CDPS-IoT: Cardiovascular Disease Prediction System Based on IoT Using Machine Learning

Jameel Ahamed[1], Abdul Manan Koli[1], Khaleel Ahmad[1]*, Mohd. Alam Jamal[1], B. B. Gupta[2]*

[1] Department of Computer Science & Information Technology, Maulana Azad National Urdu University, Gachibowli, Hyderabad (India)
[2] Department of Computer Engineering, National Institute of Technology Kurukshetra, India & Asia University (Taiwan)

Abstract

Internet of Things, Machine learning, and Cloud computing are the emerging domains of information communication and technology. These techniques can help to save the life of millions in the medical assisted environment and can be utilized in health-care system where health expertise is less available. Fast food consumption increased from the past few decades, which makes up cholesterol, diabetes, and many more problems that affect the heart and other organs of the body. Changing lifestyle is another parameter that results in health issues including cardio-vascular diseases. Affirming to the World Health Organization, the cardiovascular diseases, or heart diseases lead to more death than any other disease globally. The objective of this research is to analyze the available data pertaining to cardiovascular diseases for prediction of heart diseases at an earlier stage to prevent it from occurring. The dataset of heart disease patients was taken from Jammu and Kashmir, India and stored over the cloud. Stored data is then pre-processed and further analyzed using machine learning techniques for the prediction of heart diseases. The analysis of the dataset using numerous machines learning techniques like Random Forest, Decision Tree, Naive based, K-nearest neighbors, and Support Vector Machine revealed the performance metrics (F1 Score, Precision and Recall) for all the techniques which shows that Naive Bayes is better without parameter tuning while Random Forest algorithm proved as the best technique with hyperparameter tuning. In this paper, the proposed model is developed in such a systematic way that the clinical data can be obtained through the use of IoT with the help of available medical sensors to predict cardiovascular diseases on a real-time basis.

Keywords

I. Introduction

Since the last few decades, the World witnessed an increase in the death rate due to cardiovascular diseases (CVD). In the United States of America, one person is dying every minute due to the arrest of CVD. Numerous researchers had tried the classification techniques of machine learning to diagnose CVDs to help medical practitioners in improving the health care system world-widely. According to the World Health Organization (WHO), an estimation of 30% of global deaths is caused by CVDs over three-quarters of which catch a spot in low and middle-income countries. CVDs also killed 25% of the population in the age group of 25-69 years in India[1]. Internet of Things (IoT) is usually referred to as physical things connected to the internet world with limited storage and processing ability. It is still struggling with performance, interoperability, security, and privacy issues challenges with a huge scope for improvement in the near future [2], [3], [4]. IoT is also proven as a benchmark for healthcare systems wherein smart objects are used to continuously monitor the patient for particular diseases [2]. Smart objects comprise biomedical sensors which gather health-related information and pass it to the physician through cloud/edge for further diagnosis. Hence, IoT helps in bridging the gap between patient and Physician located at any geographical location [5], [6]. Cloud computing has the nearly endless capacity of processing power and storage capacity, which is a more advanced technology to a specific extent to resolve the technical issues in the IoT while data mining is an intelligent technology that is used to extract new information by examining huge voluminous datasets [7]. It can be utilized to make certain decisions, estimates, and predictions using different machine learning algorithms. In the present era, most of the data in the medical sector are collected via the computerized system but not utilized world-wide for analysis. It is stacked up in a database like old handwritten records and put to no use. This data can be harnessed to predict diseases such as Cancer, CVDs, Diabetes, Dengue, etc., [2], [8]. Thus, an innovative Information technology (IT) paradigm is proposed, in which IoT and machine learning with cloud computing are the three interrelated technologies integrated together to overcome both current and future world obstacles related to the health-care system and termed as IoT-ML-Cloud paradigm. Medical assisted technologies and healthcare services are nearly related and remedial to the public welfare for better health-care facilities. The collaboration of cloud computing and IoT for the applications of

* Corresponding author.

E-mail addresses: gupta.brij@gmail.com (B. B. Gupta), khaleelahmad@manuu.edu.in (K. Ahmad).

modern medical assisted technologies play an important role in the prediction of chronic diseases. The development of public cloud (cloud computing) in hospitals, which has benefits like high security, improved efficiency, virtualization, reliability, and scalability can promote resource sharing, cost savings, medical monitoring, management, and administration system with high efficiency and accuracy.

II. Related Work

Many research studies have been done for the prediction of diseases using artificial intelligence, machine learning algorithms, IoT, etc. In paper [9], the authors proposed a healthcare monitoring system using a random forest algorithm assisted by the IoT. In this paper, different diseases like Heart diseases, Diabetes, and Breast Cancer were predicted and achieved a maximum accuracy of 97.26% on the Dermatology dataset using the Random Forest algorithm. In paper [10], the authors proposed a hybrid recommender system for cardiovascular diseases using IoT in a cloud environment and also introduced the multiclass classification problem which can predict the eight types of cardiovascular diseases [11] [12]. Further, the model improved accuracy by 98% using the feature selection techniques. In paper [11], the authors worked on lung cancer and built an IoT-based prediction system using Fuzzy Cluster-Based Segmentation and Classification. The proposed system was developed in a MATLAB environment that focused to classify the lung X-ray images. The proposed model got 85% accuracy. In paper [13], the authors developed an online web-based platform for Clinical Decision Support System (CDSS) which is based on Optimal Deep Neural Networks (DNN). They developed a Cloud-based CDSS framework for the prediction of Chronic Kidney Diseases (CKD) with their level of severity. They employed IoT sensors to get the clinical data from disease-affected patients and predict the condition of normal & abnormal and achieved the highest accuracy of 98% using DNN. In paper [14], the researchers built an android-based monitoring system that can monitor the heart rate of cardiac patients. Decision Tree algorithm was used to develop the model which can trigger an alarm if a patient has an abnormal heart rate. In the paper [15], the authors proposed an IoT-based system that is used to early detect cardiac disease using machine learning algorithms. In this work, they took heart data from the MAX30100 Sensor connected to Arduino and sent the observed data to Oxywatch for further analysis. In paper [16], the researchers analyzed IoT safety, security, and privacy feature including the requirements of a security issue, use patterns, and attack taxonomies from the health care prospect. They also invented a wearable health care environment to decide how people get facilities by economic and societies in terms of sustainability for handling noisy, missing values using Decision tree-based C4.5 classifier. In the paper [17], the authors used some statistical methods like Nearest Neighbor Imputation, Hot Deck imputation, Cold Deck Imputation, Substitution, and Mean Substitution for prediction. In [18], the authors worked with 6 predictive state imputations, the frequency-based imputation used by the C4.5 algorithms, and reduces modeling for a classification tree. In the paper [19], the authors suggested the approach of only numerical values to assign the disappeared values that can able to handle categorical attributes, and in last, they compare with the different factors i.e. cost, space, time, etc., which further helps in analyzing the data [19] [20]. In the paper [21], the researchers performed the comparison between decision trees, k-nearest neighbors' algorithm (K-NN), and Naïve Bayes algorithm using Rapid miner, WEKA, Tanagra, Knime, and Orange tools on Dataset of Liver Patients [22]. In the paper [23], the authors proposed a system to deal with missing values in test datasets and training datasets for predicting diabetic disease in patients. Paper [24] showed a comparison of different machine learning algorithms with the help of RapidMiner, Tanagra, WEKA tool, Knime and Orange. In paper [25],

the diabetic dataset was analyzed with Classification and Regression Trees (CART), Support Vector Machine (SVM), K-NN, J48 algorithms for finding classifications. The accuracy rate of K-NN was 53.39% for correct classification and incorrect classification was 46.605%.

III. Findings From Related Work

We did not find any reliable model which can accurately predict cardiovascular disease in developing nations. Most of the researchers developed the cardiovascular disease models based on single and two algorithm combinations which are not strong enough to make accurate predictions in developing nations. To overcome these research gaps, we performed the following:

I. We used three different feature selections and also the five classifiers for cardiovascular disease prediction. Till date, we have not found any research work which utilized such classifiers for feature selections and cardiovascular disease predictions. Mostly, research scholars used only one or two algorithms in their research work.

II. We did not find any research work pertaining to cardiovascular disease prediction based upon clinical data for the area of Jammu and Kashmir (India) using machine learning techniques, therefore, this research work may be considered as the first cardiovascular disease prediction model for such developing areas.

III. We applied machine learning techniques for cardiovascular disease because it is considered to be one of the best and most innovative techniques for prediction work and is also used for other disciplines like cancer disease prediction, stock market predictions and weather forecasting, etc.

IV. Machine Learning Algorithms

In this research, five types of machine learning algorithms are used for the development of the disease prediction model. The algorithms for disease prediction model are k-Nearest Neighbor (KNN), Decision Tree classifier (DTC), Support Vector Machine (SVM), Random Forest (RF), and Naïve Bayes (NB). The main aim of these algorithms is classification. As our dataset has an output class label, hence these supervised algorithms have been chosen because they work with class label problems.

A. K-Nearest Neighbors

K-Nearest Neighbors is the most simple classifier that stores all the available cases and classifies new instances based on nearest majority neighbors (for example distance function) [26]. K-NN has been adopted in statistical estimation and pattern recognition was already considered as a non-parametric tool. K-NN assumes that identical classes exist in shut proximity i.e entities that are similar, exist together. In the name of the K-NN algorithm, the alphabet 'K' means the number of nearest neighbors for determining the class of an instance, as shown in Fig. 1 [9].

In this algorithm, two parameters are obligatory that is neighborhood cardinality (K), and measure to evaluate the similarity. Mathematically, we can calculate the Euclidean distance between two points. The Euclidean distance equation is given below.

$$\rho(x, x') = |x - x'| = \sqrt{\sum_{k=0}^{n} (xi - xi')^2}$$

(1)

K represents the number of the data point in the KNN algorithm that is close to the new data point [11]. For instance, if K = 1, 2, 3 then it will choose one nearest neighbor, two nearest neighbors three nearest neighbor data points respectively.
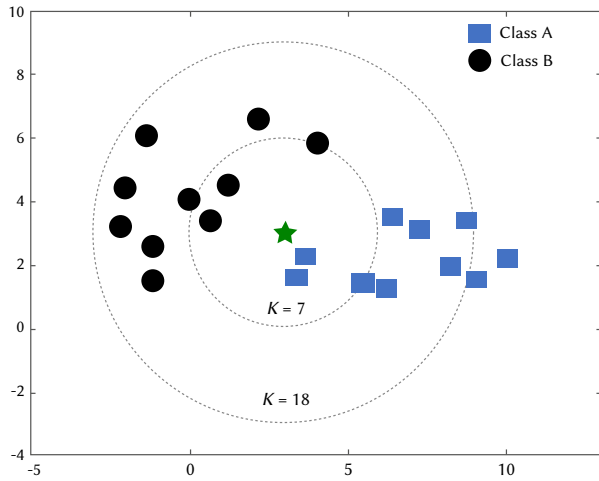
Fig. 1. K-Nearest Neighbors.

Now, it will classify the data point based on the majority of voting [27]. This idea is portrayed in Fig. 1. After classification, the new data point is represented by the green star point using the nearest neighbor technique, as depicted in Fig. 1. Here two classes A and B which is represented by a sky color rectangle and black color circle respectively. For K = 7, the star is close to the rectangle, hence the KNN algorithm classified it as a class A.

Before applying KNN, data should be cleaned and no outlier or noise value should be presented in the dataset. And also, the variable should be normalized otherwise the higher value variables can lead the model bias.

B. Random Forest

Random forest is a classification algorithm and it constructs multiple decision trees during the training phase[13]. The random forest algorithm takes a final decision to choose the trees based on the majority of voting [14]. It reduces the risk of overfitting of the model based on the utilization of multiple trees [10]. It works efficiently on a large database and produces highly accurate results [27].

Steps of Random Forest algorithm:

- Pick random R data objects from the training dataset.
- Create a DT (Decision Tree) for the K data point.
- Picks the n-tree subset from the newly generated trees and execute step 1 and step2
- Take the decision or result based on the majority of voters.'

Fig. 2 portrays the working of the Random Forest algorithm. We made three different Decision trees from the dataset, but the final classification algorithm takes decisions based upon the majority of voters.
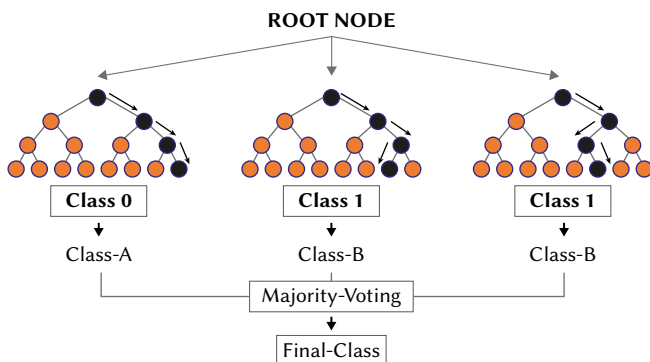

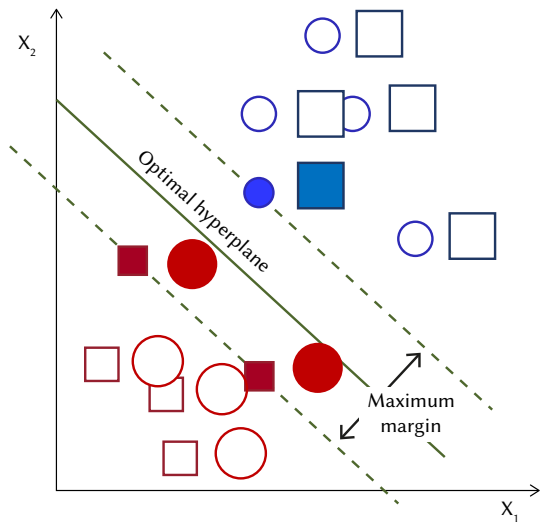Fig. 2. Random Forest Algorithm.

Fig. 3. Support Vector Machine.

C. Support Vector Machine (SVM)

SVM is a classifier or a classification algorithm that uses the hyperplane to classify the data [15]. The primary objective of SVM has to find the best and optimal hyper-plane in n-dimension space because it requires such type of hyper-plane which has a maximum margin [16]. Hence, the thumb rule is choosing the hyper-plane, which separates the two classes better.

In Fig. 3, the circle and rectangle are two classes and separated from each other by hyper-plane. But the queries raised here that how many hyper-planes are possible and which one has to choose out of possible hyper-planes. It will be chosen based upon the maximum margin between the classes. For the non-linear data points, SVM uses the kernel trick to draw the hyper-plane between them. Hence, it can say that SVM also works for non-linear data points [27].

D. Decision Tree

A decision tree can be used for both regression and classification problems. It is a non-parametric supervised learning method. It is a tree-structured classifier where the characteristic of a dataset is defined by internal nodes and branches of the tree represent the decision rules and the outcome is defined by each leaf node [28].
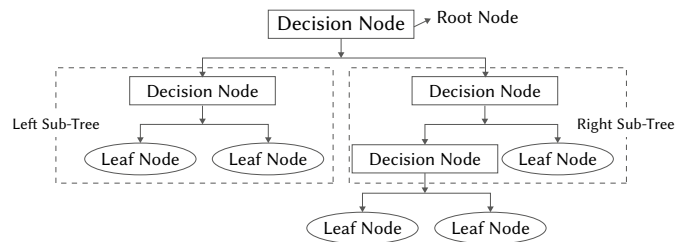

Fig. 4. Decision Tree.

There are two nodes in the Decision tree, which are the leaf node and the decision node. Decision nodes have several branches and are used to make some decisions while leaf nodes are the outcome of such decisions and do not have any further branches. A test or decision is carried out based on features of the dataset.

A decision tree asks a question, and it further partitions the tree into sub-trees based on the appropriate response (Yes / No), as portrayed in Fig. 4. A decision tree includes categorical data (YES/NO) as well as numeric data [29].

V. Primitive Statistics

The aim of this research work has to predict heart diseases which is a very delicate factor and risky as well. If done properly, it can be used by the medical administration for the proper diagnosis of patients who are suffering from CVDs. The dataset is obtained from Jammu and Kashmir (India), which is a real-world dataset. Thereafter, a combination of five ML-based classification algorithms used viz-a-viz Naïve Bayes, K-Nearest Neighbors, Decision tree, Random Forest, and Support Vector Machine.

The heart disease prediction model was developed using the anaconda tools which can handle large datasets with Jupyter notebook in python language. The developed model is then validated by medical domain experts.

To initialize the development of the model following suitable steps were performed:

A. Data Collection

This is the first step of model development in this step's dataset related to cardiovascular disease is collected from Jammu and Kashmir (India). Information and description of the dataset are given in Table I.

TABLE I. Attributes Information and Description of Dataset

S.No.	Attribute	Description
1	Age	Age in Years
2	Sex	Gender
3	trestbps	Resting blood pressure (in mm Hg
4	chol	serum cholesterol in mg/dl
5	fbs	fasting blood sugar > 120 mg/dl
6	restecg	resting electrocardiographic results
7	thalac	maximum heart rate achieved
8	Alcoholic	Alcohol Consumption
9	oldpeak	ST depression induced by exercise relative to rest
10	Ca	number of major vessels (0-3) colored by fluoroscopy
11	Slope_0	the slope of the peak exercise ST segment
12	Stress_0	Stress level
13	Thal_0	Thalassemia

B. Data Preprocessing:

In this step, the collected data is preprocessed using various methods like mean, mode, and median.

C. Feature/Attributes Selection Techniques:

After the dataset is cleaned and is ready for the mined purpose. We applied various attributes/features selection algorithms like Embedded Methods, Wrapper Methods, and Filter Methods. Then we take the mean of all the utilized methods and based upon the obtained mean we observed the weightage of each attribute.

D. Classification Algorithms:

Supervised learning algorithms like SVM, k-NN, Decision Tree, etc. were employed for the development of the proposed model.

VI. Experimentation

The proposed heart disease prediction model based on IoT-Cloud-ML consists of four-tier architecture starting from data collection, data storage, data pre-processing & analysis, and development of Graphical User Interface (GUI). Tier 1 focused on collecting health care data using IoT or wearable devices. Tier 2 and Tier 3 focused on cloud computing where data storage service and machine learning models are deployed. Tier 4 focuses on the front-end graphical user interface from where a patient or a doctor can view and interpret the results.

Tier 1 – Data Collection: Data collection tier is employed for collecting the patient's data using different IoT medical sensors and wearable devices. These IoT-based healthcare devices attached to the patient's body continuously collect clinical data. These devices then transmit the data to the cloud using various communication technologies like Wi-Fi, ZigBee, GPRS/LTE. Algorithm 1 shows IoT device initialization, data collection in a continuous manner, and transmission of data to the cloud database. In this system, we used 'THINGSPEAK Cloud' to collect the medical sensor data with the help of the NODEMCU server. The data which we used in our work was taken from Jammu and Kashmir as a repository system. Further data collected from IoT-based devices will be employed in this model for real-time prediction [30].

Algorithm 1: IoT device initialization and sending data to the cloud

Step 1: *attach the IoT medical sensors in patients' body and Start Device Initialization*

Step 2: *configure the nodemcu with ssd and password and connect with wifi*

Step3 : *attach nodemcu with sensor to work as a mediator*
If(nodemcu==connectedwithwifi)
{
  If(sensor==attachedwithnodemcu)
  {
    Read data and send it to the cloud
  }
  Else
  {
    Error: Sensor is not attached
  }

} else {
  Error: Nodemcu is not connected
}

Tier 2- Data Storage: IoT devices employed in Healthcare, Industrial IoT have the objective of sensing patient clinical data and transmit it to data storage service in a continuous manner [31], [32], [33]. But it is a very complicated task to sense and send such data by conventional data processing tools and techniques. IoT devices and physical personal computers do not have sufficient storage to store and process the voluminous clinical data of cardiovascular patients etc., [34]. Hence, the proposed method used cloud computing for this task. In the proposed model, HEROKU cloud PostgreSQL is used for storing and pre-processing the data. HEROKU PostgreSQL cloud is a data storage service widely used for data storage as described in Algorithm 2.

Algorithm 2: Storage of Patient Data onto Heroku PostgreSQL database

Step 1: *initialize PostgreSQL in Heroku cloud*
Step 2: *setup PostgreSQL with user name and password*
Step 3: *if the database and table not created* **Then**
create a database and table.
Step 4: *configure IOT sensor data by using a thing speak write API key to insert*
Step5:   *if   (thingspeakDiseaseData==True   or PatientPersonalData==True)*
{
  Insert data into the table
}
Else {
  Error: no data to insert}

Tier 3 – Data Analysis on a cloud: Data analytics tier uses the machine learning models and techniques to predict the output class label. The proposed methodology implemented various types of machine learning algorithms and chose one of them which gives the best and accurate results. The dataset is divided into training and testing datasets. The trained data was treated with different classification algorithms like Decision Tree, SVM, KNN, Random Forest, and Naive based to further test the model for prediction of cardiovascular diseases as described in Algorithm 3.

Algorithm 3: Predication of the diseases using the proposed model

Step 1: *select the patient of which we need to predict the disease*
Step 2: *check the patient data and its attribute*
If(Any Attribute==null)
{
insert the required data into the database
}
Else {
Predict the disease and show the result}

Tier 4 - GUI on a cloud: This tier provides the frontend interface which can be accessed by both patient and the doctor (expert). The patient can view their clinical results using patient_id. This GUI was developed using the python flask library. The flask application is deployed on the cloud so that it can easily communicate with other services on the cloud.

Fig. 5 shows the overall proposed system composed of a four-tier system and able to communicate with each other by integrating Cloud computing, Machine Learning, and the Internet of things in a precise manner.
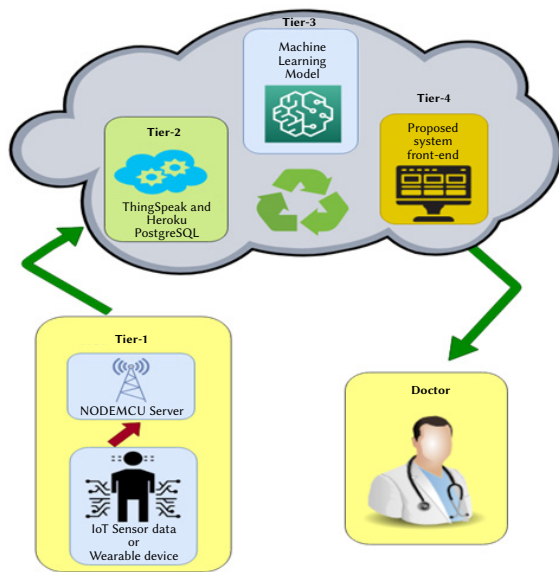

Fig. 5. Proposed Methodology Diagram.

VII. Results and Discussion

In this section, the authors applied different feature selection techniques like filter method, embedded method, and wrapper methods, and finally, their mean weightage is obtained for all attributes for their use in the development of the model, which are depicted in Table II and Fig. 6.

The results of the developed model are described in this section. From Fig. 6, it is clear that the persons having a high old peak has more chance of getting heart disease as compared to the Alcoholic person which has less chance of having heart disease based on weightage. In this way, the weightage scale can be analyzed.

TABLE II. Attributes Along With Their Mean Weightage

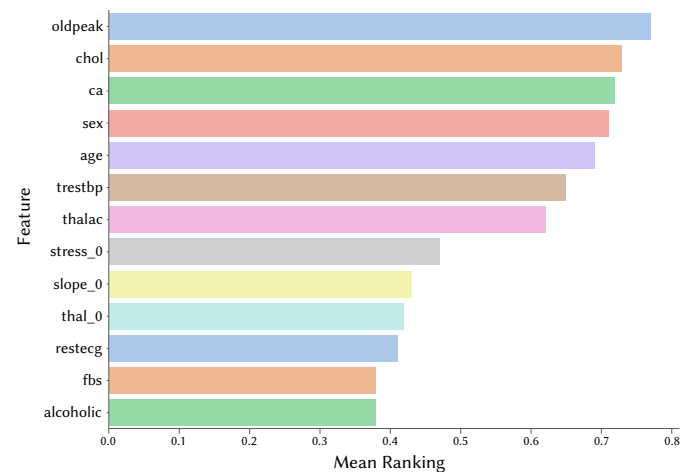| Feature Selection techniques	Filter Method	Embedded Method	Wrapper Method	Mean
age	0.98	0.18	0.92	0.69
sex	0.81	0.82	0.50	0.71
trestbp	0.99	0.20	0.75	0.65
chol	1.00	0.18	1.00	0.73
fbs	1.00	0.13	0.00	0.38
restecg	0.97	0.00	0.25	0.41
thalach	0.93	0.11	0.83	0.62
Alcoholic	0.08	0.73	0.33	0.38
oldpeak	0.73	0.90	0.67	0.77
Ca	0.61	0.97	0.58	0.72
Slope_0	0.98	0.14	0.17	0.43
Stress_0	0.00	1.00	0.42	0.47
Thal_0	1.00	0.18	0.08	0.42

Fig. 6. Attributes along with their Weightage.

After obtaining the mean weightage of all the attributes, various Machine learning classifiers like Decision tree (DTC), Random Forest search (RFS), K-nearest neighbor (KNN), Support vector machine (SVM), and Naïve Bayes (NB) are utilized in various situation and tested the performance of each classifier.
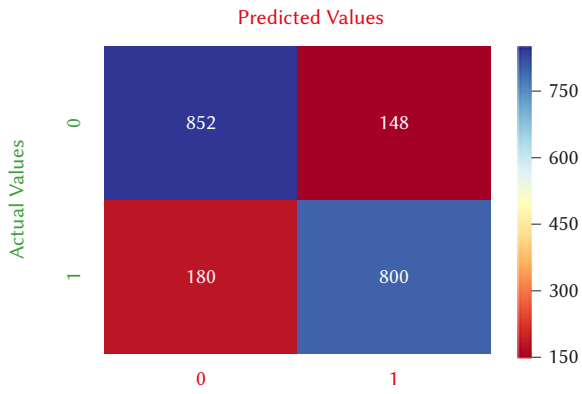
The confusion matrix along with the AUROC curve of a Decision tree is shown in Fig. 7.

Fig. 7 depicted the confusion matrix and AUROC curve for the decision tree, and its use to measure the performance of the classifier in terms of Accuracy, Precision, Recall, and F1-Score. Firstly, the Decision tree algorithm was applied to the test data set and found 83.43% accuracy with 84.38% precision, 81.63% recall, and 82.98% F1-Score. It means that the decision tree model has the capability of differentiating the person having heart disease or not is 83.43%.

The confusion matrix along with the AUROC curve of the Random forest Model is shown in Fig. 8.

Fig. 8 depicts the confusion matrix along with the AUROC curve for the Random Forest model. After applying Random Forest Algorithm on the test dataset, it was found that the (RF) model has 87.72% accuracy with 86.16% precision, 89.59% recall, and 87.84% F1-Score. It means that the Random Forest model has the capability of differentiating the person having or not having heart disease is 86.16%.

Confusion Matrix of Decision tree on Test DataSet

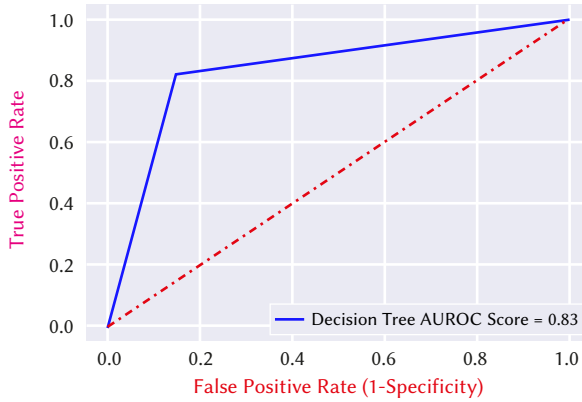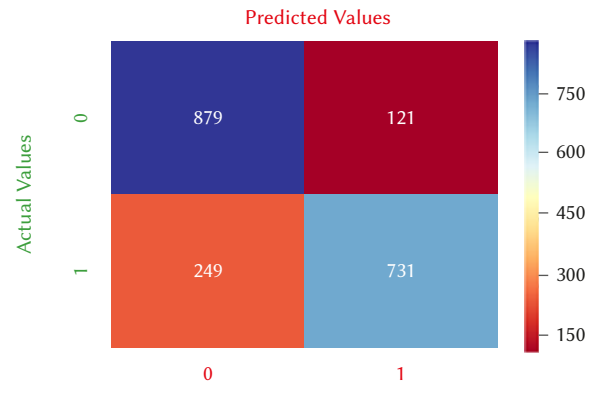Predicted Values

AUROC Curve For Decision Tree Model

Fig. 7. Confusion matrix and AUROC curve of Decision Tree Model.

Confusion Matrix of Decision tree on Test DataSet

Predicted Values

AUROC Curve For Random Forest Model

Fig. 8. Confusion matrix and AUROC curve of Random Forest Model.

Confusion Matrix of KNN Classifiers on Test DataSet

Predicted Values
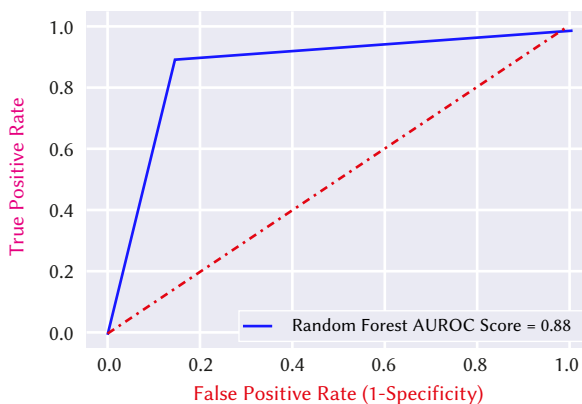

AUROC Curve For K-Nearest Neighbors Model

Fig. 9. Confusion matrix and AUROC curve of K-Nearest Neighbor Model.

The confusion matrix along with the AUROC curve of the K-Nearest Neighbors algorithm is shown in Fig. 9.

Then, the K-Nearest Neighbors algorithm was applied and found 81.31% of accuracy on the test dataset with 85.79% of precision, 74.59% of recall, and 79.80% of F1-Score. It means that the K-Nearest Neighbor model has the capability of differentiating the person having or not having heart disease is 81.31%. The confusion matrix along with the AUROC curve of the K-Nearest Neighbor model is shown in Fig. 9.

The confusion matrix along with the AUROC curve of the Support Vector Machine model is shown in Fig. 10.

Fig. 10 shows the confusion matrix along with the AUROC curve of the Support Vector Machine model. The obtained results of (SVM) are much better than KNN, in which the accuracy is 85.40% with precision 84.24%, Recall 86.73% and F1-Score is 85.47% respectively. It means that the Support Vector Machine model has the capability of differentiating the person having or not having heart disease is 85.40%.

The confusion matrix along with the AUROC curve of the Naïve Bayes model is shown in Fig. 11.

Finally, a Naïve Bayes classifier was applied and found 82.62% of accuracy without hyperparameter tuning. It means that the Naïve Bayes model has the capability of differentiating the person having or not having heart disease is 82.62%. The obtained precision of Naïve Bayes is 80.57%, with Recall and F1-Score 85.51% and 82.97% respectively. The confusion matrix along with AUROC is shown in Fig. 11. Overall results of all models are shown in Table III.
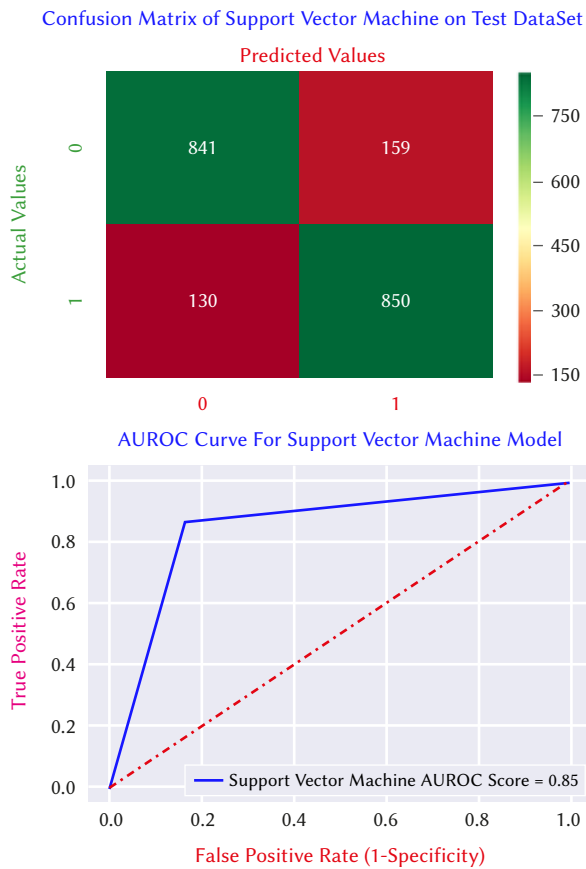
Confusion Matrix of Support Vector Machine on Test DataSet

Fig. 10. Confusion matrix and AUROC curve of Support Vector Machine Model.

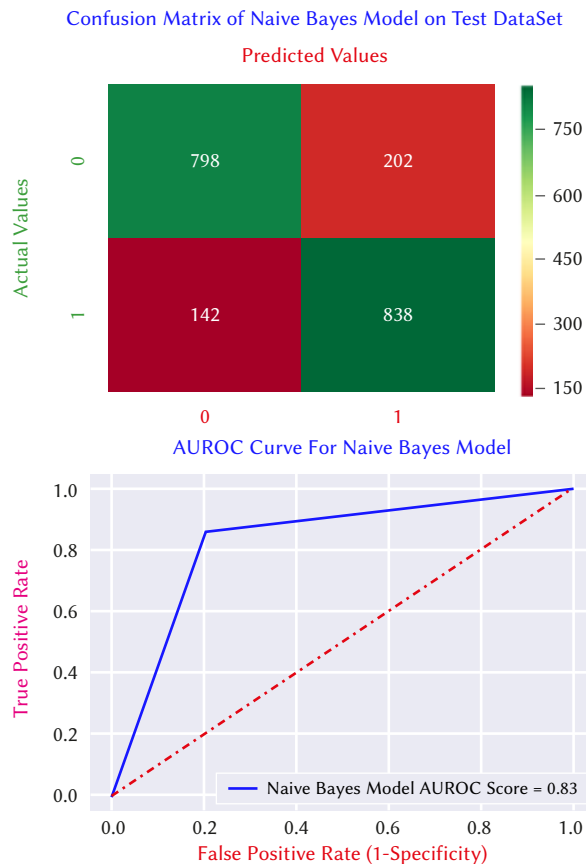**Confusion Matrix of Naive Bayes Model on Test DataSet**

Fig. 11. Confusion matrix and AUROC curve of Naïve Bayes Model.

For the refinement of results, hyperparameter tuning was applied which uses the Grid Search method of Scikit-Learn. The conventional approach of performing hyperparameter optimization has been Grid search, or a parameter sweep, which is simply a comprehensive search of candidate parameter values over all appropriate values in the specified search space [19]. After all possible parameter combinations which are calculated for a model; the finest combination will be retained. Grid search gives the training to an algorithm for all possible combinations using two hyperparameters sets (learning rate and several layers) and measures the performance using a cross-validation technique [21]. These validation techniques ensured that the trained model gets most of the patterns from the dataset. Grid Search uses a manual method to apply the parameter to the algorithm [23]. Naïve Bayes does not support Grid Search because it doesn't have any parameter. Hence, we applied Parameter Tuning on the other algorithms.

TABLE III. Performance Comparison of Different Proposed Heart Disease Prediction Models

Performance Metrics	Comparison of Heart Disease Prediction Models				
	Decision Tree	Random Forest	KNN	SVM	Naïve Bayes
AUROC	0.834163	0.877459	0.812459	0.854173	0.826551
F-1 Score	0.829876	0.878439	0.799035	0.854701	0.829703
Accuracy	0.834343	0.877273	0.813131	0.854040	0.826263
Recall	0.816327	0.895918	0.745918	0.867347	0.855102
Precision	0.843882	0.861629	0.857981	0.842418	0.805769
Error Rate	0.165657	0.122727	0.186869	0.145960	0.173737

From the above results, we came to the conclusion that without parameter tuning of Naïve Bayes performed better which gave 82.63% accuracy, and with hyperparameter tuning, the Random Forest search gave 87.72% accuracy which is higher than all other proposed classifiers. So, for this research work, the Random Forest model is the best for implementation and model building.

The experimental result is shown in Table III which shows that the Random Forest model outperforms other proposed models. So, for this research work the Hyperparameter Random Forest model is selected for making final outputs because of its higher accuracy and lesser error rates. The performance of the proposed model is tested and compared with existing models that portray very promising results.

VIII. Conclusion and Future Work

The main aim of this research was to develop an efficient cardiovascular disease prediction model for Jammu and Kashmir (India). Machine learning algorithms with IoT were utilized towards this aim. The prediction model is developed based on parameters. The parameters are selected after consulting with the domain experts. This research work came out to the conclusion that the Naïve Bayes algorithm performed better without hyper tuning of parameters while the Random Forest model is proven as an efficient technique with hyper parameter tuning. Because the Hyperparameter Random Forest model has higher accuracy and low error rates in comparison to all other selected models thus we select the Hyperparameterized Random Forest model for making the final output. After applying the machine learning classifiers to the heart disease dataset, it appears not astonishment that the complex classifiers like SVM and Random Forest brought better results with the loftiest accuracies of 0.8772% and 0.8540% as compared to the K-Nearest neighbor 0.8131%, Naïve Bayes 0.8262%, and Decision tree 0.8343% respectively. It is well worth emphasizing that in most instances hyper-parameter tuning is a must prerequisite to get sturdy results out of these classifiers. For future

improvements, our preference will be that further research should certainly be conducted to simulate this proposed model by adding new parameters with hybrid methods like ensemble techniques and more real-world datasets. These techniques can also be applied for predicting other fields like weather forecasting, election predictions, sales predictions, and other Bioinformatics predictions, etc.

Hence, we conclude this work by stating that the combination of IoT, Machine learning, and Cloud computing can be proven as a future reality for the prediction of diseases in general and cardiovascular diseases.

IX. Current and Future Developments

In this research work, a novel technique is developed for predicting cardiovascular disease for Jammu and Kashmir (India) based upon parameters. Machine learning algorithms with python language were utilized towards the objective of this research. While the results of this model are promising, if we add more parameters the accuracy of forecasting may increase. In this model, we had predicted cardiovascular disease for Jammu and Kashmir. In future, the further research will be forecasting the election outcomes for the whole of India's well. Currently, this model can predict cardiovascular disease only for Jammu and Kashmir but using this methodology researchers will extend this research work to other developing areas also by incorporating certain parameters pertaining to those areas. Further researchers can develop the new algorithm or combine all the algorithms using an ensemble technique in order to achieve better accuracy with low error rates.

References

[1] V. Fuster, B. B. Kelly, eds., "Promoting cardiovascular health in the developing world: a critical challenge to achieve global health," *National Academies Press*, 2010, pp 465.

[2] P. Kaur, R. Kumar, M. Kumar, "A healthcare monitoring system using random forest and internet of things (IoT)", 2019, pp. 19905-1916.

[3] I. Cvitić, D. Peraković, M. Periša, M. Botica, "Novel approach for detection of IoT generated DDoS traffic," Wireless Networks, 2019, pp. 1-14, doi: 10.1007/s11276-019-02043-1.

[4] A. Tewari, B.B. Gupta, "Security, privacy and trust of different layers in Internet-of-Things (IoTs) framework," Future generation computer systems", 2002, pp 909-920.

[5] I. Cvitic, D. Perakovic, M. Perisa, M. Botica, "Definition of the IoT device classes based on network traffic flow features", in 4th EAI International Conference on Management of Manufacturing Systems. Springer, Cham, 2020, pp 1-17.

[6] C. L. Stergiou, K.E. Psannis, et al. "IoT-based Big Data secure management in the Fog over a 6G Wireless Network", IEEE Internet of Things Journal, 2020, doi: 10.1109/JIOT.2020.3033131

[7] P.K. Senyo, E. Addae, R. Boateng, "Cloud computing research: "A review of research themes, frameworks, methods and future research directions", International Journal of Information Management, vol. 38, no. 1, pp 128-139, 2018.

[8] Y. Deepthi, Kalyan K.P., Vyas M., Radhika K., Babu D.K., Krishna Rao N.V. "Disease Prediction Based on Symptoms Using Machine Learning". In: Sikander A., Acharjee D., Chanda C., Mondal P., Verma P. (eds) Energy Systems, Drives and Automations. Lecture Notes in Electrical Engineering, vol. 664, Springer, 2020.

[9] F. Jabeen, M. Maqsood, M. A. Ghazanfar, F. Aadil, S. Khan, K. Kim, "An IoT based efficient hybrid recommender system for cardiovascular disease," Peer-to-Peer Networking and Applications, vol. 12, no. 5, pp 1263-1276, 2019.

[10] D. Palani, K. Venkatalakshmi, "An IoT Based Predictive Modelling for Predicting Lung Cancer Using Fuzzy Cluster Based Segmentation and Classification," Journal of medical systems, vol. 43, no. 2, pp. 1-12.

[11] A. Benjemmaa, H. Lti, M. Ben Ayed, "Design of Remote Heart Monitoring System for Cardiac Patients", In International Conference on Advanced Information Networking and Applications. Springer, Cham, 2019. pp.

963-976.

[12] Hashi, E. K., Zaman, M. S. U., Hasan, M. R. "An expert clinical decision support system to predict disease using classification techniques". International Conference on Electrical, Computer and Communication Engineering (ECCE), 2017. pp. 396-400.

[13] P. P. Chavda, "Early Detection of Cardiac Disease Using Machine Learning," In 2nd International Conference on Advances in Science & Technology (ICAST) 2019.http://dx.doi.org/10.2139/ssrn.3370813.

[14] K. Saravananathan, T. Velmurugan, "Analyzing Diabetic Data using Classification Algorithms in Data Mining," Indian Journal of Science and Technology, vol. 9, no. 43, pp 1-6, 2016.

[15] K.U. Sreekanth, K.P. Nitha., "A Study on Health Care in Internet of Things. International", Journal on Recent and Innovation Trends in Computing and Communication, vol. 4, no. 2, 2016.

[16] D. Dziak, B. Jachimczyk, W. Kulesza, "IoT-Based Information System for Healthcare Application: Design Methodology Approach," Applied Sciences, vol. 7, no. 6, pp. 596, 2017.

[17] L. Yehia, A. Khedr, and A. Darwish, "Hybrid Security Techniques for Internet of Things Healthcare Applications," Advances in Internet of Things, vol. 5, no. 3, pp. 21–25, 2015.

[18] G. Huang, K. Huang, T. Lee and J. Weng, "An interpretable rule-based diagnostic classification of diabetic nephropathy among type 2 diabetes patients", BMC Bioinformatics, vol. 16, no. 1, pp. 1-10, 2015.

[19] Iyer, A., Jeyalatha, S., Sumbaly, R. "Diagnosis of diabetes using classification mining techniques". arXiv preprint arXiv:1502.03774. 2015. DOI:10.5121/ijdkp.2015.5101.

[20] Md. Shahriar Hassan, Atiqur Rahman, Ahmed Wasif Reza, "Health status from your body to the cloud: The behavioral relationship between IoT and classification techniques in abnormal situations," Sensors for Health Monitoring, Elsevier 2019 Jan 1, pp. 135-156.

[21] S. M. Riazul Islam, M. HumaunKabir, M. Hossain, Daehan Kwak, Kyung-Sup Kwak, "The Internet of Things for Health Care: A Comprehensive Survey," IEEE Access, vol. 3, pp. 678–708, 2015.

[22] Hashi, E. K., Zaman, M. S. U., Hasan, M. R. "An expert clinical decision support system to predict disease using classification techniques". In 2017 International Conference on Electrical, Computer and Communication Engineering (ECCE), 2017, pp. 396-400.

[23] P. Patidar, A. Tiwari, "Handling Missing Value in Decision Tree Algorithm," International Journal of Computer Applications, vol. 70, no. 13, pp. 31-36, 2013.

[24] A. Naik, L. Samant, "Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime," Procedia Computer Science, vol. 85, pp. 662-668, 2016.

[25] Saravananathan, K., Velmurugan, T. "Analyzing Diabetic Data using Classification Algorithms in Data Mining," Indian Journal of Science and Technology, vol. 9, no. 43, pp. 1-6, 2016.

[26] A.M. Koli, M. Ahmed, "Machine Learning Based Parametric Estimation Approach for Poll Prediction", Recent Advances in Computer Science and Communications, vol. 14, no. 4, pp. 1287 - 1299, doi: 10.2174/266625 5813666191204112601.

[27] Thanh Noi, P., Kappas, M." Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel-2 imagery. Sensors", vol. 18, no. 1, pp. 18, 2018.

[28] Podgorelec V, Kokol P, Stiglic B, Rozman I. "Decision trees: an overview and their use in medicine", Journal of medical systems, vol. 26, pp 445-463, 2002.

[29] Elmachtoub, A., Liang, J. C. N., McNellis, R. "Decision trees for decision-making under the predict-then-optimize framework". In International Conference on Machine Learning 2020, pp. 2858-2867.

[30] Kumar, P. M., & Gandhi, U. D. "A novel three-tier Internet of Things architecture with machine learning algorithm for early detection of heart diseases," Computers & Electrical Engineering, vol. 65, pp. 222-235, 2018.

[31] Chen PT, Lin CL, Wu WN. "Big data management in healthcare: Adoption challenges and implications", International Journal of Information Management, vol. 53, 102078, 2020.

[32] Islam M. S., Islam M. T., Almutairi A. F., Beng G. K., Misran N., Amin N., "Monitoring of the human body signal through the Internet of Things (IoT) based LoRa wireless network system," Applied Sciences, vol. 9, pp. 1884, 2019.

[33] Dejene, D., Tiwari, B., Tiwari, V. "TD2SecIoT: Temporal, Data-Driven and Dynamic Network Layer Based Security Architecture for Industrial IoT," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 6, no. 4, pp. 146-156, 2020, doi: 10.9781/ijimai.2020.10.002.

[34] Harish, B. S., Roopa, C. K. "Automated ECG Analysis for Localizing Thrombus in Culprit Artery Using Rule Based Information Fuzzy Network," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 6, no. 1, pp. 16-25, 2020, doi: 10.9781/ijimai.2019.02.001.

Jameel Ahamed

Jameel Ahamed is a student of Doctorate of Philosophy in Department of Computer Science and Engineering, National Institute of Technology, Srinagar, J&K. He is also working as an Assistant Professor in the Department of CS and IT, Maulana Azad National Urdu University, Hyderabad-Telangana with over six years of experience in academic and research. His current research interests include the internet of things, wireless sensor networks, and computer networks.
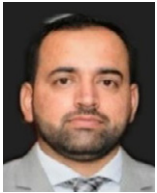
Abdul Manan Koli

Abdul Manan Koli completed his BSC from the University of Jammu and MCA from the University of Jammu. He is currently pursuing PhD from the department of CS&IT, MANUU, Hyderabad. He has worked on numerous projects in the field of machine learning and Data Mining. He has published several conference and journal papers in national and international journals. His areas of research include Political forecasting, Machine Learning, and Big Data Analytics.

Khaleel Ahmad

Khaleel Ahmad is currently an Assistant Professor in the Department of Computer Science and Information Technology, Maulana Azad National Urdu University, Hyderabad, India. He has visited the National Defence University of Malaysia, Malaysia as a visiting faculty. He holds a Ph.D. in Computer Science & Engineering and M.Tech in Information Security. His research areas are Blockchain Technology, Cyber Security, Cryptography, and Opportunistic Network. He has 30 published papers in refereed national/international journals and conferences (viz. Elsevier, ACM, IEEE, Springer, InderScience, CMC Journal-Tech Science Press, Bentham Science), 10 book chapters (Springer, CRC Press, IGI Global, Wiley). He has edited two books for Taylor & Francis titled "Opportunistic Networks: Mobility Models, Protocols, Security & Privacy" and "Emerging Security Algorithms and Techniques" and one book for Springer titled "Functional Encryption". He has completed a successful project of Rs. 1 Lakh and Five Thousand (INR). He has supervised 4 M.Tech and 4 Ph.D. in progress. He has delivered guest lectures at the Central University of Haryana, Telangana University, and JNTU Hyderabad and also chaired the session at an international conference in Malaysia and India. He has also served as Guest Editor of the International Journal of Innovative Computing and Applications (InderScience), Journal of Recent Patents on Engineering, Bentham Science and Scalable Computing: Practice and Experience (SCPE).

Mohammad Alam Jamal

Mr. Mohammad Alam Jamal completed his B. Tech and M. Tech. from Maulana Azad National Urdu University. His areas of interest include Machine Learning, the Internet of Things, and Deep Learning.

Brij B. Gupta

Prof Brij B. Gupta received the PhD degree from Indian Institute of Technology (IIT) Roorkee, India. In more than 16 years of his professional experience, he published over 400 papers in journals/conferences including 30 books and 08 Patents with over 14000 citations. He has received numerous national and international awards including Canadian Commonwealth Scholarship (2009), Faculty Research Fellowship Award (2017), from the Govt. of Canada, MeitY, GoI, IEEE GCCE outstanding and WIE paper awards and Best Faculty Award (2018 & 2019), NIT Kurukshetra, respectively. Prof Gupta is also serving as Distinguished Research Scientist with LoginRadius Inc., USA which is one of leading cybersecurity companies in the world, especially in the field of customer identity and access management (CIAM). He is also selected in the 2021 and 2020 Stanford University's ranking of the world's top 2% scientists. He is also a visiting/adjunct professor with several universities worldwide. He is also an IEEE Senior Member (2017) and also selected as 2021 Distinguished Lecturer in IEEE CTSoc. Dr Gupta is also serving as Member-in-Large, Board of Governors, IEEE Consumer Technology Society (2022-204). Prof. Gupta is also leading IJSWIS, IJSSCI and IJCAC, IGI Global, as Editor-in-Chief. Moreover, he is also serving as lead-editor of a Book Series with CRC, World Scientific and IET press. He also served as TPC members and organized/special session chairs in ICCE-2021, GCCE 2014-2021 and TPC Chair in 2018 INFOCOM: CCSNA Workshop and Publicity Co-chair in 2020 ICCCN. Dr Gupta is also serving/served as Associate/Guest Editor of IEEE TII, IEEE TITS, IoT, IEEE Big Data, ASOC, FGCS, etc. At present, Prof. Gupta is working as Director, International Center for AI and Cyber Security Research and Innovations, and Full Professor with the Department of Computer Science and Information Engineering (CSIE), Asia University, Taiwan. His research interests include information security, Cyber physical systems, cloud computing, blockchain technologies, intrusion detection, AI, social media and networking.