EasyChair Preprint

№ 7429

Disease Detection Using ML/AI

Rukhsun Ara Parvin

February 7, 2022

Disease Detection using ML/AI:

RUKHSUN ARA PARVIN

parvinrukhsunara@gmail.com

Table of Content-

1. **Abstract**

Now a days' different type of diseases is the major reason for increasing mortality rate among us. Manual diagnosis takes a huge time and effort. To reduce the time and effort, it is now become important to develop automatic diagnosis system, disease detector for early detection of different types of disease. Data mining technique is the one which contribute a lot in development of such system. This paper is relative study on various machine learning models which are Support Vector Machine (SVM), Logistic Regression, Decision Tree, Naïve Bayes, MLP, Random Forest, K Nearest Neighbour (KNN), XG-Boost etc. These are done on the dataset taken from UCI repository, Kaggle and many other Platform. With respect to the result of accuracy, precision, recall, sensitivity, specificity, false positive rate, and True positive rate the efficiency of each algorithm is measured and compared. The aim is to analysis of diseases by machine learning classifier for sufficient decision making in health care. The aim of this paper is explanation of Machine learning, & different type of classifiers & their comparison.

2. **Keyword**- Machine Learning, Artificial Intelligence, Diseases.

3. **Introduction**- There are lot of programming language in computer science. Like Python, R, MATLAB etc. Among them Python language is the strongest language among all, which consisting different type of packages, libraries which are already on it. Most of the models has implemented through python language. This era is a pandemic era. Nowadays COVID is the most infected diseases globally and we are now in the 2nd phase of the pandemic and the world is moving towards the 3rd wave of it. So, it is become important to predict and detect the characteristics and gene segmentation of COVID otherwise it will maintain and increase its continuous death rate just like 2nd and 1st phase of the pandemic. SARS-COV2, MERS COV causes corona virus disease. The risk factor is close contact (within 6 feet), coughs, sneezes, breathes, talks, touch etc. Different type of machine learning or artificial technique helps to find such kind of COVID19 symptoms. Early detection is possible only for machine learning.

   Cardio vascular disease (CVD) is another disease where ML draw a vital sign for detecting this. The reason behind this disease due to high blood pressure, diabetes, extreme level smoking, Over thinking or hypertension, variation of BMI (Body Mass index), Obesity, Inactivity, High level cholesterol, unhealthy diet, excessive alcohol and many other reason. Different kind of heart problems like Arrhythmia, Atherosclerosis, heart defect, CAD (coronary artery disease), Heart infection, Heart attack, Heart failure, Stroke, Very much chest pain, Excessive rate of heart beat, Herat pressure, Frustration, fbs, Heart block, and many other. Machine learning can detect it easily by the help of individual record what they have symptoms, characteristics, attributes like their sex, height, weight, level of cholesterol & glucose, blood test, blood pressure (systolic, diastolic), etc.

   Breast Cancer is the most often identified cancer and major reason for increasing mortality rate
   among women. Manual diagnosis takes huge time and the lesser amount of systems, there is
   important to develop automatic diagnosis system for early detection of breast cancer. Data mining technique which contribute a lot in development of such system. For the classification of Benign and Malignant Tumours, & early detection of breast cancer we have used classification technique of machine learning where machine is learned from some past data which can predict the category of new input. Not only breast cancer but there also blood cancer, skin cancer, lung cancer can also be detected by Machine learning technique.

   For any type of Disease detection and prediction it's important to test blood. Hence Blood Disease detection also important. Different types of Machine learning classifier help to

detect disorder of blood. Except this type of Diseases AI or Machine learning model can detect Diabetes, Plant disease detection, also.
Different type of DDS (Disease detection system) has been generated where information is inserted into android app. In DDS real time database are used by some pre trained machine learning algorithm. Here datasets are deployed in firebase.

4. **Machine Learning & AI**- When one's improve behaviour by some past experience we will say learning has occurred. Machine learning is basically study of computer algorithm that allow computer programs to automatically improve through experience. And here experience is called as data. It's a type of Artificial Intelligence a data analysis technique which act like human. By different type of Data analysis technique meaningful information is extracted from data. And fresh data improve various Machine learning model. Fresh data means outlier, noise free which will be in standard form. Based on fresh data Machine learning generate different type of model which are useful for any type of Diseases detection, forecasting, prediction. It can be act as Decision maker, Detector, any real time problem solver. There is two concepts learner and reasoned where Learner will build the model and reasoner use this model. And finally a particular solution is computed. There exist cloud as well as big data which make Machine learning more smart. We will see how a computer program is differing from the machine learning-

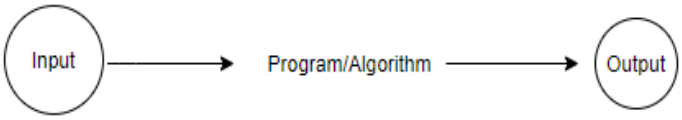

Figure 1- How a program works

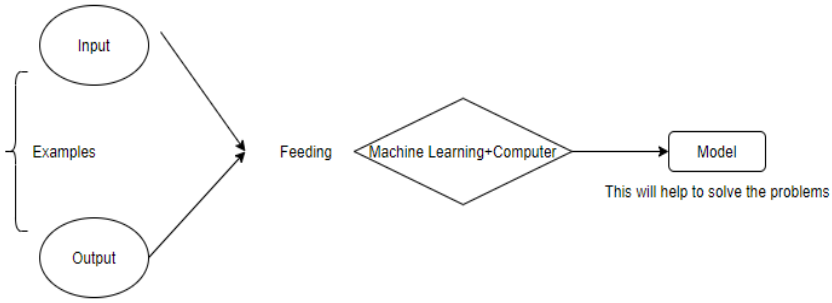

Figure 2- How data feed in to machine learning

Health hazard is the most critical issue in human life. And it is the reason for variable routine in our daily life. Machine learning techniques has taken care of our life from medical report, test by early diagnosis. We are able to known what should we do, which exercise we have needed, what test should require everything can be predicted by Machine learning.

AI is basically when a Machine want to mimic cognitive function of Humans by some programmed rule or protocol. This AI gives instruction to machine that how they behave in certain situation. The difference is Machine will never tired' by doing some work artificially they are more intelligent than human. Machine has more problem solving approach rather than one human. Basically we want some system & software in such a way that they mimic the Human behaviour, AI will have accomplished by studying & learning how a human brain thinks, the way of learning, decide, work, while they solve

some problems in that exact style a machine will do that. Which outcome occurred in this study we used it as basis of developing intelligent system as well as smart software.

4.1 Machine learning types

There are four machine learning technique- Machine Learning technique can be categorized into four main head that is

a. Supervised Learning technique

b. Unsupervised Learning technique

c. Semi-supervised Learning technique

d. Reinforcement Learning technique

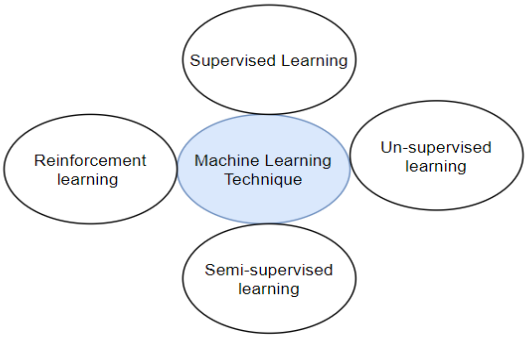

Figure 3- Machine learning technique

Supervised Learning is a subset of Machine Learning that enables us to forecast the output for unknown or future data. It is the job of Machine Learning to discover a function from labelled data. Labelled data is a dataset that contains both independent and dependent variables. I have a dataset that has a large number of active variables.

In supervised learning, we utilise data, which may include both the input and the matching output. That is, we may have an input x and a matching output y for each data instance. And from this machine learning system, a model will be constructed so that given a new observation x, it will attempt to determine what the related y is.

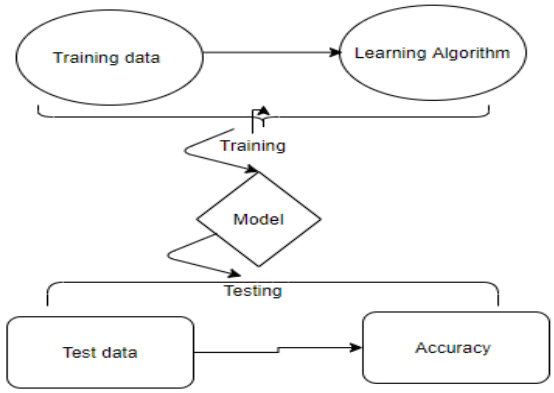

Figure- 4

- We split our data into training and test sets.
- Using the training set, we train our computer.
- Then it's necessary to create a Machine learning model
- Now for checking the performance of that model to get accuracy it's important to test data
- Now we will pass that data and will see the accuracy

Let's assume input features are $\{x_1, x_2, \ldots, x_M\}$, y is a target feature. This values are considered in a training examples for each example. Now also assume a new training example is given where only input features are given. No target values are here. Then We need to predict the value for that target features which is unknown to us. If the value of target value is discrete in range, then it will be called as classification. If it will be continuous range, then it will be called as Regression.

If A is dataset that is labelled with samples x_i and each of this inputs are labelled with y_i such that $A=\{(x_1, y_1),\ldots,(x_n, y_n)\}$, $x_i \in X$, $y_i \in Y$. And the relationship is f:X→Y for each sample $y_i = f(x_i)$ where f= unknown labelling function. For binary classification Y=0,1 and for multi class classification Y consists of possible K levels.

Unsupervised machine learning is the process of training a computer utilising data that will not be categorised or labelled and enabling the algorithm to operate on the data without direction. Without any previous training data, the machine's job is to categorise unsorted data according to similarities, patterns, and differences.

Unlike supervised learning, this method does not include an instructor. That is, the machine will get no training. Thus, the machine is limited in that it will be in restricted mode for the purpose of discovering hidden structures in unlabeled data, which will be accomplished by us.

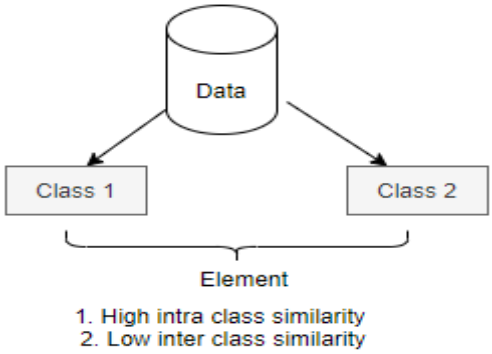

Figure- 5 If data has split in two class what properties they will require

In contrast to supervised learning, we lack any preset characteristic. On the basis of the data that is provided, we will attempt to construct our own class. And will make every effort to ensure that whatever class is produced has a high intra-class similarity and a low inter-class similarity. This method is useful for classifying the input data based on its statistical characteristics.

Reinforcement learning is a subset of Machine Learning in which the learning system observes its surroundings and learns the optimal behaviour by attempting to maximise some concept of cumulative.

Features:1 Agents monitor their surroundings, choose and carry out certain activities, reward, and are rewarded in turn (or penalties in certain cases).

Features:2 Over time, the agent develops a strategy or policy (choice of behaviours) that optimises its benefits.

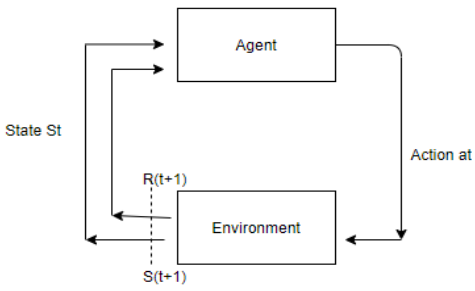

Figure- 6 Reinforcement learning

Semi-supervised learning is a hybrid method that combines supervised and unsupervised learning using labelled and unlabeled data. For data-driven supervised learning. We will generate some function from the supervised data, and if we also have unlabelled data in addition to the labelled data, we may attempt to generate a better function.

Machine learning is concerned with developing a model composed of many disciplines and conditions structured as a probability estimation function. This is a subfield of Artificial Intelligence known as machine learning.

There are many types of machine learning classifiers that are helpful for identifying diseases, diagnosing them, predicting their outcomes, and detecting them early. The usage of Deep Neural Networks is made possible by the presence of GPUs (Graphics processing units) in machine learning.

4.2 Machine learning & A.I History

Some machine learning classifiers are given based on time period.

year	History
1950s	Samuels checker-playing program
1956s	A.I. was first introduced based on theoretical concept
1960s	1. Neural Network i.e. Rosenblatt's perception 2. Pattern Recognition
1970s	Natural language processing(symbolic)
1980s	1. Decision tree & rule learning 2. Learning 3. Planning 4. Problem solving approach
1981s	Back propagation algorithm
1985s	Multilayer Perceptron(MLP)
1990s	**1.** Support vector machine(SVM) **2.** Reinforcement Learning **3.** Ensemble learning technique **4.** Bayes Net Learning
1997s	Ada Boost classifier
2000s	Kernel version SVM
2001s	Random Forest
2005	Neural Network with Deep layer
2006	Deep learning was coined
2009	Self-driving car was built by Google

2010	Deep learning was used commercially as well
2011	Watson wins Jeopardy
2014	Human vision surpassed
2015	Machine translation system by neural network

Table 1- Machine Learning & AI History

4.3 Domain/ Application

- **Diseases detection/ prediction-** There are lot of field and application of Machine Learning in our life. Specifically, the area of Medical perspective. Like as Future forecasting, sentiment analysis, diagnose diseases, Early detection & prediction of various diseases like various types of cancer detection like breast cancer, lung cancer, blood cancer, skin cancer, diabetes, plant diseases detection etc. In this pandemic situation machine learning helps for handling covid19 by early detecting. This basically makes healthcare smarter.

  - ✓ Machine learning also detect chronic diseases, Liver disorder, Heart diseases, Hepatitis, Parkinson's disease. The symptoms are different from each other. Machine learning use this symptom as input and detect that particular diseases earlier. Here different type disease symptoms, Patient records, different type of lab measurement, pathological test, Blood test, DNA tests, previous report is used as inputs in Machine learning for diagnose a disease. And as a result disease has detected. By the help of this records various datasets have created & from that future patients have cured early.

  - ✓ Not only medical history ML is used but also useful for image recognition, object detection, Robot control, for natural language processing, Speech recognition, fraud detection, & many domain & application machine learning has & AI has.

- **In Speech Recognition-** Automated Speech Recognition- When we search something on Google, Chrome, Microsoft edge, Siri we just tell it by our voice they can generate output.
- **Natural language processing**- Language use for human communication like English, Hindi, etc. But in computer system we don't use Natural language. There are lot of application of NLP like automated question answer session, Chat bot, Auto correct, prediction, spam detection, sentiment analysis, machine translation, language synthesis, spam detection etc. AI can help to process Natural languages with the interaction between computers system and human.
- **Image recognition**- By given some provided data AI model can automatically find image pattern and help to process that image by gathering and organizing data. Except this in social media, Healthcare, Financial section, Agriculture, Education, Data security AI plays a significance role.
- It can be used for self-driving car, in CD for recommending some essential product.
- **In social media** like Facebook, Instagram, YouTube, Tube Mate, Pandora they are aware which videos, song we want to next & which we want listen it happens just because of Artificial Intelligence.
- **Human activity recognition**- Machine learning is useful for recognizing human activity. For activity capture smartphone, pulsometer, smartphone, Kinect sensor is used. Various machine learning model like neural network helps to differentiate people's activity, their gesture, movement based acceleration of body, acceleration of gravity, body angular speed etc. Which can help to prevent obesity.

- **Advance Machine learning on google cloud-** Now a day's advance machine learning model used in google cloud for their services. This google cloud machine learning has made a huge revolution in business. Machine learning gives this opportunity to observe data in a new way and for new perspective. That makes us smarter. In google, they use Machine learning for optimizing their products by the help of Maps (real time traffic). They are able to detect spam on Gmail, real time face recognition in photos etc.
- **In Agriculture -** In Agriculture, Machine learning is being used for getting more success & transforming business. Machine learning & AI gives a magical way that farmers can deal with different types of plant, animals and their specific diseases. There is some awesome machine learning model can detect plant diseases earlier. For plant species selection, recognition, Quality of crop, soil & water for plantation ML model can give more accurate result to farmer and makes their decision more effectiveness.
- **In computational genetics-** Machine learning can deal with large dataset. It has some algorithms which can make best prediction on data. In bioinformatics Machine learning, AI creates significance role here. DNA sequencing of different species is also possible for machine learning.
- **In Cyber security-** Machine learning has great impact on cybersecurity. Different types of vulnerabilities, threats, attacks can be detected by Machine learning model. To make confidential information more secure & private; AI, ML can handle this thing & identify them.
- **In financial services-** In financial service ML & AI has key role. For customer service section, Fraud detection in their services are possible just because of ML & AI service Chatbots helps which in financial services for customer request, customer reviews, question answers section, bill payments.
- Not just this area Machine learning & A.I. has wide range application in many other things like E-commerce, Manufacturing, AWS, Self-driving car, political campaigns etc. We don't have any concept how much ML can help us & will help in future too.

5. Role of Artificial Intelligence in Health-

Artificial intelligence has great impact on health. For diagnosis and detecting any disease AI assign machine learning. A.I. authorize doctors and medical practitioners for track out disease and health condition in accurate manner by less amount of cost and errors. Invention of robotics in technology for medical facilities has become grew up. Now a day the greatest invention of A.I. viz. robotics creates a wonderful impact for counselling diseases, for physical therapy in rehabilitation in various laboratory, hospital and research lab in medical field. Also, for operation and surgery purpose robots are used. CT scan is the most important invention in A.I filed that helps to detecting various cancer directly viz. lung cancer. Stroke can be detected by CT scan also. Magnetic resonance imaging (MRI), EEG can access various types of heart disease. Some important boundary on our daily life that A.I. makes it easier given bellow.

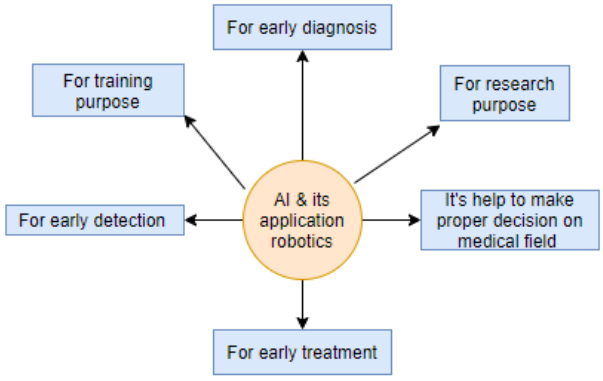

Figure- How robot helps us

5.1 Medical history on A.I.

Year	Medical History
1950s	Various medical thing was invent like X ray, Artificial Kidney, Cardiac Pacemaker and EEG
1970s	An amazing invention replacements of hip and knee artificially
1960s	Analysing of blood, ultrasound, replacement of heart valve
1980s	Surgery of Laser and vascular grafts

6. Related work-

Machine learning has wide range application. Continuous improvement of Machine learning entire world used this very much. It has high computation power also and in large datasets also. It's offer many essential resources as well for various data analysis. Here lot of research areas which are related with this topic Disease detection using ML. Heart disease can detect or diagnosis by using hybrid machine learning model. In (Khourdifi and Bahaj et al.) they predict heart disease by using different type of Machine learning model. They applied lot of optimization that include PSO which is combined with ACO. There are several study on hear disease detection. In a research (Chidambaram T) et al. predict Heart disease using Naïve bayes, AI network, Support vector machine classifier (SVM), Random forest classifier, and simple regression method. They found 98.83% prediction result by Decision tree classifier. This paper was about comparison different type of Machine learning classifier. And Decision tree gave best result. In a study, (Fahad Kamal Alsheref et al.) different type of machine learning classification algorithm like Support vector machine (SVM), K-Nearest Neighbour (KNN), Regression analysis, Decision tree is used for detection of Blood Diseases. (Naresh Kumar et al.) focuses on automated disease diagnosis. They selected three harmful diseases which are coronavirus, heart disease, diabetes. Here a android app has been used where this three disease can be detected by entering data in to that android app by using real time dataset.

Machine learning model is used for corona virus handling purpose. By the help of different type of machine learning technique to determine or predict number of death, number of recovery, affected people, also number of active cases. In china in was predicted that when covid will end by Machine learning technique. A recently published paper (Suparna Biswas et al.) A Hybrid Model based on mBA-ANFIS for COVID-19 confirmed case prediction and Forecast. In another study recently published Mathematical modelling for decision makin of lockdown during covid19 by SIR model (Suparna Biswas et al.). ES, LR, LASSO, SVM (Rustam et al, 2020) has been used for the purpose of estimating number of future affected patients. Data set was collected from GitHub (Wissel et al, 2020).

It's known that cancer is terrific disease in todays' life. And Pakistan is one of the country which has occurrence of Breast cancer. More than 83,000 cases reported. For this this early detection is important and it is the best effective mode. Machine learning helps to early detect and gives us the best outcomes. It requires an effective procedure for discriminate benign tumours from malignant ones. Here lot of related topic exist where researcher has been done lot of research on this. [Senturk & Kara, 2014] he has proposed a model for early diagnosis of breast cancer for patients. He used seven different types of machine learning algorithm for predictions, for prediction process breast cancer dataset has been collected from UCI machine learning repository. And during the process of prediction he used Rapidminer 5.0 the data mining tool to apply data mining technique on chosen algorithm.

For diabetes disease detection (Al-Zebari & Sengur, 2019) used different type of machine learning technique like Decision tree (DT), Logistic Regression (LR), DA, SVM, k-NN and ensemble technique. He used matlab for this purpose. 10-fold cross-validation has been used here. The classification accuracy obtained by individual classifier and compared. Among Decision tree (DT), Logistic Regression (LR), DA, SVM, k-NN and ensemble The best accuracy was given by LR method 77.9%. And worst accuracy was given by Gaussian SVM technique which is 65.5%. Another related work of breast cancer (Senturk et al., 2014) finding the best way for early detecting breast cancer. Dataset was collected from UCI machine learning dataset. He used SVM, NB, KNN, Decision Tree for the purpose of early detecting breast cancer. The higher accuracy was given from KNN classifier 95.15%. And SVM gave 96.40%.

(Cinarer & Emiroglu, 2019) classifying MR brain image characteristics by the outcomes of tumour classification techniques. For this the machine learning classifiers like RF (Random Forest), KNN (K-Nearest Neighbour), LDA, SVM has been used for classifying MR brain image which are n/a, multicentric, multifocal, gliomatosis. Among this classifier SVM gave highest precision rate compare to other.

Thyroid is also another disease in todays' scenario. It has many research areas. (Kousarrizi et al., 2012). By using SVM classifier and UCI machine learning dataset He got 98.62% classification accuracy. He used two datasets one collected from UCI machine learning repository and another from Imam Khomeini hospital which was collected by IDL (Intelligent Device Laboratory) of K.N. Toosi University of Technology.

Parkinson's disease (Huriharan et al., 2014) was diagnosed by neural network and SVM (Support Vector machine). He got 100% classification precision result. (Naql et al, 2020) detect Lung cancer by DL and dataset was taken from LIDC-IDIR (Menge et al, 2018). He provides automated disease detection analyser. He also provides classification for promoting radiologists' diagnosis. By using SVM (Liu et al., 2020) brain stroke has been diagnosed from 1157 patients. 83.3% accuracy was obtained. Liver disease diagnosis approach (Durai et al.,) by UCI machine learning dataset. J48, SVM, NB model was used and obtained 95.04% accuracy. The objective of this research was prediction of higher score rate for liver disease detection. Using different cancer dataset (Zebedee et al, 2018) Cancer disease was detected by Convolution neural network (CNN) base on some gene expression. Classification accuracy was 100%. Another research work is differentiating the characteristics of Alzheimer's disease. (Kulkarni and Bairagi, 2017) proposed SVM model for identifying characteristics for diagnosis of such type of diseases. 96% accuracy was obtained by this model.

7. Different type of machine learning Technique I.e. Different types of Machine Learning Algorithm

There are lot of machine learning algorithm which are developed by Machine learning scientists for detecting any disease, early diagnosis of disease & prediction instead of manually detecting. Lot of machine learning technique such as Random forest classifier, K-Nearest-Neighbour, Support vector machine, Gaussian naïve bayes or Bayes theorem, Decision tree classifier, Linear Regression, Artificial Neural network, Ensemble learning classifier, Logistic Regression, K-Means clustering, C-Means clustering, Principle component analysis, Anomaly detection algorithm, J45, C4.5 and C50 model etc. are used for early detecting disease based on clinical data. Machine learning can use for distinguish the cancerous cell and Non-cancerous cell. It can distinguish pneumonia. ML can detect early growth of Tumour in brain as well malignant and benign growth in Breast, lungs.

Machine Learning Algorithm	
Supervised Machine learning Algorithm	Unsupervised Machine learning Algorithm

Classification	Regression	Clustering
Support Vector Machine	Linear Regression	K-Means, Fuzzy C-Means, K-Medoids
Discriminant analysis	SVR	Hierarchical
Naïve Bayes	Ensemble Methods	Gaussian Mixture
K-Nearest Neighbour	Decision Tree	Neural Network
	Neural Network	Hidden Markov M
		DBSCAN, Agglomerative Clustering

Table 2

7.1 **Support Vector Machines (SVM)-** Vapnik and Alexey introduced Support vector machine 1995s. (Berhard Boser et al., 1992) developed Kernel trick on MMH or Maximum-Mergin-Hyperplanes. (Corinna Cortex et al., 1993) proposed soft margin concept. For classification and regression purpose SVM analyse data. Main purpose of SVM is to generate a decision boundary. This line separates data points into two classes. This decision boundary is called hyperplane. For hyperplane SVM select those points which will be near to that boundary. This points are called Support Vector. For non-linearly separable data Kernel trick is used. In Kernel trick inputs are taken as in low dimension and transform it in higher dimensional space. Which means that non-separable data will be converted to separable data. Polynomial kernel, Gaussian radial basis function, Gaussian kernel, Laplace RBF kernel, Hyperbolic tangent kernel, and ANOVA radial basis kernel are all examples of SVM kernels.

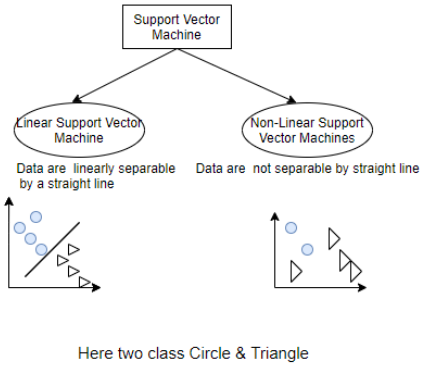

Figure 4

7.2 Logistic Regression- The term "logistic regression" refers to an expansion of the term "linear regression." Probabilistic value is assigned using logistic regression. Logistic function is used in Logistic regression for mapping the predicted values into probabilities. And output is executed through logistic function (Kousarrizi et al., 2012) Logistic regression use the concept of Threshold value. Based on threshold value output can either 0 or 1 (Kousarrizi et al., 2012, Abdulazeez, 2018). Logistic function, logistic(η) = $\frac{1}{1+e^{-\eta}}$. The Odds ratio concept in Logistic Regression making computation easier. Odds ratio = $\frac{p}{1-p}$ where p= probability that an event occurs and (1-p)= probability that the event will not occure.

7.3 Linear Regression- Linear regression classifier in machine learning generates a linear relationship in between dependent & independent variable. Independent

variable can one or more. This model works well in regression but fails in classification. Linear regression can be represented as y=a+bx+c, where y= Target Variable (i.e. Dependent Variable), x= (Predictor Variable) Independent Variable, a= intercept, c=random error. Here x, y are training datasets.

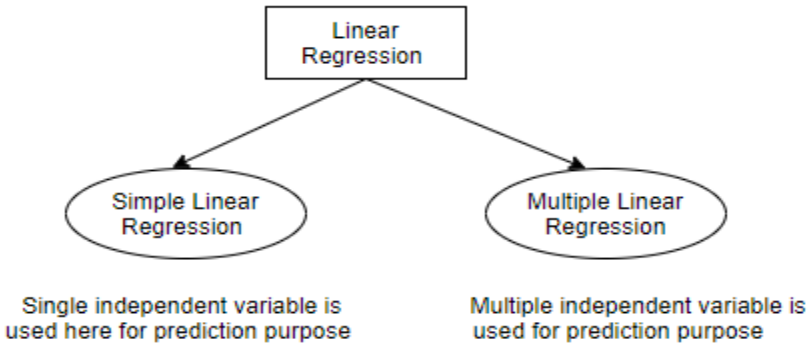

Figure 5

7.4 K-NN- K-NN means k-Nearest-Neighbour classifier. It is non-parametric classifier (1970's). It is non-parametric in approach as there are no assumption or hypothesis. We have to select K number of neighbours. Then need to calculate Euclidean distance. And Based on similarities or nearest neighbours K-NN choose the K nearest neighbours. Need to count all data points in each category in k neighbours. If there will a new point, it's need to assign in that category which has minimum Euclidian distance or has maximum similarities. K-Nearest-Neighbour is used for-

- Pattern Recognition
- Statistical estimation

7.5 DTs- This decision tree classifier is based on a tree structure and has a non-linear function type. It is composed of a root node, an internal node, and a leaf node. This decision node makes decisions about route selection. And the test node defines the example's class.It can be performed on the basis of Greedy. This Greedy approach helps to minimizing depth of tree. It is a non-parametric machine learning classifier (Al-Zebari et al., 2019). That attribute will provide the best prediction result which has highest information gain. And it will select as a root node. And for Gini index that attribute will select for split which will have lowest Gini index and highest reduction impurity. We will prefer smallest DTs always.

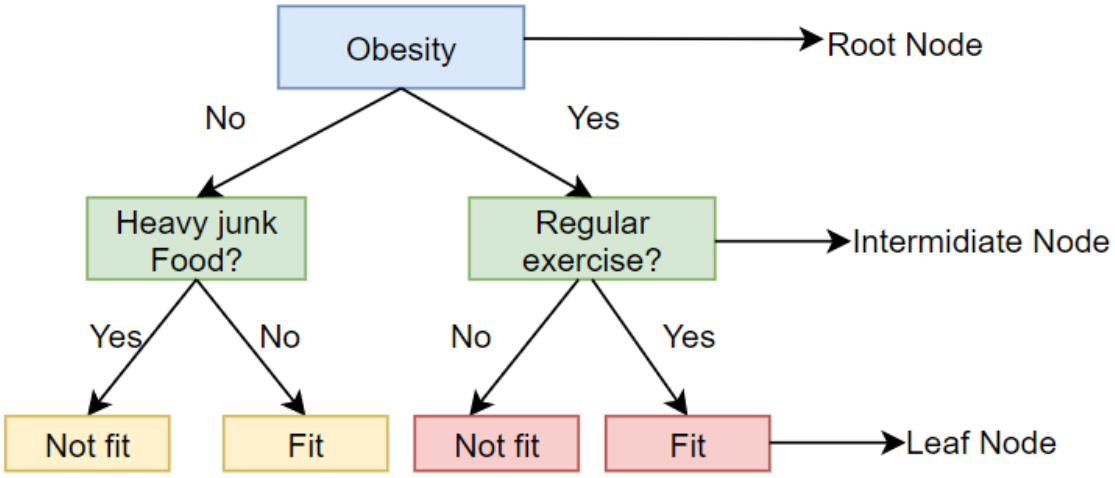

Figure 6- Decision Tree

7.6 SVR- SVR is Support Vector Regression. svr is SVM which can support linear regression as well as Non-Linear regression. Which principle use SVM, SVR use that's one. But the SVR is used for continuous value where SVM for working with classification or discrete value. The main concept is distinguishing the best fitted line.

7.7 Naïve Bayes- The naive Bayes method is based on Bayes' Theorem.Features which are classified by Naïve Bayes classifier, all of this features are independent & equal. It is probabilistic base classifier. Main purpose of this classifier is generating conditional probability.

Bayes theorem says, $P\left(\frac{w}{x}\right)$= $(P\left(\frac{x}{w}\right) P(w))$ | P(x). Here P(x) is same for all classes. $P\left(\frac{x}{w}\right) P(w))$ needs to be maximized. It's called maximization problem. If a priori probability P(W) is not known, then it is assumed that all classes are equally likely.

Join probability can be explained as-

P (W, X1, X2,….,Xn)

=P (X1 | X2,….,,Xn, W). P(X2,…,Xn, W)

= P (X1 | X2,….,,Xn, W). (X2 | X3,….,,Xn, W)P(X3,…,Xn, W)

= P (X1 | X2,….,,Xn, W). (X2 | X3,….,,Xn, W)……P(Xn|W). W

As all features are independent so,

P (X1 | X2,……..,Xn, W) = P(X1 | W)

So, P (W | X1, X2,….,Xn) α P (W , X1, X2,….,Xn)

=P (X1 | W). P (X2 | W) , . . . , P (Xn | W). P (W)

= P (W) $\prod_{i=1}^{n} P(Xi|W)$ - - - - - - - - - - - - Naïve Bayes formula

7.8 Ensemble learning classifier- It is a Machine learning algorithm which will construct a set of Machine learning classifier that will help to classify new data point based on vote of predictions. This Ensemble technique composition of multiple classifier & it will be more reliable compare to a single classifier. Basically Bayesian averaging is the primary ensemble technique.

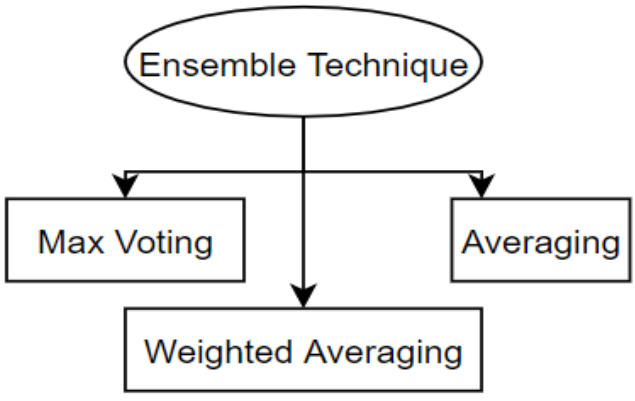

Figure- 7

In case of Max Voting when predictions will perform based on voting. And majority votes are considered for the final result. In case averaging just average predictions are considered as final prediction result. In weighted averaging technique different machine learning model are imposed with various weights demark of each machine learning model for final prediction.

Except this, stacking classifier is also ensemble learning classifier. This staking classifier is also used for prediction (on the test data) purpose where lot of models are used for making a new model. Bagging is bootstrapping aggregating. Bagging is a classifier where need to bootstrapped the data set for making decision. Multiple training datasets sets (Bag) are created here by random sampling with replacement. And each of bags a single machine learning classifier is trained. Boosting classifier is used for correcting errors of previous model in each model. This Bagging and Boosting are most important classifier that are used in machine learning.

Bagging classifier	Boosting Classifier
Bagging meta- estimator	AdaBoost
Random Forest	GBM
	XGBM
	Light GBM
	CatBoost

Table- 3

7.9 K- Means & K- Medoids- Both are Clustering algorithm. Main goal of this clustering algorithm are minimizing the sum squared distance in between data points and cluster centre in which cluster we want to assign that data points. Number of clusters will already be known to us ('K'= Number of cluster). Objective is compute the optimal centroid. In K-Medoids medoid means a particular point of cluster. Dissimilarities are measured in between medoid with other data points of that same cluster. And objective is this dissimilarity should be minimum.

Except this type of cluster there are also DBSCAB (Density- based spatial clustering of application with noise), Agglomerative clustering, Fuzzy C- Means clustering belong in Un-Supervised Machine Learning Algorithm.

7.10 Deep Learning- It is a method for unsupervised machine learning. It may have worked with organised or unstructured data. The concept is same like ML and AI that it's build The learning algorithm by mimic the human brain. It is implemented through the concept of Neural Network. Neural network mimic the concept of Biological Neuron. This Neurons are said as brain cell.

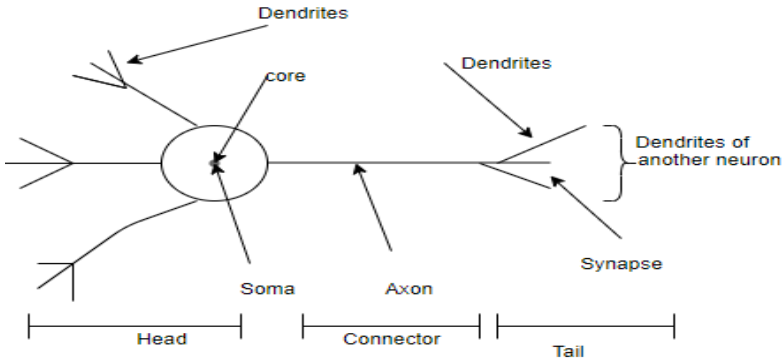

Figure 8- Biological Neural Network

From this diagram dendrite are used for providing input to a neuron. In cell body there are a nucleus which perform some function. Then output will move through axon & it will go on axon terminal and then this neuron will fire that output towards the next neuron. These two neurons will never be connected to each other. There will be huge gap between them. This gap is called Synaptic gap. In this way a BNN works.

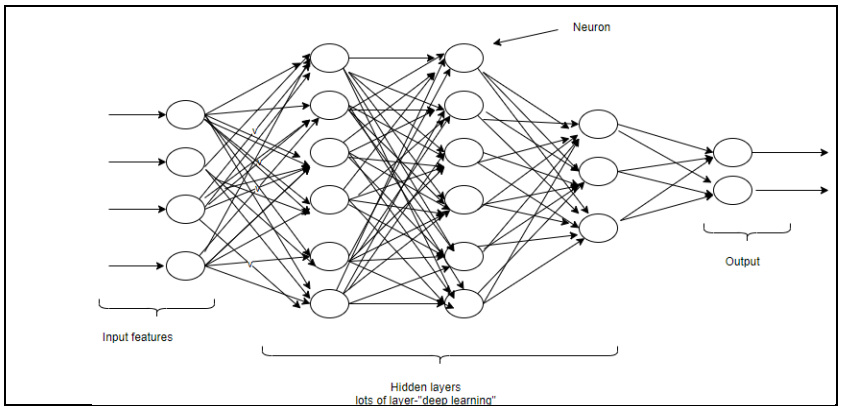

Figure- 9 Deep Neural Network

In ANN there are several inputs ($x_1, x_{2,\ldots,} x_n$) like BNN. There are some randomized taken weights ($w_1, w_2, .., w_n$). Now this inputs will provide to the processing element (like cell bode in BNN). Here weights and input will have summed up by multiplying them. This summation i.e. $s=x_1 * w_1 + x_2 * w_2 + .. + x_n * w_n$. This S is called as a Transfer Function F(S). Then this F(S) will go through some activation function. This activation function is nothing but Threshold value. And the final output of neuron is dependent on it. Then it final result will return value 1 If final result is greater than Threshold value, i.e. neuron will fire at that case otherwise it will generate 0 value i.e. neuron will not fire. There are lot of activation function like sigmoid activation function, step function as well. In this way ANN works which mimic the BNN. In ANN if actual output and desired output are not same that means there are some error. In in Backpropagation algorithm we will try to minimize that error. We will continue this process until we get our desired output as our actual output. By the help of Backpropagation weights will be adjusted. This process called learning.

Deep networks are nothing but neural network with multiple hidden layer. In this learning every node/ neuron will be interconnected to each other. There can be multiple hidden layer. Due to present of lot "Hidden Layer" the concept deep has come. It is not possible in Machine learning.

8. Important Keywords-

- **Accuracy-**
  (True Positive or (TP)) + True Negative or (TN))/ (True Positive or (TP)) + True Negative or (TN)) + (False Positive or (FP)) + (False Negative or (FN))
- **Precision-**
  (True Positive or (TP))/ (True Positive or (TP)) + (False Positive or (FP))

- **Recall**-
  (True Positive or (TP))/ (True Positive or (TP)) + False Negative or (FN))
- **F1-Measure**- (2*Precision*Recall)/ (Precision + Recall)
- **True Positive** (TP)- When we will get targeted result and obtain result both true.
- **True Negative (TN)**- When we will get targeted result and observed result both are negative or no.
- **False Positive (FP)-** When actual value is negative and predicted value is positive.
- **False negative (FN)**- When actual value is positive and predicted value is negative.
- **Sensitivity**- (True Positive(TP))/ (True positive (TP) + False negative (FN))
- **Specificity**- (True negative)/ (False positive + True negative)
- **Confusion matrix**- The performance of a model is computed based on it. Dimension should be m X m where m= # target class.

9.<u>Discussion</u>- In this review paper discussion various Machine learning model used for early diagnosis. Many researchers have been developed such kind of Detector or predictor machine where diseases are detected earlier. Due to having some problem on Machine learning most of researcher used Deep learning algorithm for better achievement of Accuracy, precision, recall value. It's noticed that for corona virus handling (Suparna Biswas et al., Rustam et a., 2020, Wissel et al.,) used ML model. For Breast cancer detection and prediction (Senturk & Kara, 2014, Al-Zebari & Sengur., 2019, Sentark et al., 2014,) used LR, DA, SVM, K-NN, ensemble technique, NB for detection & they successfully got a good classification report. Except this Khourdifi and Bahaj et al., Chidambaram T et al., Fahad Kamal Alsheref et al., Naresh Kumar et al.,) used PSO combined with ACO, Naïve Bayes, AI network, SVM, RFC, Simple Regression method, DTC. And got higher accuracy. Another researcher used DTC for blood disease detection. Another researcher (Huriharan et al., 2014) He got successfully 100% classification precision result by using Neural network and support vector machine. Except this various researcher who worked in various Machine learning model for detection diseases that are included in Related work. Except this various researcher who research in diseases detection that will be discussed below.

- **Heart Disease**- (Chithambaram T, Logesh Kannan N, Gowsalya M) they used machine learning model K- Nearest Neighbour, Random Classifier, Correlation and Support vector machines for getting better accuracy for detecting Heart Diseases. They used datasets with 70K data, 11 features which was collected from kaggle.com. Data has been pre-processed by Standardscaler. Binary values used for scaling data. By using K-NN they gained 63.4% score. Not too much better. By using RFC classifier, they gain 71.4% accuracy which is better than K-NN. Also for processing data RFC take little time compare to K-NN. 68.4% accuracy given from DTC classifier. Basically Gini-index was used for prediction. Finally, 72.5% accuracy was given from SVM (Linear kernel), 86.2% by SVM kernel (Gaussian).

On other side, (O.E. Taylor) gained 98.83% prediction accuracy by using DTC classifier. The entire paper about comparing different type of ML model like KNN, Random Classifier, DTC. They have been successfully predicted score for future reference. That Heart disease may occur & based on that ML model which gave best accuracy compare to other that particular ML model can be made for the objective yield checks that an individual having Heart illness or not.

- **Breast Cancer**- (Habib Dhahri, Eslam AI Maghayreh, Awais Mahmood, Wail Elkinai, Mohammed Faisal Nagi 2019). For breast cancer diagnosis they chose dataset from UCI Machine learning repository. Lot of researcher has used this same dataset for detecting breast cancer. Which consist of total 569 breast cancer patients (M= 212, B= 357). By Standard scaler classifier data has been pre-processed. For features selection PCA (Principle component analysis has been used) for relevant features by randomized SVD. Ensembles model was used for this diagnosis with 10- fold cross validation. And they successfully detect breast cancer.

RESEARCHERS'	Chosen model for Breast cancer diseases detection	Result		
Siham A. Mohammed, Sadeq Darrab, Salah A. Noaman, Gunter Saake 2020 and their paper title is Analysis of Breast Cancer Detection using Different Machine learning technique.	Decision Tree (J48), Naïve Bayes, SMO (Sequential Minimal Optimization, 10-Fold cross validation on WBC and Breast Cancer dataset.	J48	NB	SMO
		75.52%	71.67%	69.58%

Pre-processing technique is basically if missing values present in dataset then based on pre-processing technique they are removed. And after pre-processing technique maximum time it takes good result compare to accuracy before pre-processing. This is chosen just because of getting good prediction result. There are lot of pre-processing technique like MinMaxScaler, Standard scaler etc. They got result different accuracy as they used discretization filter 7 times. For 7-time discretization filter they obtained at last 98.20% for J48, 76.61% for NB and 95.32% for SMO. Accuracy significantly increased for j48 but not for NB and SMO.

And then After pre-processing technique they got result for j48= 74.82%, NB= 75.53% and SMO= 72.66%. They got 98.20%, 76.61%, 95.32% for j48, NB, SMO after 7 times resampling which is significantly good. These all about breast cancer dataset. They used same technique on WBC dataset also. And they got 99.24%, 99.12% and 99.56% for J48, NB, SMO classifier by using resample filter technique. Using pre-processing technique, they got 95.91%, 97.37%, 96.78% for J48, NB, SMO. So, from this paper it is obvious that for obtaining good prediction result resample filter, pre-processing technique and discretization played a vital role for disease detection.

- **Blood Diseases Detection**- (Fahad Kamal Alsheref, Wael Hassan Gomma 2019). They used benchmark dataset with 668 records with 28 parameters for detecting blood disease based on blood test result. They got highest accuracy result 98.16% by using LogitBoost classifier. Support vector gave lowest accuracy. In between high and low Random forest gave 97.12%, DT gave 97%, Regression gave 96.54%, KNN gave 92.97%, Bayesian network gave 92.86%. MLP contributed 91.80%. They compute metrics such as TP Rate, FP Rate, Precision, Recall, F-measure, MCC, and ROC curve to determine the correctness of the model. Additionally, they calculate the PRC Area and the mean absolute error.

- **COVID19, Heart disease, diabetes-** Fear of COVID19 increase rapidly day by day. In this situation early detection can control death rate. (Naresh Kumar et al., 2021) they sketched automated diseases (COVI19, Heart disease, diabetes) diagnosis machine learning model using Logistic Regression. In their model a real time dataset has been used where data are feed to mobile app & based on data detection is possible. For COVID19 prediction relevant features were chosen travel history location, country, age, sex, symptoms for prediction purpose. It is consisting of 13174 data points.

  For heart disease detection they used 70K data points. Among them relevant features like gender, age, height, weight, glucose, smoke, alcohol, cholesterol, systolic blood pressure, diastolic blood pressure has been used for heart disease detection. From this relevant features systolic blood pressure, diastolic blood pressure, age, cholesterol is major reason for heart disease according to their heat map.

  For Diabetes detection total 768 data points has been used. Among them relevant features are pregnancies, Basel metabolic rate, age, blood pressure for prediction purpose. And this features are reason for diabetes according to their heat map. Logistic regression has been used & stored on firebase for all dataset.

  By their proposed model they successfully detect disease by asking some question just a few second. In their work, they split their dataset into 3 portion training (75%), testing (25%) & validation (10%). Accuracy & F- measures for this proposed model has been predicted which is given bellow-

Diseases	Accuracy	F- Measures
Covid19	1.4765%	1.2782
Diabetes	1.8274%	1.7264
Heart diseases	1.7362%	1.3821

  So, the prediction result is comparatively good. In future we will try feed more data on android app & will try deployed firebase dataset like this.

- **Parkinson's disease**

  Another disease called Parkinson's disease for which no particular test is generated. When nerve cells of brain die Parkinson's disease occurs. Lot of

symptoms are here like Tremor, changes of speech, writing, Loss of movements or slower movement, Constipation, Weak muscular etc. Early detection important. Various researcher works on it by Machine learning model, Artificial model. Machine learning is the only way which can detect this disease within few second. (V. Ulagamuthalvi et al., 2020) used machine learning model for identification of Parkinson's disease using XGBoost classifier & Logistic Regression for purpose of classification. They used UCI dataset which consisting of lot of audio signals. For pre-processing data Min-max scaler has been used. By XGBoost classifier 96% accuracy has been occurred where Logistic Regression gave only 79% accuracy. So, XGBoost classifier works well in Parkinson's disease identification. Except this more researcher works on Parkinson's disease which are given bellow-

RESEARCHERS'	Chosen model for Parkinson's disease detection	Result
Mohammad et al., 2014	SVM, Random Tree, BPNN (Back propagation neural networks) or FBANN (feed forward back propagation neural network), 10-fold cross validation on 195 dataset (voice samples).	97.37% by FBANN Among 195 patients total number of Parkinson's patients= 23.
Ramani et al., 2011	Logistic Regression, LDA, Random tree, SVM on UCI dataset	More than 90% through LDA & Random tree.

Except this disease more no of diseases are detected by ML/AI model which are given-

Diseases	Researcher's	Technique & Result
Chronic Kidney Disease prediction	S. Revathy et al., 2019	They used SVM, RFC, DTs for prediction. After training and pre-processing data DT gave 94.16% accuracy. Where among 120 # instances 113 instances were correctly classified. SVM gave good result better than DT was

		98.33%. And among 120 instances 118 instances were correctly classified. RFC gave best result compare to them which is 99.16%. And 119 instances were correctly classified. So, Random forest is the best model for Chronic disease prediction. They predict the result based on confusion matrix.
Diabetes prediction	Hasan Tahir Abbas et al., 2019	The main purpose of paper is defining demography for preventing from this disease. They used SVM and 10-Fold cross validation for gaining better result. And they got 84.1% validation accuracy where 81.2% is recall no to much better but significantly good.
Fatty liver disease prediction using Machine learning technique.	Cieh-Chen Wu et al., 20018.	Total number patients are 577. Random Forest, Naïve bayes, ANN, LR has been used for prediction. And RFC gave the best accuracy result with 87.48%. This prediction system can take step for preventing this disease. Early diagnosis possible also. Among 577 patients 377 patients were correctly detected. They used 10-Fold cross validation also for achieving the better result. So, RFC is good classifier for predicting liver disease.

Thyroid disease prediction	Shaik Razia et al., 2017	Actually they gave a review on Thyroid disease. Datasets size 72k where features were 22. There ere no missing value on dataset so they got result. They obtain 99.23% accuracy for Decision tree. Multiple linear regression gave 91.59%, SVM gave 96.04% and NB gave 6.31% the worst result given from NB. S, for Thyroid prediction we can say DTs is a good choice.

10. <u>Conclusion-</u> Machine learning has capability of handling large amount of data. And accuracy, precision and recall depend on the quality of dataset. There are lot of machine learning techniques helpful for automatic diseases detection. And all algorithm has different procedure for detection. Different type diseases experiment by Machine learning model has been conducted for different disease like covid19, heart disease, different type of cancer etc. So, lot of disease has been discussed here for the purpose of decreasing high risk from disease. I have discussed about Cancers, Chronic Kidney disease, Hepatitis, Thyroid, Liver, Blood Disease, COVID19, Diabetes, Heart disease etc. I have explored and give a review from various researchers what they did on DDS (Disease detection system).

Different researcher got different and excellent result. It's necessary all diseases should detect within limited time. But I have observed in various ML model gave worse result just because of high dimensional data in which case early detection is not possible. I have observed (Chithambaram T et al., 2019) they got 63.4% score for heart diseases detection which is not too good by using K-NN classifier (not too much relevant features). The dataset was huge. And also KNN took huge amount of time for processing it. In this review paper I understood every ML classifier that researcher have chosen worked well & gave good result. It's shown Decision tree gave best prediction result 98.83% for Heart disease detection which was good accuracy result compare to other model. For diabetes disease detection 77.9% was the prediction accuracy which is moderate not too good. So, it's need to do more work on it. We will try to build ensemble model, Deep learning model on diabetes dataset that will give best classification result. SVM worked well in Breast cancer UCI machine learning dataset gave accuracy result 96.40%. SVM gave 98.62% result on Thyroid disease detection which is awesome & amazing. Another disease that for Parkinson's diagnosed 100% classification precision result was given by SVM. SVM gave 83.3% accuracy for brain stroke diagnosed. J48 models gave 95.04% accuracy liver disease diagnosis. For Alzheimer's disease researcher got 100% classification accuracy, another researcher got 96% accuracy by using SVM. So, in nutshell we can conclude

that SVM is a good classification algorithm & smart compare to other which showed high accuracy. And also it's noticed that SVM is used by every researcher frequently. Also I have noticed 10-fold cross validation has been used for every classifier. But for Parkinson's disease detection FBANN gave prediction result good which is significantly good compare to SVM which provided good specificity only that was 100%. The ML algorithm which has some good side and bad side also.

Classifiers	Pros.	Cons.
Support Vector machine	1. It can be useful for linear separable dataset as well non-linear separable dataset as it consists of Kernel function. 2. Chances of overfitting less. 3. I have seen maximum researcher used SVM classifier for disease detection and prediction and got max time best result. Fewer amount of time SVM gave worst result.	1. As it is suitable for linear separable dataset when datasets are in non-separable kernel trick need to apply. Kernel convert input datasets (lower dimensional) into higher dimensional output sets. And hence computational time will be taken as large.
Naïve Bayes	1. Those features are not suitable for ML model can be removed by NB classifier. 2. It takes less amount of time.	1. For training purpose, it requires large datasets. So, if in case of small size dataset can run on this model result can come as a bad.
NN	Error can be adjusted by learning.	Datasets should be large 1. Lot effort
DTs	It requires less effort If missing values are present in dataset it will not effect on DTs during training.	It requires lot time for training.
RFC	Random forest consisting of decision tree. It can act Ensemble classifier which can give good result. Missing values are automatically handled by it.	It takes huge time for training.

For high dimensional data ML unable to deal with it. And classification report not too much good if Ml is applied in Big data. As for high dimensional data we need huge input data as well as output data. Features are unique character that sample

has. It can be considered as variables. In AI these variables are nothing but features. A big problem of Machine learning is features extraction. Machine learning models are not efficient for automatically generate features. But the Deep learning will automatic generate features. Deep learning model are capable of for focus on the right features. As Deep learning has huge computation power that's why it can solve the dimensionality problem that it can deal with high D data. So, hence in future It should to try detecting diseases by DEEP learning as well.

12. References

[1] (Nareen O.M.Salim & Adnan Mohsin Abdulazeez, 2021). Human Diseases Detection Bases On Machine Learning Algorithms: A Review.

[2] (Meherwar Fatima & Maruf Pasha, 2017). Survey of Machine Learning for Disease Diagnostic.

[3] (Shaik Razia, Swathi Prathyusha, Vamsi Krishna, & Sathya Sumana, 2018). A review on disease diagnosis using machine learning techniques.

[4] (Md. Mohaimenul Islam, et al., 2018). Prediction of Fatty Liver Disease using Machine Learning Algorithms.

[5] (K. Subhadra & Vikas Boddu, 2019). Neural Network Based Intelligent System for Predicting Heart Disease.

[6] (Hasan Tahir Abbas, Lejla Alic, Marelyn Rios, & Muhammad A Abdul-Ghani, 2019). Predicting Diabetes in Healthy Population through Machine Learning.

[7] (S.Revathy, B.Bharathi, P.Jeyanthi, & M.Ramesh, 2019). Chronic Kidney Disease Prediction using Machine Learning Models.

[8] (Siham A. Mohammed, Sadeq Darrab, Salah A. Noaman, & Gunter Saake, 2020). Analysis of Breast Cancer Detection Using Different Machine Learning Techniques.

[9] (V. Ulagamuthalvi, G. Kulanthaivel, G. Sri Nikhil Reddy, & G. Venugopal, 2020). Identification of Parkinson's Disease Using Machine Learning Algorithms.

[10] (Naresh Kumar, Nripendra Narayan Das, Deepali Gupta, Kamali Gupta, & Jatin Bindra, 2021). Efficient Automated Disease Diagnosis Using Machine Learning Models.

[11] (Fahad Kamal Alsheref & Wael Hassan Gomaa, 2019). Blood Diseases Detecting using Classical Machine Learning Algorithms.

[12] (Habib Dhahri, Eslam Al Maghayreh, Awais Mahmood, Wail Elkilani, & Mohammed Faisal Nagi, 2019). Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms.

[13] (Chithambaram T, Logesh Kannan N, & Gowsalya M). Heart Disease Detection Using Machine Learning.