improper or erroneous utilization of antimicrobial medications, or utilization of inadequate definitions of medications, and untimely treatment intrusion often leads the individual to the MDR-TB.

According to the World Health Organization (W.H.O.), Pakistan ranked 5th among the 30 TB high-burden countries globally. The TB disease yearly in Pakistan targets approximately 510,000 individuals. TB infected individuals can taint up to 10 to 15 other individuals through close contact throughout a year. Moreover, according to the WHO, Pakistan positions fourth most astounding predominance of MDR-TB around the world. Hence, devoid of legitimate treatment up to 66% of TB infected individuals will face the death [107]. In order to control over this dreadful disease and to minimize the death ratio of TB infected patients; it has been a necessity to propose an automated solution for early-stage diagnosis of TB.

For this purpose, several researchers have applied different machine learning algorithms as we have seen during literature review. However, the primary challenge is to diagnose the TB disease at an early stage by means of its signs, symptoms and risk factors. To achieve the aforementioned goal, Decision Tree (DT) algorithm would be applied. So far, the literature review has shown that DT algorithm has been applied in combination with other algorithms for the diagnosis of TB disease. Thus, the proposed research aims to apply the DT as a sole technique in order to measure its accuracy. The research study will be carried out in Sindh, province of Pakistan. Furthermore, our study will fill the following research gaps found in the literature:

a) To diagnose the TB disease at an early stage by means of its signs, symptoms and risk factors.
b) To analyze genuine datasets in order to guarantee the validity and verification of the proposed solution.
c) To take into account different factors including accuracy, precision, F-score in order to assure the reliability of the proposed solution.
d) To perform statistical testing in order to show the significance the proposed solution.

B. The proposed system for Tuberculosis disease diagnosis using ML

After reviewing the state-of-the-art, we proposed a framework for TB disease diagnosis using ML, as shown in the Fig. 4.

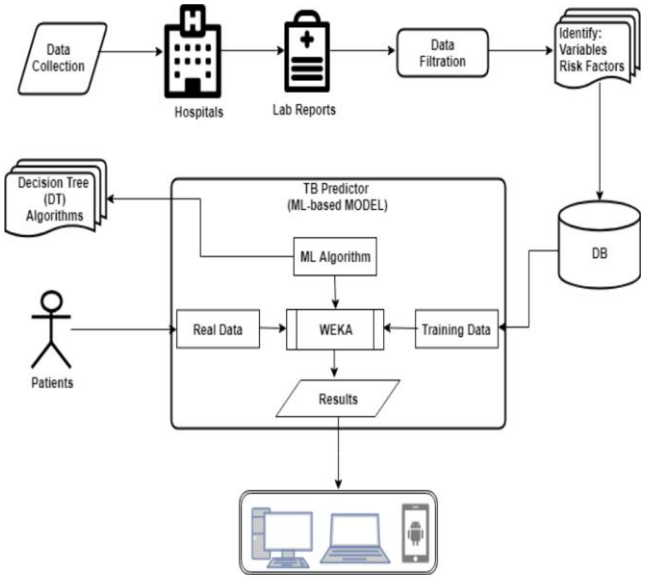

Figure 4. The framework of the proposed system.

1. At the first stage, data will be collected from hospitals in the form of patients' laboratory reports.
2. At the second stage, collected data would be filtered in order to identify the key variables and crucial risk factors causing TB disease.

3. At the third stage, filtered data would be saved into a database in order to create the training data.
4. At the fourth stage, an ML-based model is created on the training data to predict the risk level of the given new patient real data by applying Decision Tree algorithms. Moreover, we consider the Weka machine-learning tool for implementation of the model.
5. As a final point, the results would be displayed on screen.

5. CONCLUSION

In this paper, we have enlightened the impact of machine learning algorithms and practices in the context of healthcare informatics. We have classified machine learning healthcare informatics (ML-HI) procedures into four classes named as ML-HI types, ML-HI approaches, ML-HI paradigms and ML-HI algorithms. In this preliminary study, we have discussed the strengths and weaknesses of the studied techniques in order to help the healthcare research community to apply the appropriate technique in the healthcare domain. Furthermore, we have highlighted research directions and challenges specifically driven to Tuberculosis disease diagnosis. We introduced a proposed framework for TB disease diagnosis using ML.

ACKNOWLEDGMENT

A preliminary version of this paper [108] was presented at the International Conference on Intelligent Technologies and Applications, INTAP-2018, Springer CCIS (932), Islamia University Bahawalpur, Pakistan, October 23-25, 2018. Priyanka's and Vivekanand's work were supported for their MPhil/MS studies at Institute of Mathematics and Computer Science, University of Sindh, Jamshoro, Pakistan.

REFERENCES

[1] Jadad, A.R., OGrady, L.: How should health be defined? Br. Med. J. 337, a2900 (2008)

[2] Health Informatics Defined. (2016, December 29). Retrieved from http://www.himss.org/health-informatics-defined

[3] O'donoghue, J., & Herbert, J. (2012). Data management within mHealth environments: Patient sensors, mobile devices, and databases. Journal of Data and Information Quality (JDIQ), 4(1), 5.

[4] Mettler, T., & Raptis, D. A. (2012). What constitutes the field of health information systems? Fostering a systematic framework and research agenda. Health Informatics Journal, 18(2), 147-156.

[5] Parry, D. (2014). Health informatics. In Springer Handbook of Bio-/Neuroinformatics (pp. 555-564). Springer, Berlin, Heidelberg.

[6] O'Donoghue, J., O'Kane, T., Gallagher, J., Courtney, G., Aftab, A., Casey, A., ... & Angove, P. (2011). Modified early warning scorecard: the role of data/information quality within the decision making process. Electronic Journal of Information Systems Evaluation, 14(1), 100.

[7] Jordan, M. I., & Mitchell, T. M. (2015). Machine Learning : Trends, perspectives, and prospects. Science, 349(6245), 255-260.

[8] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436.

[9] Simon, P. (2013). Too big to ignore: the business case for big data (Vol. 72). John Wiley & Sons.

[10] Clifton, D. A., Niehaus, K. E., Charlton, P., & Colopy, G. W. (2015). Health informatics via Machine Learning for the clinical management of patients. Yearbook of medical informatics, 10(1), 38.

[11] Top 10 causes of death. (2018, May 22). Retrieved from http://www.who.int/gho/mortality_burden_disease/causes_death/top_10/en/

[12] Danish, M. I. (2012). Short textbook of medical diagnosis and management. Karachi, Pakistan: Paramount Books.

[13] Bambra, C., Gibson, M., Amanda, S., Wright, K., Whitehead, M., & Petticrew, M. (2009). Tackling the wider social determinants of health and health inequalities: evidence from systematic reviews. Journal of Epidemiology & Community Health, jech-2008.

[14] Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., & Sun, J. (2016, December). Doctor ai: Predicting clinical events via recurrent neural networks. In Machine Learning for Healthcare Conference (pp. 301-318).

[15] Holzinger, A. (2016). Machine Learning for health informatics. In Machine Learning for Health Informatics (pp. 1-24). Springer, Cham.

[16] Kononenko, I. (2001). Machine Learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in medicine, 23(1), 89-109.

[17] Hassanalieragh, M., Page, A., Soyata, T., Sharma, G., Aktas, M., Mateos, G., ... & Andreescu, S. (2015, June). Health monitoring and management using Internet-of-Things (IoT) sensing with cloud-based processing: Opportunities and challenges. In 2015 IEEE international conference on services computing (SCC) (pp. 285-292). IEEE.

[18] Kose, I., Gokturk, M., & Kilic, K. (2015). An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance. Applied Soft Computing, 36, 283-299.

[19] Waghade, S. S., & Karandikar, A. M. (2018). A Comprehensive Study of Healthcare Fraud Detection based on Machine Learning . International Journal of Applied Engineering Research, 13(6), 4175-4178.

[20] Holzinger, A. (2016). Interactive Machine Learning for health informatics: when do we need the human-in-the-loop?. Brain Informatics, 3(2), 119-131.

[21] Holzinger, A., Plass, M., Holzinger, K., Crişan, G. C., Pintea, C. M., & Palade, V. (2016, August). Towards interactive Machine Learning (iML): applying ant colony algorithms to solve the traveling salesman problem with the human-in-the-loop approach. In International Conference on Availability, Reliability, and Security (pp. 81-95). Springer, Cham.

[22] Amershi, S., Cakmak, M., Knox, W. B., & Kulesza, T. (2014). Power to the people: The role of humans in interactive Machine Learning . AI Magazine, 35(4), 105-120.

[23] Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., & De Freitas, N. (2016). Taking the human out of the loop: A review of bayesian optimization. Proceedings of the IEEE, 104(1), 148-175.

[24] Thornton, C., Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2013, August). Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 847-855). ACM.

[25] Mendoza, H., Klein, A., Feurer, M., Springenberg, J. T., & Hutter, F. (2016, December). Towards automatically-tuned neural networks. In Workshop on Automatic Machine Learning (pp. 58-65).

[26] Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., & Hutter, F. (2015). Efficient and robust automated Machine Learning . In Advances in Neural Information Processing Systems (pp. 2962-2970).

[27] Olson, R. S., Urbanowicz, R. J., Andrews, P. C., Lavender, N. A., & Moore, J. H. (2016, March). Automating biomedical data science through Tree-based pipeline optimization. In European Conference on the Applications of Evolutionary Computation(pp. 123-137). Springer, Cham.

[28] Olson, R. S., Bartley, N., Urbanowicz, R. J., & Moore, J. H. (2016, July). Evaluation of a tree-based pipeline optimization tool for automating data science. In Proceedings of the Genetic and Evolutionary Computation Conference 2016 (pp. 485-492). ACM.

[29] Swearingen, T., Drevo, W., Cyphers, B., Cuesta-Infante, A., Ross, A., & Veeramachaneni, K. (2017, December). ATM: A distributed, collaborative, scalable system for automated Machine Learning . In IEEE International Conference on Big Data.

[30] Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L. (2017). Disease prediction by Machine Learning over big data from healthcare communities. IEEE Access, 5, 8869-8879.

[31] Ensemble learning. (2018, July 05). Retrieved from https://en.wikipedia.org/wiki/Ensemble_learning

[32] Savio, A., García-Sebastián, M. T., Chyzyk, D., Hernández, C., Graña, M., Sistiaga, A., ... & Villanúa, J. (2011). Neurocognitive disorder detection based on feature vectors extracted from VBM analysis of structural MRI. Computers in biology and medicine, 41(8), 600-610.

[33] Ayerdi, B., Savio, A., & Graña, M. (2013, June). Meta-ensembles of classifiers for Alzheimer's disease detection using independent ROI features. In International Work-Conference on the Interplay Between Natural and Artificial Computation (pp. 122-130). Springer, Berlin, Heidelberg.

[34] Gu, Q., Ding, Y. S., & Zhang, T. L. (2015). An ensemble classifier based prediction of G-protein-coupled receptor classes in low homology. Neurocomputing, 154, 110-118.

[35] Maglogiannis, I. G. (Ed.). (2007). Emerging artificial intelligence applications in computer engineering: real word ai systems with applications in ehealth, hci, information retrieval and pervasive technologies (Vol. 160). Ios Press.

[36] Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised Machine Learning : A review of classification techniques. Emerging artificial intelligence applications in computer engineering, 160, 3-24.

[37] Tomar, D., & Agarwal, S. (2013). A survey on Data Mining approaches for Healthcare. International Journal of Bio-Science and Bio-Technology, 5(5), 241-266.

[38] Parmar, C., Grossmann, P., Bussink, J., Lambin, P., & Aerts, H. J. (2015). Machine Learning methods for quantitative radiomic biomarkers. Scientific reports, 5, 13087.

[39] Coates, A., Ng, A., & Lee, H. (2011, June). An analysis of single-layer networks in unsupervised feature learning. In Proceedings of the fourteenth international conference on artificial intelligence and statistics (pp. 215-223).

[40] Berkhin, P. (2006). A survey of clustering data mining techniques. In Grouping multidimensional data (pp. 25-71). Springer, Berlin, Heidelberg.

[41] Hastie, T., Tibshirani, R., & Friedman, J. (2009). Unsupervised learning. In The elements of statistical learning (pp. 485-585). Springer, New York, NY.

[42] Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Scientific reports, 6, 26094.

[43] Krishnapuram, B., Williams, D., Xue, Y., Carin, L., Figueiredo, M., & Hartemink, A. J. (2005). On semi-supervised classification. In Advances in neural information processing systems (pp. 721-728).

[44] Wang, Z., Shah, A. D., Tate, A. R., Denaxas, S., Shawe-Taylor, J., & Hemingway, H. (2012). Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised Machine Learning . PLoS One, 7(1), e30412.

[45] Tan, P. N. (2006). Introduction to data mining. Pearson Education India.

[46] Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

[47] Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. IEEE transactions on systems, man, and cybernetics, 21(3), 660-674.

[48] Tayefi, M., Tajfard, M., Saffar, S., Hanachi, P., Amirabadizadeh, A. R., Esmaeily, H., ... & Ghayour-Mobarhan, M. (2017). hs-CRP is strongly associated with coronary heart disease (CHD): A data mining approach using decision tree algorithm. Computer methods and programs in biomedicine, 141, 105-109.

[49] Abdar, M., Zomorodi-Moghadam, M., Das, R., & Ting, I. H. (2017). Performance analysis of classification algorithms on early detection of liver disease. Expert Systems with Applications, 67, 239-251.

[50] Shouman, M., Turner, T., & Stocker, R. (2011, December). Using decision tree for diagnosing heart disease patients. In Proceedings of the Ninth Australasian Data Mining Conference-Volume 121 (pp. 23-30). Australian Computer Society, Inc..

[51] Vapnik, V. N. (1999). An overview of statistical learning theory. IEEE transactions on neural networks, 10(5), 988-999.

[52] Vapnik, V. (1998). The support vector method of function estimation. In Nonlinear Modeling (pp. 55-85). Springer, Boston, MA.

[53] Mountrakis, G., Im, J., & Ogole, C. (2011). Support vector machines in remote sensing: A review. ISPRS Journal of Photogrammetry and Remote Sensing, 66(3), 247-259.

[54] Shmilovici, A. (2009). Support vector machines. In Data mining and knowledge discovery handbook (pp. 231-247). Springer, Boston, MA.

[55] Adankon, M. M., & Cheriet, M. (2009). Support vector machine. In Encyclopedia of biometrics (pp. 1303-1308). Springer, Boston, MA.

[56] Polat, K., Güneş, S., & Arslan, A. (2008). A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine. Expert systems with applications, 34(1), 482-487.

[57] Magnin, B., Mesrob, L., Kinkingnéhun, S., Pélégrini-Issac, M., Colliot, O., Sarazin, M., ... & Benali, H. (2009). Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI. Neuroradiology, 51(2), 73-83.

[58] Huang, C. L., Liao, H. C., & Chen, M. C. (2008). Prediction model building and feature selection with support vector machines in breast cancer diagnosis. Expert Systems with Applications, 34(1), 578-587.

[59] Zhang, H. (2004). The optimality of naive Bayes. AA, 1(2), 3.

[60] Kazmierska, J., & Malicki, J. (2008). Application of the Naïve Bayesian Classifier to optimize treatment decisions. Radiotherapy and Oncology, 86(2), 211-216.

[61] Pattekari, S. A., & Parveen, A. (2012). Prediction system for heart disease using Naïve Bayes. International Journal of Advanced Computer and Mathematical Sciences, 3(3), 290-294.

[62] Bhuvaneswari, R., & Kalaiselvi, K. (2012). Naive Bayesian classification approach in healthcare applications. International Journal of Computer Science and Telecommunications, 3(1), 106-112.

[63] Linear regression. (2018, July 03). Retrieved from https://en.wikipedia.org/wiki/Linear_regression

[64] Logistic regression. (2018, July 03). Retrieved from https://en.wikipedia.org/wiki/Logistic_regression

[65] Kurt, I., Ture, M., & Kurum, A. T. (2008). Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. Expert systems with applications, 34(1), 366-374.

[66] Thirumalai, C., & Manzoor, R. (2017, April). Cost optimization using normal linear regression method for breast cancer Type I skin. In Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of (Vol. 2, pp. 264-268). IEEE.

[67] Saleheen, D., Scott, R., Javad, S., Zhao, W., Rodrigues, A., Picataggi, A., ... & Kastelein, J. J. (2015). Association of HDL cholesterol efflux capacity with incident coronary heart disease events: a prospective case-control study. The lancet Diabetes & endocrinology, 3(7), 507-513.

[68] Peterson, L. E. (2009). K-nearest neighbor. Scholarpedia, 4(2), 1883.

[69] Adeniyi, D. A., Wei, Z., & Yongquan, Y. (2016). Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method. Applied Computing and Informatics, 12(1), 90-108.

[70] Chen, H. L., Huang, C. C., Yu, X. G., Xu, X., Sun, X., Wang, G., & Wang, S. J. (2013). An efficient diagnosis system for detection of Parkinson's disease using fuzzy k-nearest neighbor approach. Expert systems with applications, 40(1), 263-271.

[71] Deekshatulu, B. L., & Chandra, P. (2013). Classification of heart disease using k-nearest neighbor and genetic algorithm. Procedia Technology, 10, 85-94.

[72] Rai, P., & Singh, S. (2010). A survey of clustering techniques. International Journal of Computer Applications, 7(12), 1-5.

[73] Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern recognition letters, 31(8), 651-666.

[74] Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. Annals of Data Science, 2(2), 165-193.

[75] Singh, D., & Reddy, C. K. (2015). A survey on platforms for big data analytics. Journal of Big Data, 2(1), 8.

[76] Zheng, B., Yoon, S. W., & Lam, S. S. (2014). Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. Expert Systems with Applications, 41(4), 1476-1482.

[77] Escudero, J., Zajicek, J. P., & Ifeachor, E. (2011, August). Early detection and characterization of Alzheimer's disease in clinical scenarios using Bioprofile concepts and K-means. In Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE (pp. 6470-6473). IEEE.

[78] Balasubramanian, T., & Umarani, R. (2012, March). An analysis on the impact of fluoride in human health (dental) using clustering data mining technique. In Pattern Recognition, Informatics and Medical Engineering (PRIME), 2012 International Conference on (pp. 370-375). IEEE.

[79] Mukhopadhyay, A., Maulik, U., Bandyopadhyay, S., & Coello, C. A. C. (2014). A Survey of Multiobjective Evolutionary Algorithms for Data Mining: Part I. IEEE Trans. Evolutionary Computation, 18(1), 4-19.

[80] Oreski, S., & Oreski, G. (2014). Genetic algorithm-based heuristic for feature selection in credit risk assessment. Expert systems with applications, 41(4), 2052-2064.

[81] Guo, H. Y., Zhang, L., Zhang, L. L., & Zhou, J. X. (2004). Optimal placement of sensors for structural health monitoring using improved genetic algorithms. Smart materials and structures, 13(3), 528.

[82] Shah, S., & Kusiak, A. (2007). Cancer gene search with data-mining and genetic algorithms. Computers in biology and medicine, 37(2), 251-261.

[83] Yan, H., Zheng, J., Jiang, Y., Peng, C., & Xiao, S. (2008). Selecting critical clinical features for heart diseases diagnosis with a real-coded genetic algorithm. Applied soft computing, 8(2), 1105-1111.

[84] Basheer, I. A., & Hajmeer, M. (2000). Artificial neural networks: fundamentals, computing, design, and application. Journal of microbiological methods, 43(1), 3-31.

[85] Were, K., Bui, D. T., Dick, Ø. B., & Singh, B. R. (2015). A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. Ecological Indicators, 52, 394-403.

[86] Amato, F., López, A., Peña-Méndez, E. M., Vaňhara, P., Hampl, A., & Havel, J. (2013). Artificial neural networks in medical diagnosis.

[87] Abbass, H. A. (2002). An evolutionary artificial neural networks approach for breast cancer diagnosis. Artificial intelligence in Medicine, 25(3), 265-281.

[88] Raith, S., Vogel, E. P., Anees, N., Keul, C., Güth, J. F., Edelhoff, D., & Fischer, H. (2017). Artificial Neural Networks as a powerful numerical tool to classify specific features of a tooth based on 3D scan data. Computers in biology and medicine, 80, 65-76.

[89] Bhardwaj, A., & Tiwari, A. (2015). Breast cancer diagnosis using genetically optimized neural network model. Expert Systems with Applications, 42(10), 4611-4620.

[90] Deng, L. (2012). Three classes of deep learning architectures and their applications: a tutorial survey. APSIPA transactions on signal and information processing.

[91] Chen, X. W., & Lin, X. (2014). Big data deep learning: challenges and perspectives. IEEE access, 2, 514-525.

[92] Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural networks, 61, 85-117.

[93] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ... & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. Medical image analysis, 42, 60-88.

[94] Ravı, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., & Yang, G. Z. (2017). Deep learning for health informatics. IEEE journal of biomedical and health informatics, 21(1), 4-21.

[95] Liu, S., Liu, S., Cai, W., Pujol, S., Kikinis, R., & Feng, D. (2014, April). Early diagnosis of Alzheimer's disease with deep learning. In Biomedical Imaging (ISBI), 2014 IEEE 11th International Symposium on (pp. 1015-1018). IEEE.

[96] Acharya, U. R., Fujita, H., Oh, S. L., Hagiwara, Y., Tan, J. H., & Adam, M. (2017). Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals. Information Sciences, 415, 190-198.

[97] Dietterich, T. G. (2000, June). Ensemble methods in Machine Learning . In International workshop on multiple classifier systems (pp. 1-15). Springer, Berlin, Heidelberg.

[98] Shigei, N., Miyajima, H., Maeda, M., & Ma, L. (2009). Bagging and AdaBoost algorithms for vector quantization. Neurocomputing, 73(1-3), 106-114.

[99] Tu, M. C., Shin, D., & Shin, D. (2009, October). Effective diagnosis of heart disease through bagging approach. In Biomedical Engineering and Informatics, 2009. BMEI'09. 2nd International Conference on (pp. 1-4). IEEE.

[100] Morra, J. H., Tu, Z., Apostolova, L. G., Green, A. E., Toga, A. W., & Thompson, P. M. (2010). Comparison of AdaBoost and support vector machines for detecting Alzheimer's disease through automated hippocampal segmentation. IEEE transactions on medical imaging, 29(1), 30-43.

[101] Ilhan, H. O., & Celik, E. (2016, October). The mesothelioma disease diagnosis with artificial intelligence methods. In Application of Information and Communication Technologies (AICT), 2016 IEEE 10th International Conference on (pp. 1-5). IEEE.

[102] Yahiaoui, A., Er, O., & Yumusak, N. (2017). A new method of automatic recognition for tuberculosis disease diagnosis using support vector machines. Biomedical Research, 28(9).

[103] Er, O., Yumusak, N., & Temurtas, F. (2010). Chest diseases diagnosis using artificial neural networks. Expert Systems with Applications, 37(12), 7648-7655.

[104] Er, O., Yumusak, N., & Temurtas, F. (2012). Diagnosis of chest diseases using artificial immune system. Expert Systems with Applications, 39(2), 1862-1868.

[105] Weng, C. H., Huang, T. C. K., & Han, R. P. (2016). Disease prediction with different types of neural network classifiers. Telematics and Informatics, 33(2), 277-292.

[106] Kumar, P. J., Clark, M. L., & Feather, A. (2017). Kumar & Clarks clinical medicine. Edinburgh:Elsevier.

[107] Use of high burden country lists for TB by WHO in the post-2015 era: Summary. Retrieved from http://www.who.int/tb/publications/global_report/high_tb_burden countrylists2016-2020summary.pdf

[108] P. Karmani, A.A. Chandio, I.A. Korejo, M.S. Chandio (2019) "A Review of Machine Learning for Healthcare Informatics Specifically Tuberculosis Disease Diagnostics". In: Bajwa I., Kamareddine F., Costa A. (eds) Intelligent Technologies and Applications. INTAP 2018. Bahawalpur, Pakistan, October 2018.Communications in Computer and Information Science, CCIS, Vol 932. pp. 50-61, Springer, Singapore, March 2019, doi=10.1007/978-981-13-6052-7_5

[109] V. Karmani, A. A. Chandio, P. Karmani, M. Chandio, and I. A. Korejo. "Towards Self-Aware Heatstroke Early-Warning System Based on Healthcare IoT." In 2019 Third World Conference on Smart Trends in Systems Security and Sustainablity (WorldS4), pp. 59-63. IEEE, 2019.

[110] Korejo, IA, K. Brohi, AA Chandio, FA Memon, and AR Nangraj. "An Adaptive Mutation Strategy for Real Coded Genetic Algorithm." Sindh University Research Journal-SURJ (Science Series) 49, no. 2 (2017).

Priyanka Karmani is a MPhil/MS scholar in computer science of Institute of Mathematics and Computer Science at the University of Sindh, Jamshoro, Sindh, Pakistan. She is Lecturer in Cadet College Petaro Jamshoro. Her research interests include machine learning, data mining, internet of things, healthcare, cloud computing, big data.

Aftab Ahmed Chandio received a PhD. Degree from Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China in 2016. He is currently associate professor of Institute of Mathematics and Computer Science at the University of Sindh, Jamshoro, Sindh, Pakistan. His research interests include cloud computing, big data, parallel and distributed systems, scheduling, energy optimization, workload characterization, and map-matching strategies for GPS trajectories. He has published more than 20 papers in journals and conferences.

Vivekanand Karmani is a MPhil/MS scholar in computer science of Institute of Mathematics and Computer Science at the University of Sindh, Jamshoro, Sindh, Pakistan. His research interests include internet of things, healthcare, cloud computing, big data.

Javed Ali Soomro received a PhD. Degree from Beijing Sports University, Beijing, China in 2018. He is currently assistant professor of Centre for Physical Education, Health and Sports Science at the University of Sindh, Jamshoro, Sindh, Pakistan. His research interests include healthcare, sports science, physical education, internet of things.

Imtiaz Ali Korejo received his Ph.D. from the Department of Computer Science, University of Leicester, United Kingdom in 2012. Dr. Imtiaz Ali Korejo received his B.Sc. (Hons) and M.Sc.(Hons) in Computer Science from University of Sindh, Jamshoro, Pakistan, in 1999, and 2000, respectively. He is currently professor of Institute of Mathematics and Computer Science at the University of Sindh, Jamshoro, Sindh, Pakistan. His research interests are evolutionary algorithms, genetic algorithms and adaptive approaches.

Muhammad Saleem Chandio received a PhD from Department of Computer Science, University of Wales, Swansea, Wales, UK, in 2002 and MSc in Mathematics from Institute of Mathematics and Computer Science, University of Sindh, Jamshoro Pakistan in 1986. He is currently professor of Institute of Mathematics and Computer Science at the University of Sindh, Jamshoro, Sindh, Pakistan.