HEALTHCARE DATA ANALYSIS USING DATA MINING TECHNIQUES FOR DISEASE PREDICTION

Navita[1]
[1]M.D. University, Ph. D Scholar, Department of Computer Science and Applications, Rohtak, Haryana, India
navitamehra55@gmail.com

Dr. Pooja Mittal[2]
[2]M.D. University,Assistant Professor, Department of Computer Science and Applications, Rohtak, Haryana, India
mpoojamdu@gmail.com

Abstract

Nowadays, the continued expansion of modern technology provides a way of emerging IoT and Artificial Intelligence together. Data mining techniques have been applied to a large number of datasets generated through various sources in order to bring out useful information and for making the prediction.The performance of any data mining techniquemay vary depending upon the type of dataset being used and the application area under consideration. Hence, finding the data mining technique which is utmostappropriate for a precise application domain and its associated datasets would be very advantageous. Major reason behind this study is that in most of the developing countries the health data is normal dataset but with the use of increasing sensor-based technologies it becomes necessary to make analysis on sensors dataset also.To fulfill such type of requirement, analyzing the performance ofwell-known data mining techniques, including Decision Tree, K-Nearest Neighbors, Naïve Bayes, Support Vector Machine, Random Forest Tree, and Logistic Regressionfor diseases prediction has been conducted. The analysis utilized four different datasets including Heart Disease, Breast Cancer, MIT_BIH Arrhythmia, and Activity Recognition datasets collected through smart sensors or clinical examination of patients.These data sets were utilized in order to evaluate and analyze the considered techniques to select the most suitable one with high prediction accuracy. Analysis results show that Random Forest Tree provides the best outcome for all the considered datasets. Hence, performance analysis among all data mining techniques is carried out extensively where accuracy, precision, recall, and f1_score is taken into action. Spyder IDE is utilized for evaluation purposes.

***Keywords:*Internet of Things (IoT); Data Mining Techniques (DMT); Disease diagnosis; Artificial Intelligence; Naïve Bayes (NB); Support Vector Machine (SVM); Random Forest Tree (RFT); Decision Tree (DT);K-Nearest Neighbors (KNN).**

1. INTRODUCTION

In the present day, the whole world faces lotsof challenges in the healthcare domain like lack of clinical specialists,a growing number of chronic diseases, increasing healthcare costs,aged population, and continually increasing vast amount of data [Arkadip (2020)]. According to the projection made by World Health Organization, the global number of estimated aged people will be 1.5 billion in 2050 [Suzman (2020)] and also states that chronic diseases like heart attack, breast cancer, cardiovascular disease, diabetes, alzheimer, etc. are the leading cause of death worldwide[Elfein (2020)].Much work has been done on the healthcare system to deal with such types of challenges e.g., integration of ML and AI in healthcare [Rath (2019)], utilization of advanced technology based resources like wearable medical devices, smart sensors, smartwatches, smartphones, etc.[Parthasarathi (2021)]. Integration of ML and AI will improve the computational efficiency in dealing with clinical data and the advanced technology based resources provide a smart way to monitor and collect the data regarding patient health. In order to make correct predictions about the particular disease, data generated from various sources must be integrated with ML/AI techniques [Rath (2019)].Data mining techniques which is the most popular term comes under ML is used by many researchers in the healthcare domain.

Many researchers have made a comparison of different data mining techniques including Naive Bayes, Random Forest Tree, KNN, Multilayer Perceptron, Artificial Neural network, Decision Tree, Logistic Regression, etc. for the

prediction of numerous diseases like breast cancer, heart disease, diabetes,covid-19, etc. on the basis of clinically collected dataset [Mandal (2020)] [Gupta (2011)].But with the advancement of IoT-based smart technology and its integration in the medical domain demand smart ML techniques that can make efficient predictions on data generated from different sensor and sensor based devices The sensor based devices in healthcare domain provide a new way of treating a patient from a remote location. These sensors are either worn or implanted on the patient's body to offer real-time monitoring of patients like cardiac monitoring, respiratory rate monitoring, blood glucose monitoring, childbirth monitoring, and neurological disorder monitoring. All such types of monitoring services are supported by distincts types of medical sensors like heart rate sensors, pulse oximeter, temperature sensors, IMU sensors, and motion sensors, etc [10]. A huge quantity of data generated through real time monitoring must be analyzed to provide efficient and timely treatment to patients using different data mining techniques. Medical sensors used for capturing the health data are represented by Fig.1.

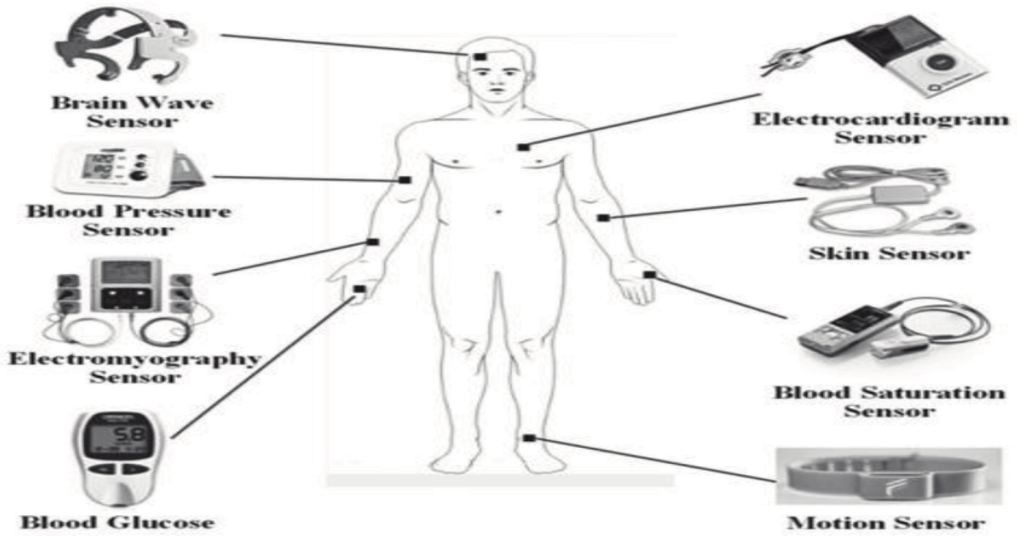

Fig. 1. Medical Sensors used for Remote Monitoring of Patients [9]

Due to the availability of a large number of data mining techniques, one common issue is to identify the best one for the identification of a particular disease.So, there may occur a need to analyze the performance of different data mining techniques on clinical as well as sensor data set to know their performance.
In this paper, we analyze and evaluate the performance of different data mining techniques on different types of health datasets like MiT_BiH Arrihymia, Heart Disease, Activity_Recognition, and Breast Cancer for the prediction of a particular disease. For comparison, we describe different data mining techniques named DT, RFT, Naive Bayes, KNN, SVM, and Logistic Regression. Although in literature there are many data mining techniques available, we consider only six well known data mining techniques that are highely utilized in the community for the prediction of different diseases. The comparison was performed by using commonly used evaluation metrics Accuracy, Precision, Recall, and F1_Score.

The major contribution of this paper is summarized below;

- Finding the best data mining techniques among the most popular and commonly used techniques
- Analysis and evaluation of various data mining techniques (KNN, Naive Bayes, Decision Tree, Random Forest Tree, Logistic Regression, SVM)
- Application of data mining techniques on real time data as well as clinically collected data for making accurate predictions of disease.
- Suggesting the most suitable technique to predict the diseases with high accuracy

Organiztion of this paper is as follows: Section 2 presents a brief survey on work done by different researchers. An overview of six data mining techniques used for analysisis given in Section 3. Section 4 describes the datasets, tool and performance evaluation metrices. Experimental result andperformancesanalysis is illustrated in Section 6. Conclusion, as well as future research directions is given in sction 7.

2. Related study

Many researchers have been trying to predict various chronic diseases by employing different DMT.[Mandal (2017)] usedDT,NB, and Logistic Regression for the detection of Breast Cancer cells where Logistic Regression achieved an accuracy of97.90% with a 10 fold cross-validation. Further, [Guptaet al.(2011)] made a performance analysis of different data mining techniques on three different data mining tools (WEKA, Tangara, and Clementine)using four different datasets. Among all applied data mining techniques SVM exhibits the highest accuracy of 96.74% for Pima Indian Diabetes, C 4.5 achieved the highest accuracy of 79.71 % for the Liver dataset, and for the Statlog heart disease dataset. SVM accomplished the largest accuracy of 99.25% with 10 fold cross-validation.However, [Senturket al. (2014)] appliedto utilize different data mining techniques (Discriminant Analysis, Artificial Neural Network, KNN, Logistic Regression, Decision tree, and Support Vector Machine) on breast cancer data set for the diagnosis of breast cancer. Among them, the SVM achieved the highest accuracy of 96.5% . [Vijayarani et al.(2015)]usedSVM and NBdata mining techniques for the prediction of Liver disease and found thata classification accuracy of 79.66% achieved by SVM which is higher than Naive Bayes.[Venkataet al. (2014)]implemented a Heart disease diagnosis model using two important techniques named Decision Tree and Naive Bayes and obtained an accuracy of 85.03% from Naive Bayes. However, [L.Sayed etal.(2018)]offereda model named 'Telemammography' for early diagnosis of breast cancer and used J48, RFT, KNN, Multilayer Perceptron Neural N/Was a classification techniques. Among all, the Decision Tree achieved the highest accuracy of 96.93%. Further, [Verma et al. (2018)]proposed healthcare monitoring framework for student by using the concept of cloud and IoT. SVM, KNN, DT, and neural network data mining techniques were used for the analysis of data regarding student health and found that among all Decision tree achieved the highest accuracy of 92.59% . On another end, [Ganesan and Kumar (2019)]designedanIoT-based diseases prediction and diagnosis model and implemented using different classification techniques such as J48, Multilayer Perceptron,support vector machine, and Logistic regressionclassifier to evaluate the performance of a model. Further,[Verma and Sood(2018)] proposed a disease diagnosis healthcare framework using different IoT technology such as temperature sensors, blood pressure sensors, etc. The proposed model was evaluated using the UCI as well as sensors dataset with the help of different classification algorithms such as Naïve Bayes, Decision Tree, Neural Network, and KNN model achieve the highest accuracy of 98.26% for the Decision Tree among all. Further,[Aich et al. (2018)]suggested a noval classification approach on the basis of decision tree for Parkinson's disease prediction. Diverse types of decision trees are used as a classification technique and found out that random forest decision tree having the highest accuracy of 98.2% among all to classify the patients suffering from Parkinson's disease. However, [Chiuchisan and Geman (2014)] proposed decision making and a home monitoring system for patients suffering from a neurological disorder. Different IoT-based technologies are used for the establishment of such a system. [Khan(2020)] designed a heart disease prediction IoT framework by using MDCNN (Modified Deep Convolution Neural Network) classifier. MDCNN techniques classified the patients suffering from heart disease into a normal and abnormal class. Results obtained from their experimentation stated that MDCNN having higher accuracy as compared to other classification algorithms. Further, [Memon et al.(2019)] proposed the diagnostic system by using machine learning techniques based diagnostic system. Analysis was done on the Wisconsin Diagnostic Breast cancer dataset by utilizing all categories of SVM. Among all categories of SVM,SVM- linear achieved the highest accuracy of 99%.

3. Overview of data Mining techniques

An overview of different data mining techniques is provided in this section to evaluate and analyze on different datasets. The performance of each technique was observed and tabulated in respect of the accuracy, recall, precision, and f1_score. A brief description of each technique is presented below:

3.1 K-Nearest Neighbors (KNN)

It is one of the easiest data mining technique used for classification purposes. The samples are classified based on the distance between them. Inany dataset, there exist two considerations in the formation of features (x) and labels (y). In the training dataset, vector x_i contains a set of features and $y_i$is the class label corresponding to x_i. For making the prediction of any observation firstof all distance between that observation and all features must be calculated. There existseveral methods for distance calculation, including Euclidean, Minkowski, and Manhattan [Vitabile (2019)][Raihan et al. (2021)]. KNN requires less computational time but does not work well with high dimensional data.

Among all, Euclidean distance is the most popular, and calculate the distance using equation 1:

$$d(x_i y_i) = \sqrt{(x_{i,1} - y_{i,1})^2 + \ldots\ldots + (x_{i,m} - y_{i,m})^2} \quad (1)$$

3.2 Support Vector Machine (SVM)

It classifies the data by findingthe best hyperplane or a decision boundary in an n-dimensional space. The hyperplane can be evaluated using equation 2 [Vitabile (2019)].

$$f(x) = a^T x + c \quad (2)$$

Here, a represents the dimensional coefficient and c is the offset.

SVM having less overfitting problems and can easily work with complex structured dataset.

3.3 Naive Bayes Classifier (NB)

It is the most effective classification algorithm used for quick and fast prediction. It expects that occurrence of certain featuresdoes not depend on other features. Naïve Bayes comes under a supervised learning approach that depends on the Bayes Theorem represented by equation 3. It also acts as a probabilistic algorithm in which predictions are made by considering the probability of an object as the base point [Vijayarani (2015)] [Vitabile (2019)]. Bayes Theorem:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \quad (3)$$

Where,

P (A|B) Posterior probability	Hypothesis A probability on the noticed event B.
P (B|A) Likelihood probability	likelihood of the evidence specified that the likelihood of hypothesis is true.
P (A) Prior Probability	likelihood of hypothesis prior to notice theevidence.
P (B)	Probability of Evidence denotes Marginal Probability.

3.4 Decision tree (DT)

A Decision tree is the simplest and the most usually used classification algorithm. It is a tree-like structure where attributes are represented by an internal node, decision rules are represented by branch and the final outcome is represented by the leaf node. Binary, as well as multi-class classification problems, can be easily solved with the help of this algorithm [Aich et al. (2018)] [Fatima et al. (2017)].Entropy is used as a splitting measure for splitting the nodes in the tree. Equation of Entropy is represented by equation (4). The problem associated with the decision tree is that it can not handle linear relationships among data [Raihan (2021)].

$$Entropy(s) = \sum_{i=n}^{n} - p_i * \log 2(p_i) \quad (4)$$

P_i is the probability of S belongs to class i.
n number of classes

3.5 Logistic Regression (LR)

Logistic regression is mainly used for the evaluation of probability where an instance highly fits a particluar class [Raihan (2021)].Logistics regression is highly useful when a dependent variable is categorical in nature. Logit function used for evaluation of probability is given by equation 5:

$$\hat{p} = h_{\phi} x = \sigma(x^{t}\theta) \qquad (5) \quad \text{Where,} \quad \sigma(t) = \frac{1}{1+\ell^{-t}}$$

3.6 Random Forest Tree (RFT)

RFT comes under the category of supervised learning algorithm which uses the concept of ensemble learning in which multiple classifiers are combined to resolve the complicated problems and expand the performance of a model. In the Random forest tree algorithm, multiple decision trees are integrated to reach the final decision.The decision tree faces a problem of overfitting which is resolved by averaging decision trees [Raihan (2021)]. It recieved the correct prediction from the individual tree and attempts to select the best result with the help of the voting procedure [Vanjana (2015)][Fatima (2017)].

All the above stated data mining techniques are further used for the experimentation purpose of this work and applied on different datasets for knowing their perfromances using different evaluation measures.

4. Data Sets, Tool, and Performance Evaluation n metrics

Details of all datasets, implementation tool, and performance evaluation metrics are described in this section.

4.1 Data Set Description

In this paper, analysis was performed on four different data set (MIT_BIH Arrhythmia, Activity Recognition, Breast Cancer, and Heart disease dataset) taken from different dataset repositories. In which two datasets were collected from IoT based sensors devices either worn or implanted on patient body such as ECG (electrocardiogram) and accelerometer and two were collected clinically by observing patient in hospitals.

1. Heart Disease Dataset:

This dataset is available on the UCI repository contains a total of 1190 records with 14 features divided into two classes [38].

Attribute No.	Attribute Name	Description
1	If	patient identification number
2	Ccf	Social Security Number
3	Age	Age in Years
4	Sex	Male=1; Female=0
5	ChestPainType	Value=1:typical angina, Value=2 Atypical angina:,Value 3: non-anginal pain,Value=4: asymptomatic
6	TrestbPs	resting blood pressure
7	Cholestrol	serum cholestoral in mg/dl
8	fbs	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
9	exang	exercise induced angina (1 = yes; 0 = no)
10	restecg: resting electrocardiographic results	Value 0: normal, Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
11	mhr	Maximum Heart Rate
12	oldpeak	ST depression induced by exercise relative to rest
13	STSlope	slope of the peak exercise ST segment, Value 1: upsloping, Value 2: flat, Value 3: downsloping
14	target	1: Yes, No:0

Table 1: UCI Heart Disease Dataset Description [38]

2. Breast Cancer Dataset

This data set also available on the UCI repository contains a total of 569 records with 32 features classified into two categories [37].

Attribute No.	Attribute Name	Description
1	ID number	Identification Number
2	radius_mean	Mean of radius(mean of distances from center to points on the perimeter)
3	texture_mean	Mean of Texture(Standard deviation of gray-scale values)
4	Perimeter_mean	Mean of perimeter
5	area_mean	Mean of area
6	smoothness_mean	Mean of smoothness (local variation in radius lengths)
7	compactness_mean	Mean of compactness (perimeter^2 / area - 1.0)
8	concavity_mean	Mean of cancavity(severity of concave portions of the contour)
9	concave points_mean	Mean of Concave Point (number of concave portions of the contour)
10	symmetry_mean	Mean of symmetry
11	fractal dimension_mean	Mean of fractal dimension ("coastline approximation" - 1)
12	Radius_se	Standard error of radius (mean of distances from center to points on the perimeter)
13	texture_se	Standard error of texture (Standard deviation of gray-scale values)
14	perimeter_se	Standard error of perimeter
15	area_se	Standard error of area
16	smoothness_se	Standard error of smoothness (local variation in radius lengths)
17	compactness_se	Standard error of compactness (perimeter^2 / area - 1.0)
18	concavity_se	Standard error of Concavity (severity of concave portions of the contour)
19	concave points_se	Standard error of Concave Points (number of concave portions of the contour)
20	symmetry_se	Standard error of Symmetry
21	fractal dimension_se	Standard error of Fractal dimension ("coastline approximation" - 1)
22	Radius_worst	Worst of radius (mean of distances from center to points on the perimeter)
23	texture_worst	Worst of texture (Standard deviation of gray-scale values)
24	perimeter_worst	Worst of perimeter
25	area_worst	Worst of area
26	smoothness_worst	Worst of smoothness (local variation in radius lengths)
27	compactness_worst	Worst of compactness (perimeter^2 / area - 1.0)
28	concavity_worst	Worst of Concavity (severity of concave portions of the contour)
29	concave points_worst	Worst of Concave (number of concave portions of the contour)
30	symmetry_worst	Worst of Symmetry
31	fractal dimension_worst	Worst of fractal dimension ("coastline approximation" - 1)
32	Diagnosis	M = malignant, B = benign

Table 2: UCI Breast Cancer Dataset Description [37]

3. Activity Recognition Dataset

This dataset was collected by mounting a wearable accelerometer on the chest of a person in order to notice the activities made by that person available on the UCI repository. capture the movements done by him. This dataset contains a total of 1926896 records with five features classified into 7 different categories [38].

4. MIT_BIH Arrhythmia Dataset

This dataset was collected from ambulatory ECG recordings, applied on 47 subjects under a study done byBIH Arrhythmia Laboratory between 1975 and 1979 available on Phsyionet dataset repository. It contains a total of 109496 records with 187 features classified into five classes [39].

4.2 Performance Evaluation Metrics

The quality of any statistical or a machine learning model can be measured by using different performance evaluation metrics. For analysis purposesAccuracy, Precision, Recall, and F1_Score are calculated and evaluated.'Confusion matrix' is one of the important metrics that is mainly used for performance evaluation purposes. It acts as a visualization tool through which the accuracy of the different classifiers is represented.

Columnsshow the predicted class and the rows show the real class as represented by Table 3.Mathematical expression associated with each metric is demonstrated by Table 4.

		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

Table 3: Confusion Matrix for Evaluating Performances of Different Data Mining Techniques [14]

- **True positive (TP)** denotes the number of positive samples that are accurately predicted.
- **True negative (TN)** denotes the number of negative samples that are accurately predicted.

- **False negative (FN)** denotes the number of positive samples that are incorrectly predicted.
- **False positive (FP)** denotes the number of negative sampes incorrectly predicted as positive.

Metric	Explanation	Formula
Recall /Sensitivity	Fraction of the no. of predicted positive instances to overall positive instances.	$\frac{TP}{TP+FP}$
Accuracy	Fraction of no. of exact prediction to the overall prediction made	$\frac{TP}{TP+TN+FP+FN}$
Specificity	Fraction of the no. of predicted negative instances to overall positive instances.	$\frac{TN}{FP+TN}$
Specificity	Fraction of the no. of predicted negative instances to overall positive instances.	$\frac{TN}{FP+TN}$
F1_Measure	The harmonic means among recall and precision	$\frac{2*\Pr ecision*Sensitivity}{\Pr ecision*Sensitivity}$

Table 4: Performance Evaluation Metrics for Different DMT [14]

4.3 Implementation Tool

Overall experimentation was performed on Intel(R) Core (TM) i3Processor up to 1.74 GHz and 8 GB RAM for implementation purposes, the most popular data science toolkit, Spyder IDE from Anaconda navigator was used.Spyder IDE is an open-source scientific environment integrated with Python as a programming language.

5. Experimental Results and Analysis

In this section,the previously studied data mining techniques have been utilized, tuned with hyperparameters, analyzed, and then compared to knowtheir performance on different datasets taken from different dataset repositories including UCI, Kaggle, and Physionet. The experimental result obtained after the implementation of distinct data mining techniques is described in this section. The schematic workflow diagram is represented in Fig. 2. After considering the data into Spyder IDE, exploration of data analysis starts, first data is pre-processed for handling missing and noisy data and features are scaled using standard scalar. After that data was split into a ratio of 0.25 which means 75% data was utilized for training purpose and 25% data was utilized for testing purpose.

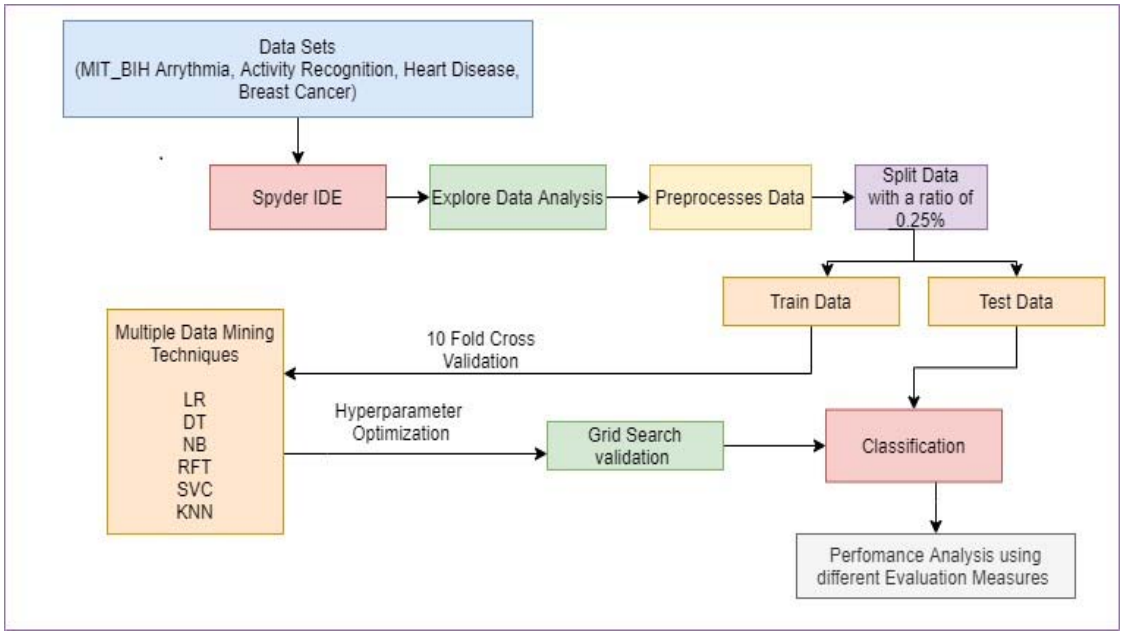

Fig. 2. Schematic Diagram Representing the Working Procedure utilized for the Prediction of Different Disease with the help of DMT

Further, six data mining techniqueswere examined by using 10-fold cross-validation on training data. Hyperparameters were tuned using the Grid search Cross Validation (GSCV) method so that highly efficient results can be achieved [33]. In last, analysis was done to find the most efficient data mining technique for making predictions.

5.1 Experimental result onHeart Disease Data Set

Table 5 and 6show the result for DMTimplemented on the heart diseases data set. On the basis ofcomparison done using different evaluation measures.It can be seen that Logistic Regression, indicated the lowest performance with an average of 0.79% f1_score, 0.82% recall, 0.78% precision , and 0.81% accuracy. In contrast NB outperformed Logistic Regressionwith the result of 0.84% f1_score, 0.89% recall, 0.81% precision , and 0.83% accuracy. KNN performed better than the first two, with an average of 0.83% f1_score, 0.85% recall, 0.82% precision , and 0.85% accuracy.

Sr. No.	DMT	Best Parameter	Accuracy Score
1	RFT	N_estimators=200	0.916966
2	SVM	C=10, kernel=rbf	0.878864
3	KNN	N_neighbours=5	0.845231
4	DT	Max_features=7	0.882222
5	LR	Solver=liblinear	0.826142
6	NB	Var_smoothing=0.28480358684	0.836255

Table 5. DMT best accuracy score with tuned hyper-parameter

SVM and Decision Tree performances are quite closer to each other. Decision Tree showed an average accuracy of 88% while SVMshowed an average accuracy of 87%. Decision Tree showed an average precision of 87% while SVM showed an average precision of 88%. Decision Tree showed an average f1_score of 83% while SVMshowed an average f1_score of 80%. RFT outperforms among all the techniques with an average of 0.88% f1_score, 0.87% recall, 0.91% precision , and0.91% accuracy.

Techniques	Accuracy	Precision	Recall	F1_Score
RFT	0.91	0.91	0.87	0.88
SVM	0.87	0.80	0.80	0.80
KNN	0.85	0.82	0.85	0.83
DT	0.88	0.87	0.80	0.83
LR	0.81	0.78	0.82	0.79
NB	0.83	0.81	0.89	0.84

Table 6: Performance metrics of different DMT for Heart Disease Dataset implemented in Python

5.2 Experimental result on Breast Cancer Data Set

Sr. No.	DMT	Best Parameter	Accuracy Score
1	RFT	N_estimators=70	0.960078
2	SVM	C=2, kernel=rbf	0.981118
3	KNN	N_neighbours=5	0.960188
4	DT	Max_features=3	0.922591
5	LR	Solver=liblinear	0.976523
6	NB	Var_smoothing=0.28480358684	0.941196

Table 7. DMT best accuracy score with tuned Hyperparameter

Table 7 and 8 show the result for data mining techniques implemented onthe breast cancer data set. On the basis of comparison done using different evaluation measures. The lowest performance shown by the Decision Tree with an

average of 0.91% f1_score, 0.85% recall, 0.87% precision , and0.92% accuracy. Incontrast, Naive Bayes outperformed Decision Tree with the result of 0.95% f1_score, 0.98% recall, 0.93% precision, and 0.94% accuracy. Logistic Regression performed better than first two, with an average of 0.96% f1_score, 0.96% recall, 0.98% precision , and 0.97% accuracy.

Techniques	Accuracy	Precision	Recall	F1_Score
RFT	0.96	0.96	0.97	0.95
SVM	0.98	0.98	0.94	0.95
KNN	0.96	0.95	0.87	0.95
DT	0.92	0.87	0.85	0.91
LR	0.97	0.98	0.96	0.96
NB	0.94	0.93	0.94	0.95

Table 8: Performance metrics of different DMT for Breast Cancer Dataset derived in Python Programming Language

Random Forest and KNN performances are quite closer to each other. Random Forest Tree and SVMshowed an average accuracy of 96%.Random Forest Tree showed an average precision of 97% while KNN showed an average precision of 95%. Random Forest Tree showed an average recall of 95% while KNN showed an average recall of 97%. SVMoutperforms among all the techniques with an average of 0.95% f1_score, 0.94% recall, 0.98% precision, and0.98% accuracy.

5.3 Experimental result on MIT-BIH Dataset

Table 9 and 10 show the result for data mining techniques applied on the MIT_BIH Arrhymia data set. On the basis ofcomparison done using different evaluation measures. It can be found that Naive Bayes, showed the lowest performance with an average of 0. 31% f1_score, 0.60% recall, 0.34% precision, and 0.50% accuracy. In contrast,Decision Tree outperformed Naive Bayes with the result of 0.77%
f1_score, 0.77% recall, 0.78% precision, and 0.94% accuracy. Logistic Regression performed better than the first two, with an average of 0.77% f1_score, 0.70% recall, 0.88% precision , and 0.95% accuracy. Random Forest, SVM and KNN performances are quite closer to each other. Random Forest Tree, SVMand KNN showed an average accuracy of 97%. Random Forest Tree showed an average precision of 96% while KNN showed an average precision of 92% and SVM showed an average precision of 96%. Random Forest Tree showed an average recall of 80% while KNN showed an average recall of 84% and SVM showed an average recall of 79%. Overall result analysis showed that Random Forest Tree outperforms among all the techniques with an average of 0.87% f1_score, 0.80% recall, 0.96% precision, and 0.97% accuracy.

Sr. No.	DMT	Best Parameter	Accuracy Score
1	RFT	N_estimators=200	0.975912
2	SVM	C=1, kernel=poly	0.971949
3	KNN	N_neighbours=5	0.972977
4	DT	Max_features=7	0.947244
5	LR	Solver=newton_cg	0.953709
6	NB	Var_smoothing=1.0	0.504055

Table 9: DMT best accuracy score with tuned hyperparameter

Techniques	Accuracy	Precision	Recall	F1_Score
RFT	0.97	0.96	0.80	0.87
SVM	0.97	0.93	0.79	0.85
KNN	0.97	0.92	0.84	0.88
DT	0.94	0.78	0.77	0.77
LR	0.95	0.88	0.70	0.77
NB	0.50	0.34	0.60	0.31

Table 10: Performance metrics for Different DMT for MIT_BIH dataset evaluated in Python

5.4 Experimental result on Activity Recognition Dataset

Table 9 shows the result for data mining techniques applied on the Activity Recognition data set. On the basis ofcomparison doneusing different evaluation measures. The lowest performance shown by the Logistic Regression with an average of0. 36% f1_score, 0.46% recall, 0.38% precision, and0.66% accuracy. Incontrast Naive Bayes outperformed Logistic Regression with the result of 0.36% f1_score, 0.46% recall, 0.38% precision, and 0.75% accuracy. KNN performed better than the first two, with an average of 0.85% f1_score, 0.83% recall, 0.86% precision, and 0.94% accuracy.Decision Tree achieved high performance than the first three with anaverage of 0.91% f1_score, 0.91% recall, 0.91% precision, and 0.96% accuracy. Random Forest, SVM performances are quite closer to each other. Random Forest Tree and SVM showed an average accuracy of 97%. Random Forest Tree showed an average precision of 96% and SVM showed an average precision of 93%. Random Forest Tree showed an average recall of 80% and SVM showed an average recall of 78%. Overall result analysis showed that the Random Forest Treeoutperforms among all the techniques with an average of 0.87% f1_score, 0.80% recall, 0.96% precision, and 0.97% accuracy.For a better view, Fig. 3-6 gives the graphical representation of the comparison done on different datasets using different data mining techniques.

Techniques	Accuracy	Precision	Recall	F1_Score
RFT	0.97	0.96	0.80	0.87
SVM	0.97	0.93	0.75	0.95
KNN	0.94	0.86	0.83	0.85
DT	0.96	0.91	0.91	0.91
LR	0.66	0.38	0.46	0.36
NB	0.75	0.38	0.46	0.36

Table 11: Performance Metrics of DMT for Activity Recognition Dataset Implemented in Python

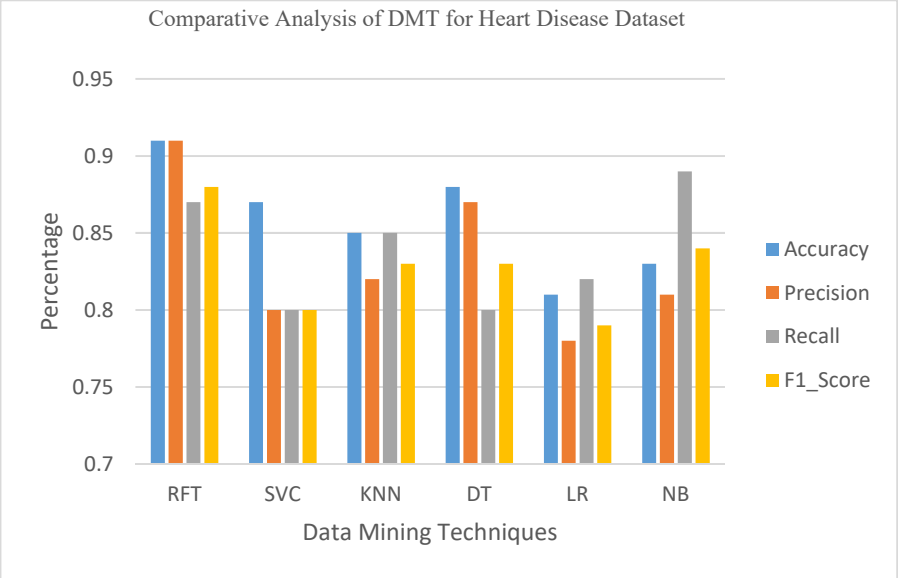

Fig. 3.Graphical Representation of Comparative Analysis of DMT for Heart Disease Dataset

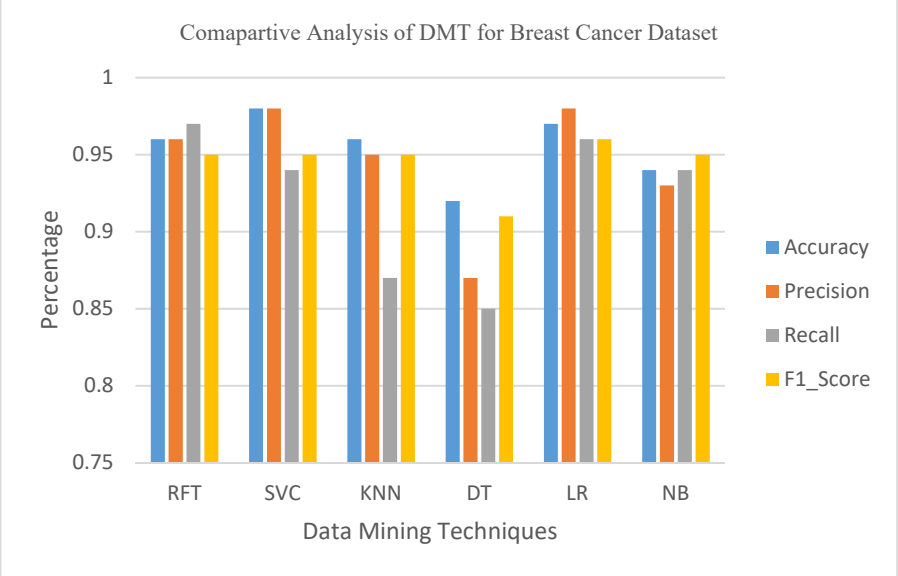

Fig. 4. Graphical Representation of Comparative Analysis of DMT for Breast Cancer Dataset

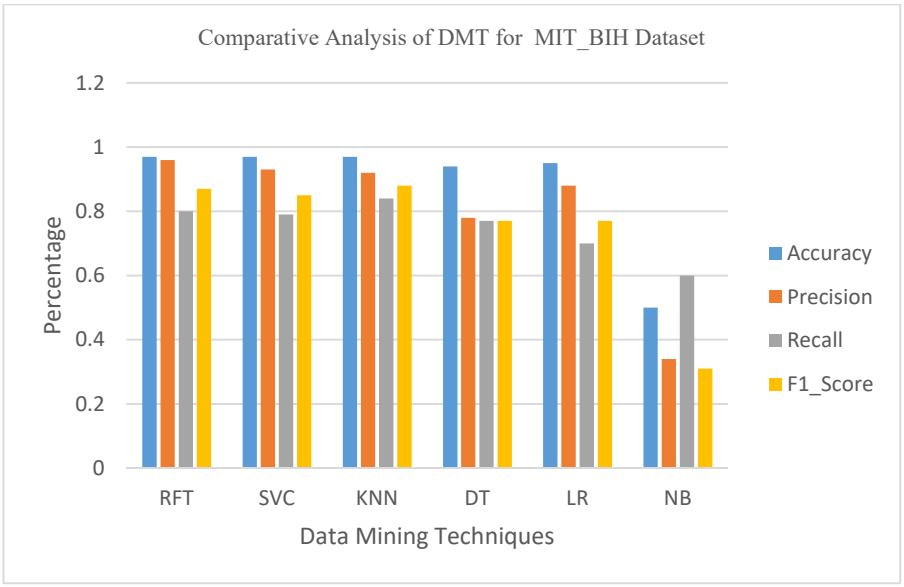

Fig.5. Graphical Representation of Comparative Analysis of DMT for MIT_BIH Dataset

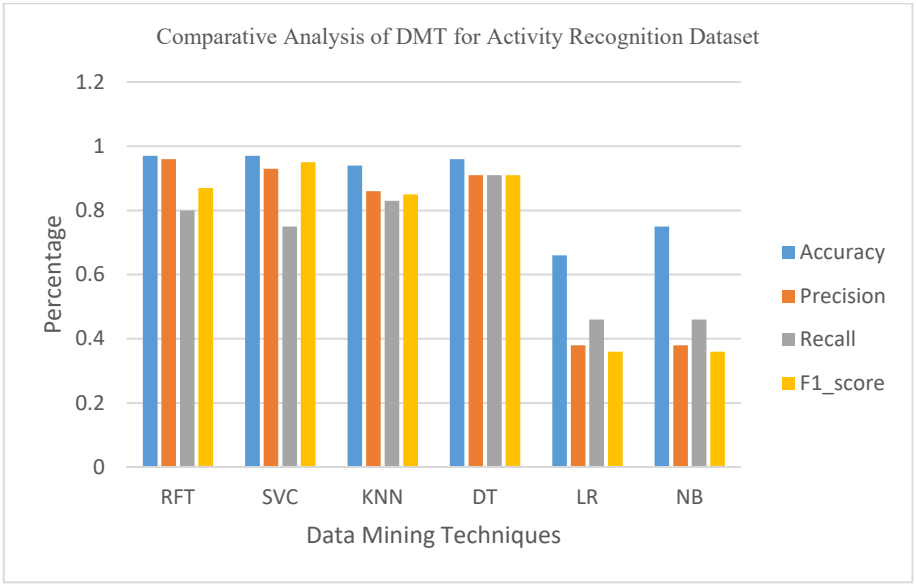

Fig.6. Graphical Representation of Comparative Analysis of DMT for Activity Recognition Dataset

6. Conclusion

With the growing number of chronic diseases, a large amount of data regarding patient health must be generated. This data must be analyzed by using some effective mining techniques for the timely forcasting of diseases. To know the effective data mining technique, a comparison between their performances must be done. The novelty of this work is mainly associated with the comparative analysis of the techniques on the basis of datasets of different nature either generated clinically or by the real monitoring of patients by using sensors, accelerometer, and ECG. The curiosity of this work is mostly connected with thefinding of the optimized data mining technique for making an accurate prediction regarding patient health by performing the comparative analysis of six mostfamiliar DMT: RFT, DT, NB, Logistic Regression and KNN by utilizing Accuracy, Recall, Precision, and F1_Score as evaluation metrics. The result shows that among all the techniques, RFT provides the best result for all types of datasets with an effective accuracy of 97%and works well with datasets belong to bothbinary as well as multiclass. Moreover, after RFT, SVM

can be considered the other promising technique which shows the result more closer to RFT. This study will assists researchers to appreciate the proficiency of data mining techniques on sensor datasets. With the growing use of sensor-based devices for monitoring patient health, it becomes necessary to find the data mining technique which can effectively work with sensor datasets. New data mining techniques can be considered and analyzed in near future for other datasets.

Acknowledgment

I highly thank to Dr. Pooja Mittal for their comments and supports in the preparation of this manuscript.

Conflicts

The authors have no conflicts of interest to declare.

References

[1] Ray. A.; Chaudhuri, A.K.(2020).Smart healthcare disease diagnosis and patient management: Innovation, improvement and skill development,Machine Learning with Applications,Volume 3,2021,ISSN 2666-8270.
[2] Suzman, R.; Beard,J.(2020).World Health Study on Global Health and Aging https://www.who.int/ageing/publications/global_health.Pdf.
[3] Elfein, J. (2020). Diseases - Statistics & Facts _ Statista, https://www.statista.com/topics/2070/diseases/.
[4] Rath, M.; Pattanayak, B.(2019). Technological improvement in modern health care applications using Internet of Things (IoT) and proposal of novel health care approach. International Journal of Human Rights in Healthcare, 12: 148–162. doi: 10.1108/IJHRH-01-2018-0007.
[5] Pattnayak, P.; Panda, A.R. (2021). Innovation of Machine learning techniques in healthcare services, Technical Advancements of Machine Learning in Healthcare (pp.1-30).
[6] Mandal, S.K. (2017). Performance analysis of data mining algorithms for breast cancer cell detection using Naïve Bayes, logistic regression and decision tree. International Journal of Engineering and Computer Science, 6(2), pp.20388-20391.
[7] Gupta, S.; Kumar; D.; Sharma, A.(2011) . Performance analysis of various data mining classification techniques on healthcare data. International journal of computer science & Information Technology (IJCSIT), 3(4), pp.155-169.
[8] Muhammad,L. J.; Ebrahem A.; Usman, S.S;Chakraborty, AA.C; Mohammed,I. A. (2021). Supervised Machine Learning models for prediction of COVID-19iInfection using Epidemiology dataset, SN Computer Science, 2(11), pp. 1-13.
[9] https://en.wikipedia.org/wiki/Monitoring_(medicine)
[10] https://www.google.com/search?q=wireless%20body%20sensors&tbm=isch&hl=en&sa=X&ved=0CB0QtI8BKABqFwoTCLCYupnGn_ICFQAAAAAdAAAAABAC&biw=1499&bih=667#imgrc=Y9sTSGd9FHsClM&imgdii=j0clUgRFkLA_4M
[11] Z.K. Senturk ; Kara, R..(2014). Breast cancer diagnosis via data mining: performance analysis of seven different algorithms. Computer Science & Engineering, 4(1), p.35.
[12] Vijayarani, S.; Dhayanand, S.(2015). Liver disease prediction using SVM and Naïve Bayes algorithms. International Journal of Science, Engineering and Technology Research (IJSETR), 4(4), pp.816-820.
[13] Venkatalakshmi, B.; Shivsankar, M.V. (2014). Heart disease diagnosis using predictive data mining. International Journal of Innovative Research in Science, Engineering and Technology, 3(3), pp.1873-7.
[14] Sayed, L.; Jabeen, S.; Manimala, S.: Elsayed, H.A.(2019). Data Science Agorithms and techniques for Smart Healthcare using IoT and Big Data Analytics. Smart Techniques for Smarter Planet. Studies in Fuzziness and Soft Computing, Springer, 211-241. https://doi.org/10.1007/978-3-030-03131-2_11.
[15] Syed, L.; Jabeen, S.; Manimala, S.(2018). Telemammography: A novel approach for early detection of breast cancer through wavelets based image processing and machine learning techniques. Advances in Soft Computing and Machine Learning in Image Processing, 730: 149-183.
[16] Verma, P.; Sood, S.K.; Kalra, S.(2018). Cloud-centric IoT based student healthcare monitoring framework. Journal of Ambient Intelligent and Humanized Computing, 9(5): 1293–1309. doi: 10.1007/s12652-017-0520-6.
[17] Ganesan M and Sivakumar N. IoT based heart disease prediction and diagnosis model for healthcare using machine learning model. IEEE International Conferenc of System Computation, Automation and Networking, 2019, 1–5. doi: 0.1109/ICSCAN.2019.8878850.
[18] Verma, P.; Sood, S.K.(2018). Cloud-centric IoT based disease diagnosis healthcare framework. Journal of Parallel Distribution and Computing, 116: 27–38. doi: 10.1016/j.jpdc.2017.11.018.
[19] Aich, S.; Younga, K.; Hui, K. L, Al-Absi, A. A.; Sain, M.A(2018) A nonlinear decision tree based classification approach to predict the Parkinson's disease using different feature sets of voice data. International Conference of Advanced Communication and Technology, 638–642. doi: 10.23919/ICACT.2018.8323864.
[20] Verma, P.; Sood, S.K.(2018). Cloud-centric IoT based disease diagnosis healthcare framework. Journal of Parallel Distribution and Computing, 116: 27–38. doi: 10.1016/j.jpdc.2017.11.018.
[21] Chiuchisan, I.; Geman, O. (2014) An approach of a decision support and home monitoring system for patients with neurological disorders using internet of things concepts. WSEAS Transaction on System, 13(1): 460–469.
[22] Khan, M A.(2020). An IoT Framework for Heart Disease Prediction Based on MDCNN Classifier. IEEE Access, 8: 4717–34727. doi: 10.1109/ACCESS.2020.2974687.
[23] Memon, M. H; Li, J. P.; Haq, A.U; Memon, M.H, Zhou, W.; Lacuesta, R. (2019). Breast Cancer Detection in the IOT Health Environment Using Modified Recursive Feature Selection. Wireless Communiation and Mobile Computing, 2019. doi: 10.1155/2019/5176705.
[24] Abdelaziz, A. Salama, A.S ; Riad, A.M; Mahmoud, A N.(2019). A Machine Learning Model for Predicting of Chronic Kidney Disease Based Internet of Things and Cloud Computing in Smart Cities. Security in Smart Cities: Models, Applications, and Challenges, Springer International Publishing, 93-114.
[25] Boukenze, B.; Mousannif, H. ; Haqiq, A. (2016). Performance of Data Mining Techniques to Predict in Healthcare Case Study : Chronic Kidney Failure Disease. International Journal of Database Management System, 8(3): 1–9. doi: 10.5121/ijdms.2016.8301.
[26] Vanaja, S ; Rameshkumar, K. (2015). Performance analysis of classification algorithms on medical diagnoses-a survey. Journal of Computer Science, 11(1): 30–52. doi: 10.3844/jcssp.2015.30.52.
[27] Fatima, M.; Pasha, M.(2017). Survey of Machine Learning Algorithms for Disease Diagnostic. Journal of Intelligent Learning System and Applications, 09(01): pp. 1–16. doi: 10.4236/jilsa.2017.91001.

[28] Khanom, N. N; Nihar, F. ; Hassan, S.S; Islam, L (2020). Performance Analysis of Algorithms on Different Types of Health Related Datasets. Journal of Physics: Conference Series, 1577(1). doi: 10.1088/1742-6596/1577/1/012051.
[29] Datar, S.; Jain, A. (2020) Design and performance analysis of ecg data compression using convolved window-based cosine modulated filter bank. Journal of Engineering Science and Technology, 15(5): 3449–3464.
[30] Vitabile , S; Michal, M.; Stoanovic, D.. (2019).Medical Data Processing and Analysis for Remote Health and Activities Monitoring. High-Performance Modelling and Simulation for Big Data Applications. Lecture Notes in Computer Science, vol 11400. Springer, ,https://doi.org/10.1007/978-3-030-16272-6_7.
[31] Asif-Ar-Raihan, Md.; Noshant, M.M.; Faisal. F.; Dip, R.R.(2021). Performance Evaluation and comapartive analysis of different machine learning algorithms in predicting cardiovascular disease, Journal of Engineering Latter, 29 (2).
[32] Awaan, F.M; Saleem, Y.; Minerva, R.; Crespi, N.(2020). A comparative analysis of machine/deep learning models for parking space availability prediction, Journal of Sensors, 332 (20), 2020, doi:10.3390/s20010322.
[33] Singh, A.; Yadav, A.; Rana, A(2013) K-means with Three different Distance Metrics. International Journal of Computer Applications, 67.
[34] Scikit Learn. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html (accessed on 25 September 2019).
[35] Activity recognition dataset https://archive.ics.uci.edu/ml/datasets/Activity+Recognition+from+Single+Chest-Mounted+Accelerometer
[36] Breast Cancer dataset https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29
[37] Heart disease dataset https://www.kaggle.com/ronitf/heart-disease-uci
[38] MIT_BIH Arrhythmia dataset https://www.physionet.org/content/mitdb/1.0.0/

About the Authors

Dr. Pooja Mittal is currently working as Assistant Professor at Department of Computer Science & Applications, M.D.University, Rohtak, India. She obtained her Ph.D from Maharshi Dayanand University. Her area of resaerch and specializtaion include Data Mining, Data Warehousing and Computer Science. She has published more than 50 resaerch papers in renowed international and SCI Journals and attended more than 30 conferences.

Ms. Navita has completed her M.Tech from GJU S&T University. She is currently pursuing Ph.D in Computer Science at Department of Computer Science & Applications, M.D.University, Rohtak. Her main research area includes Internet of Things (IoT) and Data Mining.