Volume 64, Issue 2, 2020

Journal of Scientific Research

Institute of Science,

Banaras Hindu University, Varanasi, India.

State-of-the-Art in Performance Metrics and Future Directions for Data Science Algorithms

Ajay Sharma* and Promad Kumar Mishra

Department of Computer Science, Banaras Hindu University, Varanasi, India.

ajay.sharma17@bhu.ac.in*, mishra@bhu.ac.in

Abstract: **With the advancement in cutting-edge technology, a huge volume of data is generated every day. Data science is one of the unique and most motivating area of research which is progressively popular now a days. Data science plays a vital role for analysing a massive volume of data that cannot be processed by conventional technologies in real time. Its aim is to find solutions from existing data in order to improve our existing systems to reduce time and make it cost effective. In this article we explored why we need data science, big data and data mining techniques. It gives readers clear intuition towards the basic steps required for dealing with data science and analytics problem. This work focuses more on the supervised learning. In this review we review all related articles in the field of Healthcare that needs an improvement for making our healthcare systems more reliable. This also highlights an introduction of important techniques of supervised learning in the domain of healthcare for better understanding of techniques. In particular Recommendation regarding the choice of suitable activation function and evaluation metric to improve the performance of classifiers is explained briefly.**

Index Terms: **Data science, Healthcare, Feature engineering, Classification, Performance metrics.**

I. INTRODUCTION

With the advancement in the cutting-edge technology a very huge volume of data is generated every day. According to Forbes article, by 2020 in every second there will be 1.7 megabytes of information generated for every person on this planet. In this competitive world, various companies' goal is to exploit and analyse data for finding useful advantage out of it. Availability of Data itself is solution to various existing problems. Conventional Technology databases and manual work is proven useless today due to speed and volume of data generated. Today in 21st century we are working with powerful computers with high speed network and processing availability. Data can be structured, unstructured or in semi structured from. The availability of data in various fields can be utilized properly for improving decision making capabilities of the available systems in various fields. In healthcare field data is critically important and is available mostly in unstructured from. We can analyse this data to diagnose and predict a disease in early stages on the basis of available symptoms. Also, we can predict better medicine. In business analysis we can answer questions like: Where a particular thing is sold more? While buying some particular thing what is the probability of buying some other thing? We can use available data to make better healthcare, education, transportation etc.

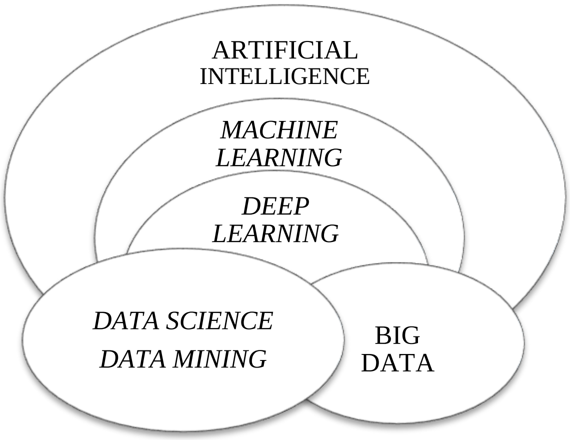

Fig.1. Relationship between AI, ML, DL, DS, DM

[*] Corresponding Author

DOI: 10.37398/JSR.2020.640232

To find solutions of above questions gives birth to Data Science which is one of the hottest research field of 21st century (*Davenport & Patil, 2013*). In "Fig. 1" we have shown the relationship between various Emerging fields with data science.

A. Data Science

Data science is an assembly of machine learning algorithms and lots of statistics to find out unknown knowledge from raw data. Data science is formulated by integration of data and science. Data can be any raw information in structured, unstructured or semi structured form that can be integrated from different sources. Whereas Science is the way of exploring and observing data. In particular, the way of extracting hidden knowledge and intuitions from data by means of scientific approaches. It is useful for processing of massive and unstructured data that is difficult to be processed by conventional techniques. Data mining is also one of closely related concept of data science. Data mining also include hundreds of algorithms used to process complex data. Data science is a widely used field now a days in business, healthcare, finance, education, transport etc. It is mainly used for predictions and decision making from raw data. Data science makes use of machine learning, deep learning, mathematics, and statistics for prescriptive as well as predictive analytics (*Raban & Gordon ,2020*). In "Fig. 2" we have shown the life cycle of data science. In Data science we analyse huge volume of data in order to improve decision making ability. Thus, analytical thinking is one of the key aspect of data science. In customer relationship management data science helps to analyse customer data in order to know behaviour and the value of the customer. Fundamental of the data science is statistics.

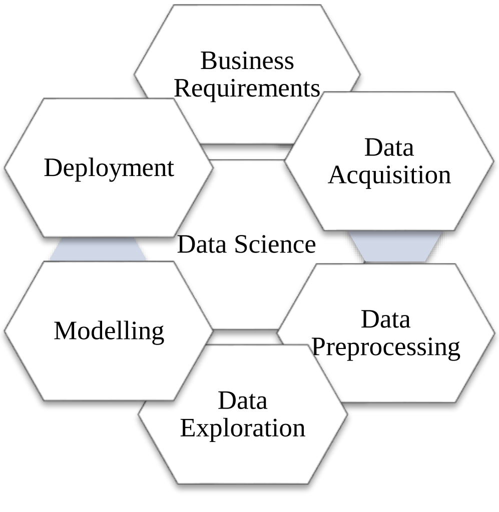

Fig. 2. Data Science Life Cycle

Suppose there are hard cyclones from last few days. So, what is the role of data science here in this issue? Data analysts can exactly predict the probability of areas to be affected more. It can also answer what will be the effect of cyclone on business and what are the things that people prefer to buy during cyclones for their safety (*Dhar, 2013*). Data science is one of the hottest jobs of industry and companies are hiring more and more data scientists. Academics institutions are also promoting it and in research it has great importance (*Leek, 2013*). Finally, we can say that data science is related to "Big Data and Data Mining" that is difficult to process by conventional data processing algorithms.

B. Big Data

In the modern era of technology data is generated by different sources quickly which is very difficult to process by traditional techniques of processing in real time. Due to continuously change in the speed and volume of data generated forced to upgrade from traditional processing systems to big data science. Big data itself expresses the data that increases exponentially with time and very hard to store, collect, maintain, analyse and visualize in real time by traditional technologies. In this real-world we know how our data is useful in decision making so it is important to store and process this huge volume of data quickly. Big data technology helps to do this work very quickly and efficiently. It helps to convert it to a manageable piece of information and answer various useful questions. Since, it is important to know where does this big data comes from. It can be machine generated or human generated. Machine generated data is recorded with the help of surveillance cameras, sensors or application server logs. It can be whether or atmospheric data, images/videos, data generated from internet/games/websites *(West,2012)*. Human generated data include emails/Reviews/blogs/Medical records/Facebook/linked-in/Twitter/contacts etc. Many companies like google, twitter, Facebook, WhatsApp etc. use this big data architecture to work

(*Bhat & Ahmed, 2016*). In the field of healthcare, public administration, business, automation etc. big data plays a vital role in solving issues of daily life. For e.g. today various E-commerce and M-commerce companies like amazon, Flip Kart gather billions of transactions per day. Everyday medical and healthcare sector stores large volume of medical records. This huge volume of data needs to be processed and analysed in Real-time. Hadoop, CouchDB and HBase resolves issue of "big data" processing and storage. The processing and storage issue of this big data is resolved using HADOOP architecture developed by yahoo and MapReduce approach developed by Google. MapReduce technique divides a large volume of input data into fixed chunks, where each chunk is processed in parallel to reduce the processing and storage time. NOSQL databases are used for the storage of big data like MangoDB, CouchDB, Cassandra etc. The need to process big data to find out hidden and useful patterns gives birth to data mining. Five characteristics that define big data are volume of data generated, velocity, veracity, variety and value (*Bollier & Firestone, 2010*).

C. Data Mining

Today, we are living in the world of big data. More than 2.5 quintillion of data is generated every day. It is a big challenge how to analyse this big data and to collect meaningful knowledge from this big data. Such a useful task is possible with the help of data mining tools. Data mining is a way of finding out hidden patterns from a very large volume of data. It tries to find out unknown patterns that are useful to answer various questions and drawing conclusions (*Hand, Mannila & Smyth, 2001*). Everyday huge amount of data is generated from business, industry, medical and healthcare sector, automation, sensors, internet etc. The data mining involves various phases to discover knowledge from data (*Koh & Tan, 2011*). In “Fig. 3” we have shown various steps involved in KDD in the form of waterfall model to discover knowledge from original data or raw data.

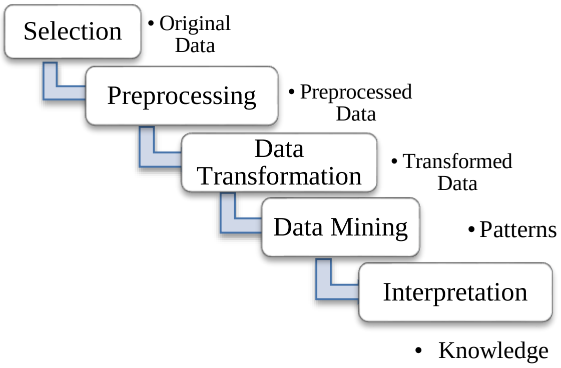

Fig. *3*. Knowledge Discovery from Data

Data collection is the first phase where data is collected using IOT sensors or by manual surveys. This stage is critically important so best possible choices should be used to collect data because its impact’s data mining process. Data ware house or database is a store where data is stored for processing after collection. This collected data in original form always contaminated with impurities and inconsistencies which is not useful for processing (*Xu et al., 2019*). Major threats of Data are noise (deviate from actual result), incompleteness (lacking attribute) and inconsistency (dependencies in code). So, data pre-processing stage is used to convert this inconsistent data useful for data mining algorithms to analyse. Various steps involved in data pre-processing in shown in “Fig. 4”.

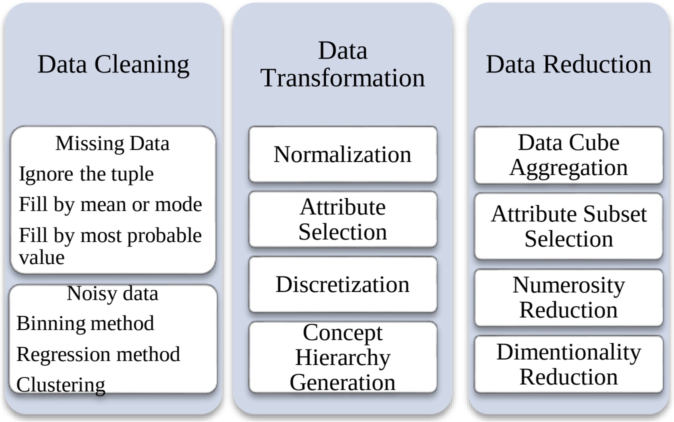

Fig. 4. Data pre-processing steps

It involves data cleaning, feature selection, feature extraction, data transformation and data reduction. Our data contains various features but every feature is not relevant one for our algorithm. According to particular application, analysts select some most relevant features that are helpful in mining and leave other poorer features. After feature extraction stage data cleaning is done. Data cleaning helps to remove inconsistent values that arise due to some error or missing value. Whereas missing values can be filled manually with most probable values, with the mean value or to leave the record in case of more missing values. Sometimes data from various sources are extracted, cleaned and integrated into a common source for processing. After removing inconsistencies from the data, it is necessary to transform data features into another feature space which is useful for data analysis. In particular, most of mining algorithms works well on low dimensional features because high dimensional features are more error prone and noisy. Data transformation means transforming the old set of values of a given attribute or set of attributes to new values of different or similar formats so that each old value can be identified with one of the new values. It is usually done to represent our data in easy formats and within a particular range for analysis purpose (*Usama et al.,1996*). Some of useful methods of data transformations are normalization, attribute selection, discretisation, and concept hierarchy generation (*Kumar & Katyal, 2018*). Also, it is difficult to analyse a huge volume of data, so after data transformation data reduction technique is applied. Aim behind data reduction is to reduce the requirement of storage space and cost of analysis. Principle component analysis (PCA) and wavelet transformation are two important methods of dimensionality reduction.

II. Literature Review

With the advancement in technology, Data science brings insights into the research field. Due to its popularity and power in analysing big data, it is one of the fundamental approaches to solve the data driven problems. Our motivation to work in this area comes from its use in big companies, massive application areas, opportunities and challenges to solve real life problems *(Kaur & Grewal, 2016).* In this field a lot of work has been carried out in almost every area like Transportation, healthcare and medicine, Finance, Banking, Education, and Agriculture etc. But healthcare is one of the prominent and emerging field because healthcare is the issue that every person faces some day. Now a days this world is in grief from virus named Coronavirus. Various data scientists use machine learning, deep learning and mining to find out quick decisions for the treatment of Coronavirus. Thus, data science is a backbone of healthcare today. This is one of the fields that also motivates me to explore this world of data science. INTERNET SEARCH is one of an exclusive application of data science. Various search engines like google, Bing, yahoo, AOL, ASK makes use of data science algorithms to give response to any search query in a fraction of seconds. At least 20 petabytes of data are processed by google every day. RECOMMENDATION SYSTEM is one of the biggest application of data science. Netflix is one of the most popular online video streaming platforms where people can see online movies, shows in real time on their personal computers. Netflix is using data science for improving their recommendation systems to provide interesting and relevant contents according to users search history. By using data science, they increased the performance by 10.06 % by ensemble of 107 algorithm. AMAZON and Flipkart both are one of the largest E-commerce Marketplaces in India. It uses data science for making recommendation systems that recommend its customer an item of recent search trends. It helps to increase performance and decrease search time. There are various applications that makes use of data science. In "Table I". We have shown various use cases where data science is tremendously used.

In the field of data science, we need to familiar with data mining, machine learning and deep learning approaches.

Table I. Use Cases of Data Science

DOMAIN	Applications / Use Cases
Internet Search	Google, Bing, Yahoo, AOL, Ask, DuckDuckGo, Yandex, Baidu, WolframAlpha, Internet Archive
Recommender Systems	Netflix, Google Play, Amazon, Twitter , imdb, LinkedIn, Flipkart, ParallelDots, Sajari
Speech Recognisation	Google Now, Microsoft Cortana, Apples Siri, Amazon Alexa, Hound
Price Comparison Websites	Pricerunner, Shopzilla, Junglee, Dealtim, Bizrate, Nextag, Pricegrabber, Pronto
Delivery Logistics	DHL, Kuehne + Nagel, usps, FedEx, UPS

In our related work we review a number of high-quality recent articles from IEEE, springer etc. in the field of healthcare. It includes heart disease, cervical cancer, breast cancer, Parkinson disease, Alzheimer, lung cancer, skin cancer, diabetes, kidney and Thyroid. *Sanjay Kumar Singh* et al. proposed a model for the classification of a *cervical* cancer using various machine learning classification algorithms. This proposed model uses Decision tree, Logistic regression, random forest, k-Nearest neighbor, perceptron etc. It calculates accuracy, precision, recall and F1-score by using confusion metric. Accuracy obtained by these algorithms are Decision tree (99.01), Logistic regression (97.84), random forest (99.83), k-Nearest neighbor (100), perceptron (98.66) etc.

Senthil Kumar Mohan et.al (2019) makes a hybrid model for the prediction of *heart* disease. Dataset for the classification of heart disease is Cleveland UCI repository dataset having 303 patient records with each record having 75 attributes. In pre-processing 6 records were discarded and 297 were considered only for research work. Out of 75 features 14 important features are considered for research. Various classifiers like decision tree, naïve Bayes, support vector machine, logistic regression and generalized linear model were initially used. Then a hybrid model is constructed by using three best classifiers (Decision tree, Random forest and generalized linear model).

varamakrishnan rajaraman et al. (2020) developed an ensemble approach using chest radiographs for the detection of *tuberculosis*. In his study (i) Pediatric pneumonia dataset (positive D 1583, negative D 4273), ii) RSNA dataset (positive D 8851, negative D 17833) and (iii) Indiana dataset (positive D 1726, negative D 2378) to get accuracy of 95% based on deep learning with modifications of various activation function.

Sanjay Kumar et al. (2018) proposed a model for the prediction of *liver* disorder by using Naïve bayes, KNN, Random forest, decision tree and adaptive boosting. Dataset is taken from UCI repository containing 345 samples with 7 features each.

QIANG XU et al. (2019) proposed a model for the detection of chronic obstructive *pulmonary* disease (COPD) by using SVM, KNN, NB, ANN with respective accuracy score of 58.6, 77.4, 67.8, 86.4 %. Dataset used in this study contains 18471 clinical records. Further we can improve this model for better prediction.

Samrat Kumar Dey et al. (2018) proposed a model to predict type 2 diabetes using ANN, KNN, NB, LR. Dataset is taken from UCI containing 2 classes. In table 2 we listed a number of reviewed papers with limitations and essential details for the application of data science in healthcare in order to improve our healthcare systems.

Manogaran et al. (2018) presented an expert system for heart diseases diagnosis using Multiple Kernel Learning (MKL) with adaptive neuro-fuzzy inference system. The proposed system used the data of 123 samples of heart disease patients each record with 7 input parameters such as age, blood pressure, blood sugar, heart rate, body temperature, cholesterol and chest pain type and 1 output parameter such as heart attack.

A diagnostic system for depression was developed by Ekong and Onibere (2015) using Takagi–Sugeno neuro-fuzzy system with case-based reasoning. The study was conducted on the medical data collected from hospital in Nigeria comprising 100 instances of depression each with 9 psychological and 7 physiological factors. The proposed model generated 144 fuzzy rules using fuzzy inference system. The generated rules were then optimized by evaluating local similarities using neural network and finally case-based reasoning was implemented for making decision support system.

Table II. Data Science Techniques in Healthcare

Author(year)	Techniques/Methodology	Clinical Context	Dataset Details	Dataset classes	Limitations
Senthil Kumar Mohan et al. (2019)	Naïve Bayes, Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, Hybrid (HRFLM)	*heart disease.*	Samples-303, Features -76 Features extracted-13 Samples discarded-6 Clinical Data	Nominal: (0-4) Absence (0)-160 Presence (1)-137	Dataset is small. Training time is more.
David A. Omondiagbe et al. (2019)	Support vector Machine, Naïve Bayes, Artificial neural network	*Breast Cancer.*	Classes -2, Samples-569 Features -32 Clinical Data	Benign—357 Malignant—212	Lack of Bagging, boosting.
Ahmed Osmanović et al. (2019)	SVM, SVM-RFE, SVM-LDA, Neural network, NN-PCA, Naïve Bayes, NB-RFE, NB-LDA	*Breast Cancer.*	Classes -2 Samples-699, Features-9 Clinical data	Benign—458 Malignant—241	Class imbalance. Lack of ensemble Learning.
Shamsul Huda et al. (2016)	SVM, Decision tree, ANNIGMA+ SVM+ Decision tree+ Bagging	*Brain Tumour*	Classes = 2 Samples = 100 MRI images	Brain tumour-6 Control -400	No cross validation.
Sanjay Kumar Singh et al. (2020)	Decision Tree, Random Forest, Perceptron, K-Nearest Neighbor, SVM	*Cervical Cancer*	Classes -2 Samples-803 Pap smear images	Benign Malignant	Lack of ensemble learning.
Sherif fayz1 et al. (2018)	PCA, RFE, Random Forest, SMOTE RF, SMOTE-RF-PCA	*Cervical Cancer*	Samples-858 Attributes-36 Features-32 Clinical Data	Target Variables-4 Hinselmann, Cytology, Schiller, and Biopsy	Class imbalance. Lack of ensemble method.
Saurabh Pal et al. (2019)	Decision Tree, Support Vector Machine, Random Forest, CART, Gradient Boosting	*Dermatology/Skin Cancer*	Classes -6 Samples -366, Features - 35 Clinical Data	psoriasis 112, seboreic dermatitis-61, lichen planus-72, pityriasis rosea -49, cronic dermatitis- 52, pityriasis rubra pilaris-20	Less Accuracy.
Muhammad Rehan Abbas et al. (2019)	Support Vector Machine, SVMR, SVML, Random Forest, K-Nearest Neighbor.	*Lung +Breast Cancer*	Datasets-6 WPBC-194, WDBC-569, WBCD-683.	Classes-2 Benign Malignant	Lack of bagging, boosting.

Table II. Continued…

Resul Das et al.
Amin Ul Haq et al. (2019)
	Logistic Regression, Decision Tree, Artificial neural network, SVM-linear, SVM-RBF	*Parkinson disease*	Paticipants-31
			Males-19, Females-12
			Samples-197, Features -23
				Classes-2
				Disease-23(M-16, F-7)
				Control-8(M-3, F-5)
					Dataset is small.
					Lack of ensemble model.

Norma Latif Fitriyani `et al. (2019)	Decision Tree, Support Vector machine, LR, MLP, K-Means+LR, DBSCAN+SMOTE+RF, CART	*Diabetes and Hypertension*	Classes -2, Samples-569, Features - 32 Clinical Data	Datasets-4 (Type 2 Diabetes, Hypertension, kidney disease, Prehypertension)	class imbalanced. Lack of ensemble method.
Samrat Kumar Dey et al. (2018) H. Roopa (2019)	SVM, KNN, GNB, ANN PCA, Linear Regression Model	*diabetes Type 2*	Classes-2 Samples-768, Features-9 Clinical Data	Diabetes (1)-268 Normal (0)-500	Class imbalance. Lack of ensemble approach.
Sampath and Saradha et al. (2015)	SVM, HNFRK Ten-fold cross validation	*Alzheimer*	Classes = 2 Samples = 150 MRI image	AD—95 Control—55	Class imbalance.
JIONGMING QIN et al. (2019)	logistic regression, random forest, support vector machine, k-nearest neighbor, naive Bayes classifier and feed forward neural network, Logistic + Random Forest	*Kidney disease*	Classes-2, Samples-400, Features-25(numerical- 11, categorical-13, target1) Clinical data	Disease-250 Normal-150	Dataset is small. Class imbalance.
Prabal Poudel et al.	Support vector machine, ANN, Random Forest	*Thyroid*	Datasets-2, Sample(D1)-675 Samples(D2)-3370, Features Extracted-30 Ultrasound images		No cross Validation.
HAKAN GUNDUZ (2019)	Deep Learning (9 layered CNN), LOPO CV	*Parkinson disease*	Paticipants-252 Males-130, Females- 120 Samples-756, Features -754 Clinical data	Classes-2 PD-188 (M-107, F-81) Control-64(M-23, F-41)	Class imbalance.
Sanjay Kumar et al. (2018)	Naïve bayes, KNN, Random Forest, K-means, c5.0, c5.0 with Adaptive boosting	*Liver Disorder*	Classes-2, variables-11 Instances-583 Clinical data	Classes-2, PD-188 (M-107, F-81) Control-64 (M-23, F-41)	Class imbalance. No cross validation.

III. DATA MINING TECHNIQUES

Data mining is a way of extracting knowledge from a large volume of data. Extraction of this knowledge is based on the supervised, unsupervised semi supervised or reinforcement learning Techniques. In this article we mainly focus on the supervised learning techniques. In supervised learning we have labeled data and learning process is based on this labeled data. Thus, model is trained by mapping all input points with their respective targets. When a new input tuple is given it analyses training data and categories that input with correct target. But in unsupervised learning data is unlabeled. Thus, machines are not trained but model learns by their own and try to find hidden patterns by our-self. In "Fig. 5" we have shown some useful supervised and unsupervised learning techniques.

A. Classification

In Statistics and Machine learning classification is a supervised learning technique. This Classification model classifies the given set of input points into a target class/label rather than predicting a quantity. This technique is good for both structured and unstructured data. One of the most useful supervised learning methods of Data Science in Healthcare sector is classification. Its main purpose is to identify the new input will fall in which class.

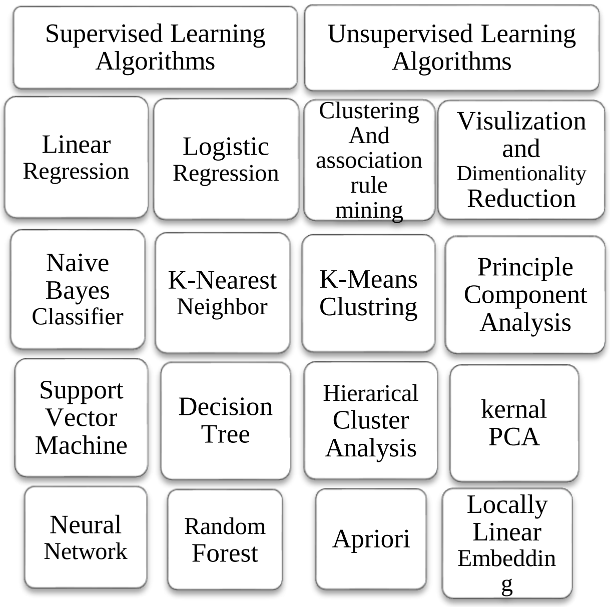
Supervised Learning Algorithms
Unsupervised Learning Algorithms
Linear Regression
Logistic Regression
Clustering And association rule mining
Visulization and Dimentionality Reduction
Naive Bayes Classifier
K-Nearest Neighbor
K-Means Clustring
Principle Component Analysis
Support Vector Machine
Decision Tree
Hierarical Cluster Analysis
kernal PCA
Neural Network
Random Forest
Apriori
Locally Linear Embeddin g

Fig. 5. Data mining techniques

In "Table. II" We clearly identify some of the important issues related to healthcare sector where we further need to improve our existing articles. Listed problems in the above table is related to heart disease, cervical cancer, Lung disease, skin cancer, Parkinson disease, Breast cancer, Automatic text classification and many more. In all these issues classification plays an important role in predicting and diagnosis of disease in early stages to avoid future risks. Modifications may include hyper parameter tuning or optimizations in the existing algorithms. In classification firstly, the classifier is trained using the training Dataset to understand the way how given input variables are associated with the class. After the classifier is trained, it is applied on test dataset then on the basis of testing dataset its accuracy is calculated. Classification problem can be binary or multiclass. In binary classification, exactly two classes should be there like "yes" or "no" but there are more than two targets in multiclass approach for example, "high", "medium" and "low.

1) Naïve Bayes classifier

NBC is one of the simplest methods of classification that is based on the Bayes theorem. In order to implement, it assumes that attribute independence property holds good i.e. occurrence of one certain feature does not disturb the occurrence of other features. If the attributed are not related then it will give more accurate results. But in majority of cases in real life predictors are dependent that effect its performance. It is widely used in field of healthcare to predict disease in patients on the basis of symptoms. Statistically, it performs better and is computational easy on large datasets. But Imbalanced dataset creates a problem and reduces the performance of a classifier. Aridas, Christos K et al. (2019) using Naïve Bayes classifier learning under unbalanced datasets and describes the ways to handle imbalanced dataset. It is based on the concept of Bayes Theorem. Mathematically if M and N are two events then it is given as:

$$\boldsymbol{P\left(\frac{M}{N}\right)=\frac{P\left(\frac{N}{M}\right)P(M)}{P(N)}} \tag{1}$$

It finds the probability of happening of event M when event N is already happened. Thus, it describes probability of some event based on the previous knowledge of some other event. *Shameem Fathima* developed a decision support system for predicting the Arboviral disease-Dengue using Naïve Bayes and SVM. It uses datasets containing 5000 records and 29 features (*Fathima & Hundewale ,2011*). But it has less accuracy so we can improve its accuracy and helps to resolve future issues related to it.

- Requires predictors to be independent of each other.
- Requires outcome is equally affected by predictors
- Fast and computationally easy.
- Used in recommendation systems, spam filtering, sentiment analysis.

2) K Nearest Neighbor

It is the simplest, non-parametric and lazy learning classification algorithm that works only on the stored available data and classifies the new input only on feature similarity measure. It classifies unlabeled data point on the bases of maximum vote of k nearest neighbor to each point. Where K refers to the count of nearest neighbor in KNN. If k=5 then for 5 nearest neighbor whichever label the maximum of its neighbor is the label of the new data point M.

- Can be used for Regression and classification Problem.
- *Scaling/Normalization*: It is based on the distance function so it is important to scale our dataset features in the same range to get rid from bias and improve performance.

How selects K?

- One way to select K using cross validation i.e. select small slice from training dataset as validation set and use it to select best value of K. Run a loop and take value of K that performs best on validation set with minimum error rate then use that value to test. In general take K=SQRT(N), N is the number of samples.
- Simple to handle multi classes and flexible to Distance choices.

Senthil Kumar Mohan et al. (2019) uses KNN classifier for the diagnosis of heart disease prediction. Kumar & Thakur (2020) published an article for the diagnosis of liver disorders Using KNN on three datasets using BUPA, ILPAD from UCI and MPRLPD datasets. Also compared KNN with fuzzy KNN on imbalanced datasets shows that KNN is not good for imbalanced dataset. *W. L. Zuo* (Zuo et al. ,2013) diagnosis a *Parkinson disease* using modified Fuzzy KNN approach that increases accuracy. The statistical estimation is based on the distance function Euclidean, Manhattan distance and Minkowski distance functions.

$$Euclidean\ = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2} \tag{2}$$

$$Manhattan = \sum_{i=1}^{m}|x_i - y_i| \tag{3}$$

$$Minkowski = \left(\sum_{i=1}^{m}(|x_i - y_i|)^q\right)^{\frac{1}{q}} \tag{4}$$

3) Support Vector Machine (SVM)

SVM is more efficient and extremely popular classification algorithm introduced in 1960. It can perform classification, regression as well as outlier detection. It was developed to resolve the decision boundary problem of logistic regression. Because decision boundary is selected arbitrarily in Logistic regression. Thus**,** its aim is to create a hyperplane in N dimensional feature space to separate data points in distinct classes. Among classes different hyperplanes can be drawn to separate classes but our intuition is to take maximum margin to select the hyperplane. The distance on both sides of the hyperplane is called as margin. The nearest data points to the hyperplane is known as support vectors. It can easily separate linear plane but not linear planes are not easily separable. In order to separate non-linear planes, it converts low dimensional space into a high dimensional space. Rajkumar, N et al. (2013) uses improved RBF kernel and SVM for feature extraction and classification of *Dermatology* disease with the accuracy 95%. SVM kernel trick are used to separate the inseparable planes that make classifier accurate. SVM include *Hyper parameters* like c, gamma and selection of the kernel. Some frequently used Kernel functions are

$$Gaussian\ kernel = \exp\left(-\frac{||x-y||^2}{2\,\sigma^2}\right) \tag{5}$$

$$Sigmoid\ kernel = \tanh(\gamma.x^TY + r) \tag{6}$$

- Can handle Linear as well as non-Linear problems.
- Highly effective and accurate in high dimensional spaces.
- Effective if count of dimensions is more than the count of samples
- Robust to outlier and noise than logistic regression.
- It is highly memory efficient classification model than any other models.
- Used in Bioinformatics, handwriting recognition, image classification.

LIAQAT AL et al. developed an improved version of article for the prediction of *heart disease* using UCI Cleveland database containing 303 instances. The proposed model uses SVM RBF with improved accuracy of 92.8% by taking 8 parameters in consideration. *Alzheimer's disease* is a common disorder that destroys brain cells**.** Huang & Lu (2013) developed a model for Alzheimer's disease using SVM-based Adaboost to that uses 100 MRI samples with accuracy score of 84.30%.

4) Decision Tree

It is a classification algorithm and is similar to a tree structure containing root, branches and leaf node. In Decision tree, every branch represents an outcome of that test, every non-leaf node represents a test on a particular attribute and every leaf node denote a class label. For the classification of the Alzheimer 's disease Zhang et al., (2014) developed an efficient model by using decision tree, KNN, SVM algorithms. In particular, the lowest split cost is chosen that makes root as a best predictor. This greedy approach is also known as CART (Classification and Regression Trees). By using the training data one at a time, it learns sequentially. Each time a rule is learned, the tuples covering the rules are removed and it continuous until the termination point is met.

$$Entropy = -\sum_{j=1}^{m} p_{ij}log_2p_{ij} \tag{7}$$

DT is a top down recursive divide and conquer approach (DAC) is used to construct a tree based on high entropy inputs.

- It can perform feature selection/variable screening.

- Capable of handling numeric and categorical data as well as multiclass problems.
- Problem of Low bias, high variance and overfitting is lowered by bagging and boosting.
- Runs fast even with more features but expensive to train.

A common nervous disorder is a Parkinson disease effects movement of a person. Athanasios Tsanas et al. **(**2012) developed a model for the treatment of Parkinson disease using improved decision tree by taking dataset from UCI machine learning repository. Also, Kaya, et al., (2017) C4.5 Decision tree for classification and feature selection for diabetes retinopathy from VOG signals.

5) Random Forest

Random forest is kind of ensemble classifier that uses decision tree algorithm in a randomize way. In this the bootstrap datasets are created with the help of sampling with replacement form original dataset. Then a number of decision trees are created using bootstrap datasets created at training time. After training when a new test tuple is given then its output will be the based-on majority vote (classification) of the all the decision trees. It can be used to solve classification (*MODE*) as well as regression (*MEAN*) problem. When the data is continuous it acts as regression and its output will be the mean of all the decision trees. Hualing et al. (2019) classifies the *air quality* using novel Random forest approach and its impacts on the healthcare. Thus, the advantage of using with multiple decision trees in RF it has low variance, less error rate hence more accurate results. Also, it reduces overfitting problem present in decision tree. For increasing performance, we can change various hypermeters. some of its hyperparameters include max_features (for splitting a particular node), n_estimators (number of trees), max_depth, min_samples_leaf, min_samples_split, bootstrap (methods for sampling).

- Can handle numeric as well as categorical data
- Low variance, less error rate and reduces overfitting.
- Increase max_features, n_estimators and max_depth improves performance but to maintain the right balance for optimal feature selection.

6) Artificial neural network (ANN)

ANN is a collection of the smallest unit called neurons(nodes) and is inspired by the biological functioning of mind. ANN forms a network that consists of input layer, hidden layers and output layer. There can be one or more than one hidden layer. In this network, nodes are interconnected and work in parallel to give output function. Nodes and connections are associated with a weight. Weights can be adjusted to increase or decrease the strength at the connection. The output is given by the activation function by adding bias value. It is good for dealing with noisy data but has high complexity and requires many parameters. It can be feedforward or feed backward neural network. It can be used for classification and pattern recognition problems. Chronic disease is one of major cause of death today. *Jiongming Qin, et al*., (2019) presented an article for novel methodologies of machine learning for early detection of chronic diseases with less error rate and predict more accurate results. In medical field, it can be used for the prediction of cancer.

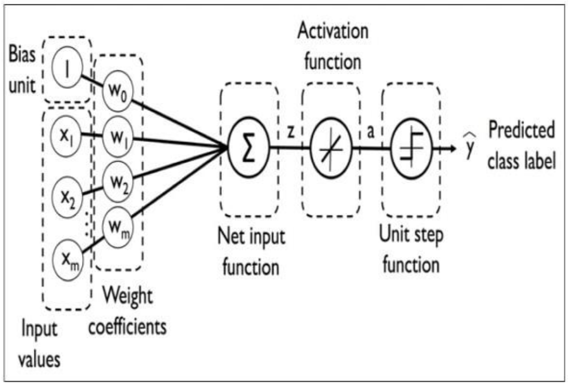

Fig. 6. Artificial Neural Network

Shariati et al. (2010) diagnosis thyroid and hepatitis disease by comparing SVM, ANN, ANFIS. In this research he uses 221 samples of thyroid and 250 samples of hepatitis. Also, *Pulmonary disease* is a lung disease that causes breathing problem which was solved by Xu et al., (2019) by using a novel approach involving ANN for diagnosis of pulmonary disease containing 18471 real clinical records with accuracy of 86.45% and f1-score of 82.93. Er et al. (2010) uses ANN for predicting a *chest disease* in a patient and also a comparative study was completed using Multilayer, probabilistic neural networks and regression model.

a) Activation Functions

Activation function plays very important task in as it is foundation of neural network .If "$\{x_1, x_2, x_3, \dots\dots\dots, x_n\}$" be the inputs with their respective weight "$\{w_1, w_2, w_3, \dots\dots\dots, w_n\}$" to a neuron, then the neuron computes " $\sum_{i=1}^{n} x_i w_i$ " . The task of activation function is to give final output by adding a bias to the weighted sum. AF is useful as it introduces *Non-linearity* so that it can deal with any nonlinear variable. Without activation function it would behave just like single layer perceptron which is incapable of computing nonlinear functions like *XOR*, *EXNOR* etc. But multi-layer perceptron can be able to separate any nonlinear function. Thus, ANN is also known as Universal function approximation. Some of the desirable features of activation function are Non-linearity, Range, and Monotonic, continuously differentiable and approximate identity near the origin. *Isa et al. (2010)* proposed an approach for diagnosis of *thyroid and breast cancer* using multilayer perceptron with various activation functions. Some useful activation Functions are listed in "Table III" are as follows:

Table III. Activations Functions

Function	Equation	Range
Linear Function	$f(y) = y$	$(-\infty, \infty)$
Heaviside step	$f(x) = \begin{cases} 0 \ \ for\ x < 0 \\ 1 \ \ for\ x \geq 0 \end{cases}$	$\{0\ ,1\}$
Sigmoid	$A = \dfrac{1}{1 + e^{-t}}$	$[0,1]$
Tanh	$f(y) = \dfrac{2}{(1 + e^{-2y})} - 1$	$(-1, +1)$
ArcTan	$f(y) = \tan^{-1} y$	$\left(-\dfrac{\pi}{2}, \dfrac{\pi}{2}\right)$
ReLU	$A(y) = \max(0, y)$	$[0\ , \infty)$
Leaky ReLU	$f(y) = \begin{cases} 0.01x \ \ for\ x < 0 \\ x \quad\ \ for\ \ x \geq 0 \end{cases}$	$(-\infty, \infty)$
SoftMax	$f(y_i) = \dfrac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}}\ for\ i = 1\ to\ k$	$(0,1]$
Gaussian	$f(y) = e^{-y^2}$	$(0,1)$

Another important thing we need to keep in mind while making ANN is the choice of activation function. Every activation function has some advantages and disadvantage.

Bircanoğlu & Arıca (2018) compares various activation functions in ANN**.** There are lot of activation function but their choice is dependent upon couple of factors. There are several issues with these activation functions. Sigmoid or logistic function has a saturation problem. Because for high input value it gives high output value and for low input value it gives low output value. Thus, reaches its peak value either maximum or minimum and vanish gradients. Sigmoid function converges very slowly. Also, it isn't zero centered. Tanh function vanish the problem of not zero centered of sigmoid function. Still it has problem of saturation and also requires lot of computation. *ReLU* is the simplest function as it output the number itself if it is positive and 0 if the number is negative. For the positive values of weighted sum, it does not saturate and it is also not zero centered. But in ReLU neuron dies for large negative values. To resolve the problems of ReLU another activation function Leaky ReLU is given. In *LReLU* due to multiplicative factor 0.01 neurons do not die. Also, it is zero centered and has no saturation problem. Sometimes when classes are associated with different probability SoftMax or logistic function can be used. *Logistic* is preferred for binary classification problem whereas SoftMax is preferred for multiclass classification problems in the output layers. On the basis of the requirement of the problem we need to use an activation function. According to thumb rule,

- Firstly, in general start with ReLU as simplest one if it doesn't perform better then switch to other one.
- Tanh and sigmoid does not preferred where network have many layers because of gradient vanishing problem.
- ReLU vanishes gradient decent problem and performs better and learn faster.
- For CNN, Treat ReLU as a standard activation function but use Leaky ReLU if dead neuron situation occurs.
- ReLU should not be used in output layer but applied only to hidden layers.
- Sigmoid is preferred in *output layer* but not in hidden layer.
- Tanh function is preferred in *RNN*.
- Radial Basis function and Sigmoid both can be effectively utilized by Support vector machine.

Yahaya & Isa (2011) proposed a hybrid model HMLP by comparing different activation functions for the classification of *cervical cancer*.

B. Regression Model

Regression model is used for predicting a *continuous quantity* and more specifically, establish if there is a statistically significant relationship between variables/features. It predicts a dependent variable based on one or more independent variables. In statistical modelling, Regression model can have multiple variables called as multivariate regression problem. In Regression end result is to predict a continuous quantity rather than a label data.

1) Linear and Logistic Regression

Linear regression is a regression model used to predict a dependent/target variable(Y) on the basics of independent /predictor variable(X). Mathematically it is represented a “$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta\_n\ X\_n\ \ (Y\ is\ a\ response\ variable\ to\ be\ predicted,\ \beta\_0\ \ is\ the\ \ y\ intercept\ )$ “ .Always there is a linear relationship between target *(continuous variable)* and predictor variable that makes a best fit straight line.

Characteristics: Continuous dependent variable type, infinite possible values, predict integer output, least square estimation method used, best fit straight line, used in forecasting sales and business domain.

Logistic regression is a supervised learning technique that can be used for both classification as well as regression problem. As Linear regression can predict only if the outcome feature is continuous in nature but if outcome feature is *categorical* then, logistic regression is used. Polat (2019) proposed a model for diagnosis of Parkinson's disease by using and comparing various activation functions for improving accuracy. Sisodia & Verma, (2017) also proposed a model for prediction of chronic kidney disease using various classification algorithms but logistic regression outperforms all with accuracy 86.73% where Fast Fourie**r** transformation is used to extract features. It can predict a binary, Multinomial and Binomial dependent variable on the basics of one or more independent variables. Thus, a linear relation is not mandatory and best fit line forms a curve rather than straight line. The logistic function or sigmoid function defined in "Table III" forms its basics.

$$LR = \frac{1}{1+e^{-(Y=\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n)}} \qquad (8)$$

Characteristics: categorical dependent variable, definite outcomes, predict binary value (1 or 0), Maximum likelihood estimation used, best fit line is curve, used in disease diagnosis, weather prediction.

C. Ensemble Learning

It is a machine learning technique that is used to improve performance and reduce error rate. After understanding above explained machine learning supervised learning models, we can select best model by evaluation metrics explained below. Then by combining the several best models we can make a new model by a technique called ensemble learning. Combining of several learning models gives better predictive power then a single machine learning model. Ensemble learning tries to minimize the bias, variance and noise that improves the accuracy and stability of the machine learning models.

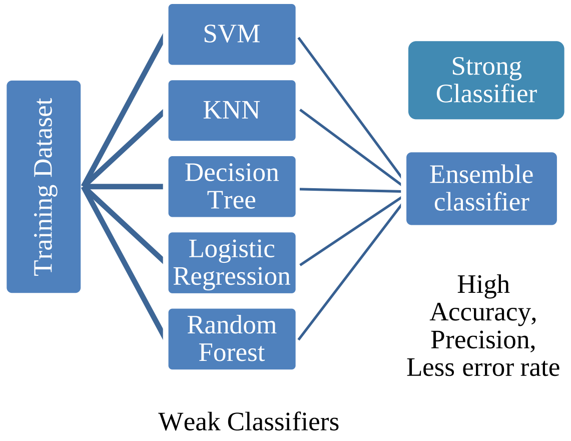

Fig. 7. Ensemble learning

In "Fig. 7" suppose we combined different weak classifiers like SVM, KNN, Decision tree, Logistic Regression and Random forest trained on different training dataset to get a unique strong classifier giving high accuracy, precision and less error rate then individual weak classifier**.** *Senthil Kumar Mohan et.al (2019)* proposed a hybrid model for the prediction of heart disease. In this research an ensemble learning approach i.e. hybrid random forest with linear model is used. Initially the implementation of Logistic regression (82.9%), naïve Bayes (75.8), decision tree (85), random forest (86.1), support vector machine (86.1) and generalized linear model (85.1) with their respective accuracy. Finally, a new hybrid model is proposed by combining decision tree, random forest and linear model. This proposed model gives an accuracy of 88.4% with error rate as reduced to 11.6.

1) Bagging

It is an ensemble learning technique in which various homogeneous weak learners learns in parallel independently from each other. Finally, all independent learners are combined to follow weighted average for predictive analysis. Aim of bagging is to reduce overall variance and to produce an ensemble model more robust then individual learner. In bagging from original training dataset D containing n records, new D_i datasets (bootstrap samples) are generated from D each containing m records by sampling with replacement. Every record has equal probability to appear in a dataset. These bootstrap samples are then given to different weak learners for training in parallel. It is also known as bootstrap aggregating. E.g. *Random Forest.*

2) Boosting

It is a sequential ensemble learning technique which generates a new learner by considering into account the success of its prior learner. In this samples are selected with replacement over weighted data. Consider dataset D containing n tuples with weights $"w_1, w_2, w_3, \dots, w_n"$ Initially dataset D_1 is created by random sampling with replacement with equal probability. Model M_1 is trained on D_1 and dataset d is tested on it. Each misclassified sample weight is increased and included in dataset D_2 for training model M_2 then again dataset D is tested by model M_2. This process continuous till M_n. After all the model are trained a new strong model is trained using the majority vote of all the models on test datasets. and their weights are redistributed after every training step. Samples that performs well on training data are assigned more weight. For e.g. *GradientBoost, AdaBoost, XGBoost, LPBoost* etc.

IV. IMPACT OF METHODS OF SPLITTING TRAINING AND TEST DATA

After the model or algorithm is developed, next step is to train and test the model to evaluate its performance.Model is trained and tested using different training and testing datasets.But we have an availability of finite dataset so how we can use it for both training and testing.In general,input vectors and their corresponding targets are given in training data.In order to get better generalizations or result one should use more and more data to train the model.But for the estimation of better error probability of classification model more and more data is used for testing.For better evaluation purpose the data used for testing has not been part of training data.We can use the whole data to train our final model/classifier only after evaluation is finished.We can split our dataset on the basis of percentage ,random selection or k-fold validation method using train_test_split function of sklearn.For the partitioning of the available dataset into training/testing dataset we can use following method:

1) *Holdout Method*

In holdout method whole dataset is divided into two parts:train and test set.Train set is used to train the classifier.Test dataset is used to know how well our model perform on unknown data.Usually the training dataset is more than the test dataset.But in general to split in 75 (training) and 25 (testing) ratio is best. It is good in case of very large datasets. In this there is only one train_test_split, thus it takes less time.Sometimes, a validation dataset is also chosen to select the best hyperparameters for better performance evaluation and model selection.Another variant of this method is repeated holdout method.In repeated holdout method random sampling is done to select train-test set and holdout method is repeated k times .Accuracy is computed by taking average of the sum of accuracy all iterations.But this method has drawback of overlapping problem.

$$A = \frac{\sum_{i=1}^{n} A_i}{n} \qquad (9)$$

2) *Cross Validation Method*

K-Fold validation is a method in which entire dataset is divided into k random splits each having same size.For testing purpose only 1 group is used while remaning k-1 groups are used for training the classifier.we will continue this process till each group acts a a test set.suppose in 4-Fold cross validation it is divided into 4 groups and process is repeated for 4 times.In cross validation technique multiple train-test splits are used so this will give better results then holdout method.Also it resolve the overlapping problem of holdout method because different split has different test set. But it will consume more time and computational power then holdout method.Mean of k errors will give final estimated error.Although for better performance or accuracy, use k Fold validation with (k=5 or K=10).This is also useful in checking if the model is overfitted or not.In this practitioner need not to choose separate validation dataset. Yadav & Shukla, (2016) uses colossal dataset to analyse the impact on performance using k-fold cross validation over holdout approach. In "Table IV" 4-Fold Cross Validation is shown with labeled 3 training group and 1 test group in each split.

Table IV. 4-Fold Validation

FOLD 1	FOLD 2	FOLD 3	FOLD 4	Split1
FOLD 1	FOLD 2	FOLD 3	FOLD 4	Split2
FOLD 1	FOLD 2	FOLD 3	FOLD 4	Split3
FOLD 1	FOLD 2	FOLD 3	FOLD 4	Split4
TRAINING			TEST	

V. STEPS FOR BEST MODEL SELECTION

When we come across solving any problem, first step is to understand the problem, collect dataset and formulate problem statement. After collecting data apply data pre-processing steps (explore the data) and feature extraction to make the dataset relevant to our Model. After the extraction of relevant feature, we have to develop our model and split our data set into training and testing according to above explained methods. Once the model is trained, we have to evaluate our model using above explained evaluation metrics. Although our developed model may contain overfitting or underfitting issue. Thus, we need to refine, optimize and re-evaluate our model. Finally, on the basics of above explained evaluation parameters best model is selected and can be further ensembled to generate new hybrid models. In "Fig. 8" we illustrated the steps required to follow in order to select best model for a problem

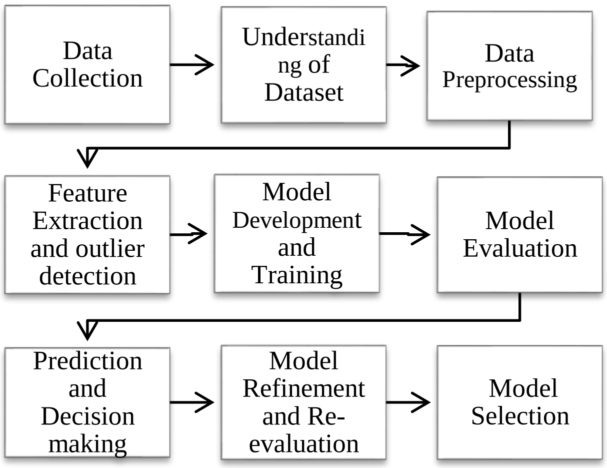
Data Collection
Understandi ng of Dataset
Data Preprocessing
Feature Extraction and outlier detection
Model Development and Training
Model Evaluation
Prediction and Decision making
Model Refinement and Re-evaluation
Model Selection

Fig. 8. Algorithmic Steps

VI. PERFORMANCE EVALUATION METRICS

Evaluation is the major step to understand whether our Machine learning algorithms performs well or not. After completing feature engineering and training the model it is important to test how effective our model performs on test dataset.

A. Classifier Evaluation Metrics

The goal of classification model is to predict unknown class label from the known one. In classification, we have different performance metrics for the evaluation. Some of them are accuracy, log-loss precision, recall etc. Every performance metrics has merits and demerits. *Sokolova et al. (2006)* an article for the better performance measure on different conditions using holdout and cross validation method. Thus, the choice to select right performance metrics is important. In classification, the problem can be binary or multiclass and the output is either class label or in probability form. For measuring the performance of a classification model confusion matrix is used. Confusion matrix is a table that can represent binary as well multiclass model. Suppose there are two classes positive (P) and negative (N). Then confusion table of size 2 X 2 having four possible outcomes is shown in "Table V".

Table V. Confusion Matrix

CONFUSION MATRIX		PREDICTED	
		POSITIVE	NEGATIVE
ACTUAL	POSITIVE	True Positive (TP)	False Negative (FP)
	NEGATIVE	False Positive (FP)	True Negative (TN)

Majority of the performance metrics are based on this table. This table is constructed between actual output and predicted output on the test dataset and represents summary of predicted outcome. In below confusion matrix entries in the diagonal (TP and TN) are correctly predicted.FP and FN is an error associated with it. FP is a type I error and FN is the type II error .

- If actual value is positive and is predicted as postive i.e postive value is correctly classified is known as TP.
- If actual value is negative and is predicted as negative is known as TN(correctly classified).
- If actual value is positive but is incorrectly predicted as negative is known as FP(incorrectly classified).
- If actual is negative but is predicted as positive is known as FP(incorrectly classified)

In particular, on the basis of type of dataset we need to select metrics to compute performance.A dataset can be balanced and imbalanced. It is said to be balanced if the number of samples that are positive is similar to the number of negative samples. Assume there are 1000 records if out of 1000 (500 positive and 500 negative or 600 positive and 400 negative) it is said to be balanced.If 800 are positive and 200 are negative it is imbalanced problem. It is essential to have balanced dataset so that there will be no bias and performance evaluation is easier.

1) Accuracy

It is a performance metrics which gives reliable result only if the dataset is nearly balanced.It is the ratio of sum of correctly classified samples to all the samples.If dataset is imbalanced recall,precision and F1-score is the better choice to use then accuracy. Most accurate results are given by TP and FP so always try to increase their values. FP, FN are type I and type II error so always minimize their values. It is not considered reliable anymore when majority of data belong to only one class.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (10)$$

2) Recall(TRUE POSITIVE RATE)

It determines that out of total actual positive values how many positive did we predict correctly.It is useful when we like to predict more positives as positive.It is more about finding how many patients that actually had heart disease is diagnosed correctly by algorithm.In recall FN is more important and we try to minimize FN.In particular,it focus more on picking all patients having heart disease rather than picking heart disease patient correctly. If out of 50 people 10 people has disease but our model said all have disease(FN=0 then FP+FN=10) recall is 100 %. It is also known as sensitivity.

$$Recall = \frac{TP}{TP + FN} \qquad (11)$$

3) Precision(positive prediction value)

It determines out of total predictive positive value how many were actually +ve. It is more precise about result. It is more about how many people that are diagonsed having heart disease actually had heart disease.In precesion FP is given more importance and we try to minimize FP to get 100% precision. In particular,it is more about capturing result correctly.

$$Precision = \frac{TP}{TP + FP} \qquad (12)$$

4) Specificity

It is a performance metric that is opposite to recall. It is about how many patient that do not have a disease,were predicted by the model not disease.In this FP is more important and it try to minimize FP. It is also known as true negative rate.

$$Specificity = \frac{TN}{FP + TN} \qquad (13)$$

5) F-Beta Score

In classification ,it is a performance metric used to measure accuracy of a model.Precision and recall both are important metrics in imbalanced problems but it is better to combine both to get more accurate output.F-Beta is based on both precision as well as recall(*Gallas & Frey,2009*). F-Beta is given as:

$$F_\beta = \frac{(1 + \beta^2)Precision \times Recall}{\beta^2 \times Precision \times Recall} \qquad (14)$$

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \qquad (15)$$

$$Harmonic\ mean = \frac{2\times X\times Y}{X+Y} \qquad (16)$$

F1-score is a performance metrics derived from F-Beta score when the value of $\beta eta = 1$. It is also known as hormonic mean of precision and recall which is m4ch better then the arthemic mean. F1-score is useful When both precision and recall values are equally important.If $precision = X\ and\ recall = Y\ then$ it is defined as:

According to thumb rule, whenever precision is given more weight (FP) than recall then reduce beta value in range 0<beta<1 to give $F_{0.5}\ measure$ .If recall is given more weight then precision then beta value is increased to give $F_2\ measure.$ F_β is applicable in various fields of information theory for the performance of query and document classification.In particular,F-measures take into account only false positive and false negative but true negatives are not given importance.

6) ROC Curve

An essential step in any machine learning algorithm is performance measurement.In classification problem for performance measurement ROC curve is a better choice. It is useful for visual comparison in classification models.Signal detection theory is the base point form where ROC curve is originated. It is commonly used metric in medical decision making systems. It can be used to check performance of binary as well as multiclasss.In ROC curve sensitivity(TPR/RECALL) is plotted against False positive rate.X-axis represents FPR and Y-axis represents TPR. ROC curve represents probability measure. Accuracy of classification model is a measure of area under ROC curve. If area under ROC curve is higher,it is better in separating classes. More area it has under it ,Better is the model associated with it. Closer it is towards 1, better it is in distuinshing classes. In worst case it is closer towards zero. It does not distinguish classes even if AUC is 0.5.Thus, ROC curve should lie between 0.5 to 1 for meaningful results.

$$FPR\ = 1 - Specificity \qquad (17)$$
$$= \frac{FP}{TN + FP}$$

7) Matthews correlation coefficient (MCC)

MCC is a classification performance metric developed for chemical structure comparison in 1975 by Matthews. In bioinformatics MCC is the most commonly used performance metric. Accuracy is not a perfect measure if classes are imbalanced then MCC is a better choice there. MCC takes into account all the four choices of 2X2 confusion metrics (TP, TN, FP, FN) by a single number even if the size of classes is not similar. Chicco & Jurman (2020) in an article describes the advantages of MCC over accuracy and F1 score, accuracy and precision. It returns -1 for total disagreement and +1 in case of perfect prediction. Although MCC is preferred to F1-score in evaluating due to two main reasons: Firstly, in case of class exchange from positive to negative and vice-versa F1 score varies but MCC does not varies. Second, In F1-Score, count of samples properly classified as negative will not be measured.

$$MCC = \frac{TP\times TN - FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

8) Log Loss

In performance evaluation of the classification model logarithmic loss is an essential performance metric. It is applicable for both binary and multiclass classification problems. It is based on uncertainty of our predicted input that measures how much it is different from the actual label. Range of our prediction input lies between 0 and 1. In particular, log loss value of a perfect model is 0 and the aim of our model is to reduce log loss close to zero.

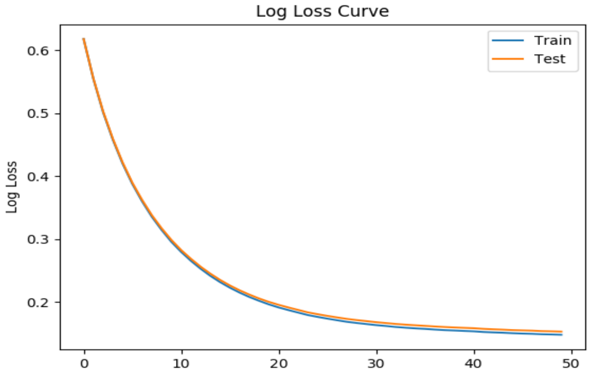

Fig. 9. Log Loss Curve

If log loss decreases the accuracy of the model increases as shown in “Fig. 9”. It clearly shows when the value of predicted probability increases, the log loss value decreases. In case of binary classification where (M=2) and p is the probability of predicting 1 then log loss is defined

$$= -(ylog(p) + (1-y)\log(1-p)) \qquad (19)$$

$$Log\ Loss = \frac{-1}{M}\sum_{i=1}^{M}\sum_{j=1}^{N} y_{ij} * \log(p_{ij}) \qquad (20)$$
$$= \begin{cases} y_{ij} = 1 & if\ i\ \epsilon\ j \\ \quad\ = 0 & otherwise \end{cases}$$

9) PR Curve

In order to measure the performance precision recall curve is widely used and is similar to ROC curve.It is used for evaluating the performance of binary classifiers.This curve is dependent on the two performance measures i.e. Recall and Precision.Recall is also recognized as true positive rate.Thus it is more concerned about all positive part of dataset Whereas precision is more concerned about predicting result correctly. It is easy to construct a precision recall curve by taking Recall along x-axis and precision along y-axis as shown in "Fig. 10".PR curve is more informative than the ROC curve. Thus sometimes it is preffered over ROC curve.

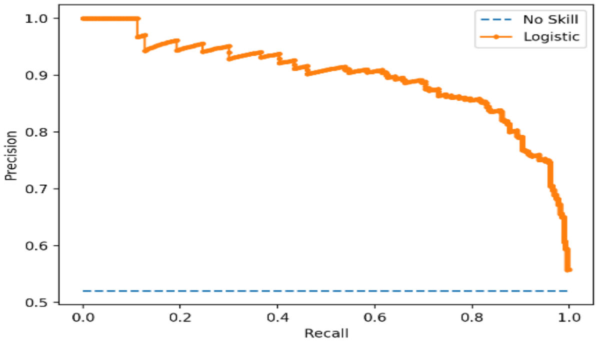

Fig. 10. Precision-Recall Curve

B. *Regression Performance Metrics*

Like classification models,for the performance evaluation of regression model we have metrics.Most commonly used metrics are Root mean square error(RMSE),Mean absolute error(MAE),Mean square error(MSE).

Root mean square error is most useful regression metrics that measures how much error our model makes in predictions. More the weight of RMSE means higher the error in predictions. Mathematically it is represented as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{m}(x^{i} - y^{i})^{2}}{m}} \tag{21}$$

Where m is the number of instances in dataset, x^i be the feature vector and y^i be the values to be predicted.

RMSE is most preferable one but it is not always used. It performs worst if our dataset contains outliers. It is similar to MSE but the root is not there. MSE is used particularly when we pay attention for certain high or low values. But even from single bad prediction MSE value is large thus RMSE is used mostly.

Mean Absolute Error is taken as an average of absolute difference among prediction and the target values. In the average of MAE all individual differences are weighted equally. Mathematically, its representation is

$$MAE = \frac{\sum_{i=1}^{m}|x^{i} - y^{i}|}{m} \tag{22}$$

Whenever there exist outliers in the dataset it is recommended to use MAE as it is less sensitive to outliers. Also, it penalizes huge errors. Thus, in case when outliers are rare RMSE performs exponentially well then MAE and should be preferred one.

VII. IMPORTANT TOOLS AND LIBRARIES

Prerequisites of data science are shown in Table VI.

Table VI. Pre-requires of Data Science

Integration development tools	Jupyter Notebook, Spyder, PyCharm, Thonny, Atom
Libraries	Scikit-Learn, Tensorflow, Keras, Pytorch
Data Analytics Tools	Pandas, NumPy, Matplotlib, seaborn, SciPy.
Data Visualization tools	TABLEAU, POWERBI, SAS, Qlik View.

DISCUSSION

The findings of this article suggest that how important our data is useful in finding solutions to existing and future arising problems. In this revolutionary world of technology people are attracted towards the field of data science. Mainly, we focused on how data science can analyse a huge volume of data that conventional technology failed to handle. In this article we survey a number of IEEE, springer high quality articles related to data science techniques in the field of healthcare that needs an improvement in future work. We mainly focused on the road map for data pre-processing, feature engineering and supervised learning techniques in healthcare. Moving onwards we discussed some of useful classification techniques and review their use cases in healthcare domain to optimize and improve healthcare systems. Finally, the right selection of performance metrics and activation functions dependent on the type of dataset and algorithm that highly impacts performance is discussed. This article will help in identifying a right problem, activation function and performance measure for the application of data science in healthcare domain.

CONCLUSION

The main aim of this article is to insight towards the identification of problems in healthcare and the roadmap for the selection of right features, algorithm, performance metrics and activation function on the basics of available problem. We review all the major problems related to our health from IEEE, Springer high quality recent articles with their essential details. Thus, it will help to identify the relevant problems in an easiest way. In the field of healthcare our data is extremely useful and contain some valuable patterns that helps to predict disease in the early stages. Data science is an emerging field that can utilize healthcare data efficiently and improves healthcare systems. After the identification of the relevant problem this provides us the roadmap for the steps involved to solve every problem effectively. Our data may contain lot of inconsistencies, firstly it guides to pre-process the dataset and to select relevant features from the ample of attributes. Also need to remove noise, inconsistencies and outliers that impacts the performance of the classifier. It helps to make dataset relevant for analysis**.** It also helps to select and select right model on the basis of the availability of the dataset because no such model performs well on every dataset. After the selection of model, it guides the right ways to split the dataset for better model development and performance. Finally, it guides to select right performance metrics for the evaluation of our model that depends upon the type of available problem i.e. in case of imbalanced problem accuracy is not considered best measure. Thus, this article helps to select the best model for a problem. In ANN also we analysed that the choice of activation function also impacts performance. Thus, it helps to select the right activation function in the different layers of ANN. To sum up the entire we conclude that this will help to identify relevant problem in healthcare and guides the various steps required to give improved solution of the problem.

In future we can select any of the reviewed problem in this article to apply data science to improve the limitations in existing problem to make healthcare systems better. Also, i recommend to improve accuracy and to use ensemble learning to develop hybrid models for better prediction. Usage of deep learning and neuro fuzzy for classification in above problems will give better results.

REFERENCES

Davenport, T., & Patil, D. J. (2013). Data Scientist: The Sexiest Job of the 21st Century-Harvard Business Review. Harvard Business Review.

Raban, D. R., & Gordon, A. (2020). The evolution of data science and big data research: A bibliometric analysis. Scientometrics, 122(3), 1563-1581.

Dhar, V. (2013). Data science and prediction. Communications of the ACM, 56(12), 64-73.

Leek, J. (2013). The key word in 'Data Science' is not Data, it is Science. Simply Statistics, 12.

West, D. M. (2012). Big data for education: Data mining, data analytics, and web dashboards. Governance studies at Brookings, 4(1), 1-10.

Bhat, A. Z., & Ahmed, I. (2016, March). Big data for institutional planning, decision support and academic excellence. In 2016 3rd MEC International Conference on Big Data and Smart City (ICBDSC) (pp. 1-5). IEEE.

Bollier, D., & Firestone, C. M. (2010). The promise and peril of big data (pp. 1-66). Washington, DC: Aspen Institute, Communications and Society Program.

Hand, D., Mannila, H., & Smyth, P. (2001). Principles of data mining. 2001. MIT Press. Sections, 6, 2-6.

Koh, H. C., & Tan, G. (2011). Data mining applications in healthcare. Journal of healthcare information management, 19(2), 65.

Xu, Q., Tang, W., Teng, F., Peng, W., Zhang, Y., Li, W., ... & Guo, J. (2019). Intelligent Syndrome Differentiation of Traditional Chinese Medicine by ANN: A Case Study of Chronic Obstructive Pulmonary Disease. IEEE Access, 7, 76167-76175.

Usama, F., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37-54.

Kumar, S., & Katyal, S. (2018, July). Effective analysis and diagnosis of liver disorder by data mining. In 2018 international conference on inventive research in computing applications (ICIRCA) (pp. 1047-1051). IEEE.

Kaur, S., & Grewal, A. K. (2016). A Review Paper on Data Mining Classification Techniques for Detection of Lung Cancer. International Research Journal of Engineering and Technology (IRJET), 3(11), 1334-1338.

Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE Access, 7, 81542-81554.

Omondiagbe, D. A., Veeramani, S., & Sidhu, A. S. (2019, April). Machine Learning Classification Techniques for Breast Cancer Diagnosis. In IOP Conference Series: Materials Science and Engineering (Vol. 495, No. 1, p. 012033). IOP Publishing.

Osmanović, A., Halilović, S., Ilah, L. A., Fojnica, A., & Gromilić, Z. (2019). Machine learning techniques for classification of breast cancer. In World Congress on Medical Physics and Biomedical Engineering 2018 (pp. 197-200). Springer, Singapore.

Huda, S., Yearwood, J., Jelinek, H. F., Hassan, M. M., Fortino, G., & Buckland, M. (2016). A hybrid feature selection with ensemble classification for imbalanced healthcare data: A case study for brain tumor diagnosis. IEEE access, 4, 9145-9154.

Singh, S. K., & Goyal, A. (2020). Performance Analysis of Machine Learning Algorithms for Cervical Cancer Detection. International Journal of Healthcare Information Systems and Informatics (IJHISI), 15(2), 1-21.

Abdoh, S. F., Rizka, M. A., & Maghraby, F. A. (2018). Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques. IEEE Access, 6, 59475-59485.

Verma, A. K., Pal, S., & Kumar, S. (2019). Classification of skin disease using ensemble data mining techniques. Asian Pacific Journal of Cancer Prevention, 20(6), 1887-1894.

Abbas, M. R., Nadeem, M. S. A., Shaheen, A., Alshdadi, A. A., Alharbey, R., Shim, S. O., & Aziz, W. (2019). Accuracy Rejection Normalized-Cost Curves (ARNCCs): A Novel 3-Dimensional Framework for Robust Classification. IEEE Access, 7, 160125-160143.

Das, R. (2010). A comparison of multiple classification methods for diagnosis of Parkinson disease. Expert Systems with Applications, 37(2), 1568-1572.

Haq, A. U., Li, J. P., Memon, M. H., Malik, A., Ahmad, T., Ali, A., ... & Shahid, M. (2019). Feature selection based on L1-norm support vector machine and effective recognition system for Parkinson's disease using voice recordings. IEEE Access, 7, 37718-37734.

Fitriyani, N. L., Syafrudin, M., Alfian, G., & Rhee, J. (2019). Development of Disease Prediction Model Based on Ensemble Learning Approach for Diabetes and Hypertension. IEEE Access, 7, 144777-144789.

Dey, S. K., Hossain, A., & Rahman, M. M. (2018, December). Implementation of a web application to predict diabetes disease: an approach using machine learning algorithm. In 2018 21st international conference of computer and information technology (ICCIT) (pp. 1-5). IEEE.

Roopa, H., & Asha, T. (2019). A Linear Model Based on Principal Component Analysis for Disease Prediction. IEEE Access, 7, 105314-105318.

Sampath, R., & Saradha, A. (2015). Alzheimer's Disease Classification Using Hybrid Neuro Fuzzy Runge-Kutta (HNFRK) Classifier. Research Journal of Applied Sciences, Engineering and Technology, 10(1), 29-34.

Rajaraman, S., & Antani, S. K. (2020). Modality-Specific Deep Learning Model Ensembles Toward Improving TB Detection in Chest Radiographs. IEEE Access, 8, 27318-27326.

Qin, J., Chen, L., Liu, Y., Liu, C., Feng, C., & Chen, B. (2019). A Machine Learning Methodology for Diagnosing Chronic Kidney Disease. IEEE Access, 8, 20991-21002.

Poudel, P., Illanes, A., Ataide, E. J., Esmaeili, N., Balakrishnan, S., & Friebe, M. (2019). Thyroid Ultrasound Texture Classification Using Autoregressive Features in Conjunction with Machine Learning Approaches. IEEE Access, 7, 79354-79365.

Zhang, Z., Zhao, M., & Chow, T. W. (2012). Binary-and multi-class group sparse canonical correlation analysis for feature extraction and classification. IEEE Transactions on Knowledge and Data Engineering, 25(10), 2192-2205.

Li, H., Wang, X., Liu, C., Wang, Y., Li, P., Tang, H., ... & Zhang, H. (2019). Dual-Input Neural Network Integrating Feature Extraction and Deep Learning for Coronary Artery Disease Detection Using Electrocardiogram and Phonocardiogram. IEEE Access, 7, 146457-146469.

Aridas, C. K., Karlos, S., Kanas, V. G., Fazakis, N., & Kotsiantis, S. B. (2019). Uncertainty based under-sampling for learning Naive Bayes classifiers under imbalanced data sets. IEEE Access.

Fathima, S., & Hundewale, N. (2011, November). Comparison of classification techniques-SVM and naives bayes to predict the Arboviral disease-Dengue. In 2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW) (pp. 538-539). IEEE.

Zuo, W. L., Wang, Z. Y., Liu, T., & Chen, H. L. (2013). Effective detection of Parkinson's disease using an adaptive fuzzy k-nearest neighbor approach. Biomedical Signal Processing and Control, 8(4), 364-373.

Kumar, P., & Thakur, R. S. (2020). Liver disorder detection using variable-neighbor weighted fuzzy K nearest neighbor approach. Multimedia Tools and Applications, 1-21.

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

Ali, L., Niamat, A., Khan, J. A., Golilarz, N. A., Xingzhong, X., Noor, A., ... & Bukhari, S. A. C. (2019). An optimized stacked support vector machines based expert system for the effective prediction of heart failure. IEEE Access, 7, 54007-54014.

Huang, L., Pan, Z., & Lu, H. (2013, August). Automated Diagnosis of Alzheimer's Disease with Degenerate SVM-Based Adaboost. In 2013 5th International Conference on Intelligent Human-Machine Systems and Cybernetics (Vol. 2, pp. 298-301). IEEE.

Zhang, Y. D., Wang, S., & Dong, Z. (2014). Classification of Alzheimer disease based on structural magnetic resonance imaging by kernel support vector machine decision tree. Progress in Electromagnetics Research, 144, 171-184.

Tsanas, A., Little, M. A., McSharry, P. E., Spielman, J., & Ramig, L. O. (2012). Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. IEEE transactions on biomedical engineering, 59(5), 1264-1271.

Kaya, C., Erkaymaz, O., Ayar, O., & Özer, M. (2017, October). Classification of diabetic retinopathy disease from Video-Oculography (VOG) signals with feature selection based on C4. 5 decision trees. In 2017 Medical Technologies National Congress (TIPTEKNO) (pp. 1-4). IEEE.

Yi, H., Xiong, Q., Zou, Q., Xu, R., Wang, K., & Gao, M. (2019, July). A Novel Random Forest and its Application on Classification of Air Quality. In 2019 8th International Congress on Advanced Applied Informatics (IIAI-AAI) (pp. 35-38). IEEE.

Qin, J., Chen, L., Liu, Y., Liu, C., Feng, C., & Chen, B. (2019). A Machine Learning Methodology for Diagnosing Chronic Kidney Disease. IEEE Access, 8, 20991-21002.

Shariati, S., & Haghighi, M. M. (2010, October). Comparison of anfis Neural Network with several other ANNs and Support Vector Machine for diagnosing hepatitis and thyroid diseases. In 2010 International Conference on Computer Information Systems and Industrial Management Applications (CISIM) (pp. 596-599). IEEE.

Xu, Q., Tang, W., Teng, F., Peng, W., Zhang, Y., Li, W., ... & Guo, J. (2019). Intelligent Syndrome Differentiation of Traditional Chinese Medicine by ANN: A Case Study of Chronic Obstructive Pulmonary Disease. IEEE Access, 7, 76167-76175.

Isa, I. S., Saad, Z., Omar, S., Osman, M. K., Ahmad, K. A., & Sakim, H. M. (2010, September). Suitable MLP network activation functions for breast cancer and thyroid disease detection. In 2010 Second International Conference on

Computational Intelligence, Modelling and Simulation (pp. 39-44). IEEE.

Olafsson, R., Witte, R. S., Jia, C., Huang, S. W., Kim, K., & O'donnell, M. (2009). Cardiac activation mapping using ultrasound current source density imaging (UCSDI). IEEE transactions on ultrasonics, ferroelectrics, and frequency control, 56(3), 565-574.

Bircanoğlu, C., & Arıca, N. (2018, May). A comparison of activation functions in artificial neural networks. In 2018 26th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE.

Yahaya, S. Z., & Isa, N. M. (2011, March). Implementation of HMLP network with different activation function for cervical cells classification. In 2011 IEEE 7th International Colloquium on Signal Processing and its Applications (pp. 266-271). IEEE.

Yadav, S., & Shukla, S. (2016, February). Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. In 2016 IEEE 6th International conference on advanced computing (IACC) (pp. 78-83). IEEE.

Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006, December). Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In Australasian joint conference on artificial intelligence (pp. 1015-1021). Springer, Berlin, Heidelberg.

He, X., Gallas, B. D., & Frey, E. C. (2009). Three-class ROC analysis—toward a general decision theoretic solution. IEEE transactions on medical imaging, 29(1), 206-215.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC genomics, 21(1), 6.

Polat, K. (2019, April). Freezing of Gait (FoG) Detection Using Logistic Regression in Parkinson's Disease from Acceleration Signals. In 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT) (pp. 1-4). IEEE.

Banko, M., & Brill, E. (2001, July). Scaling to very large corpora for natural language disambiguation. In Proceedings of the 39th annual meeting on association for computational linguistics (pp. 26-33). Association for Computational Linguistics.

Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. IEEE Intelligent Systems, 24(2), 8-12.

Sisodia, D. S., & Verma, A. (2017, November). Prediction performance of individual and ensemble learners for chronic kidney disease. In 2017 International Conference on Inventive Computing and Informatics (ICICI) (pp. 1027-1031). IEEE.

Gunduz, H. (2019). Deep Learning-Based Parkinson’s Disease Classification Using Vocal Feature Sets. IEEE Access, 7, 115540-115551.

Kumar, S., & Katyal, S. (2018, July). Effective analysis and diagnosis of liver disorder by data mining. In 2018 international conference on inventive research in computing applications (ICIRCA) (pp. 1047-1051). IEEE.

***