International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Predicting Heart Disease Risk Using an Ensemble AdaBoost Supervised Machine Learning Classifier

Hardik J. Prajapati[1], Dr. Dushyantsinh B. Rathod[2]

Submitted: 26/01/2024 **Revised:** 04/03/2024 **Accepted:** 20/03/2024

Abstract*:* The heart is a crucial component of all living beings. The diagnosis and prognosis of heart illness need enhanced completeness and accuracy, since even a little error may result in severe complications or loss of life; heart-related fatalities are many and increasing quickly each day. A system that can forecast the spread of diseases is essential for finding a solution to this issue. An example of artificial intelligence is machine learning. When it comes to predicting the outcomes of all kinds of natural catastrophes, it offers exceptional help. Using the UCI benchmark data sets for training and testing, we determine the accuracy of four machine learning algorithms—k-proximal neighbours, Naïve Bayes, voting classifier, and ADABOOST—in predicting the occurrence of heart disease. Because it comes with a wide variety of libraries and header files, the Anaconda (Jupyter) notebook is the greatest tool for implementing Python programming.

Keywords: *supervised, confusion matrix, linear regression, unsupervised, python, reinforced*

1. Introduction

One of the largest and most vital organs in the body, the heart need particular care. Most illnesses have some connection to the heart, thus it's important to be well-versed on the most relevant information for disease prediction. For this reason, it is crucial to do comparative study on this subject. It is crucial to comprehend the most useful data for illness prediction because, in the present day, the majority of patients die because their illnesses are only identified near the deadline due to the instrument's lack of accuracy.Machine learning, which relies on training and testing, is among the most effective testing approaches. [4] Machine learning is a branch of AI, which encompasses a wide range of capabilities that computers may acquire to imitate human intelligence. However, machine learning algorithms are trained to understand and make use of data, which is how the phrase "artificial intelligence" came to be used to describe the merging of these two technologies. In this project, we utilised physiological parameters as test data, including lipids, heart rate, biological sex, age, and so on. We aimed to compare the algorithms' accuracy using these parameters, as machine learning is defined as learning from natural phenomena and things. Our three algorithms used in this project were Naive Bayes, KNN, and LOR. This essay begins with a brief introduction to machine learning and cardiac issues. In Section 2, we go over the Data Mining Algorithm. A literature review constitutes the third portion. In the fourth section, we go over the proposed architecture. In Section 5, the qualities and dataset of this project are summarised. A quick look at the document's potential future scope concludes this document's overview.

1 PhD Scholar , Faculty of Engineering and Technology, Sankalchand Patel University, Visnagar, Gujarat
2 Professor & Head, Ahmedabad Institute of Technology, Gota, Ahmedabad, Gujarat
** Corresponding Author Email: hardikjp2707@gmail.com*

2. Data Mining Algorithm

There are so many Data mining algorithms, But following algorithms are referred for research study.

2.1. Naive Bayes (NB)

One method of machine learning is Naive Bayes, and it is used for classification problems. Bayes' theory of probability forms its basis. Its primary use is text classification using massive training datasets. Example tasks involve emotion detection, spam filtering, and news item categorisation. Its efficiency has made it famous. Quick model building and prediction are both made possible by the Naive Bayes approach. This is the first approach to classification of texts that has been contemplated.

2.2. K Nearest Neighbor (KNN)

Machine learning's k-nearest neighbours (KNN) method is simple. In this post, we will go over the basics of the kNN method and how to use R for kNN modelling. The dataset must be generated before the knn() function can be used in A. After making a future prediction using the kNN method, you must assess the model's diagnostic performance. When describing the kNN algorithm, average precision is the most common metric to employ. There are many elements that affect the success of the model. These include the k-value, the distance estimation, and the choice of suitable predictors.

.

2.3. ADABOOST

AdaBoost approach used as an Ensemble Method in Machine Learning is the AdaBoost algorithm, which stands for Adaptive

Boosting. Adaptive Boosting is the process of reassigning weights to each instance, with larger weights given to instances that were mistakenly categorised. Boosting is a tool for supervised learning that helps decrease bias and variation. Learners progress in a sequential manner, which is the basis of its operation. With the exception of the first, all succeeding learners are offspring of learners that came before them. Basically, it's a way to turn poor learners into strong ones. Similar to boosting but slightly modified, the AdaBoost algorithm achieves its goals.

2.4. Voting Classifier

The Voting Classifier is an example of a machine learning model that trains a group of models to make a prediction about a class by adding up the probability that each model would choose that class. All it does is compiles the results from each classifier that the Voting Classifier has received and utilises the most heavily voted one to predict the output class.

3. Literature survey

Varun Kumar et al.[1] proposed his architecture with an accuracy of up to 85%, the convolutional neural network method analyses the peril of early heart disease using structured data. In addition, photos and unstructured data can be handled using the CNN technique.

P. K. Gupta et al.[2] tested several machine learning techniques to detect heart diseases using the dataset. Using modified random forest, they achieved 86.84% accuracy. This approach works well in real time, and adding additional data with CNN and deep learning algorithms might improve its accuracy.

Brahami et al.[3] used healthcare dataset categorization, a machine learning priority. We explored Logistic Regression, Adaptive Boosting, and Multi-Objective Evolutionary Fuzzy Classifier. All individually, Majority Voting is 80.20% accurate, Logistic Regression the lowest, and AdaBoostM1 the highest.

Rakesh Kumar et al.[4], in his study, the accuracy for the four different machine learning algorithms was examined; KNN delivered the best outcome, with an accuracy of 87%.

Shamsheela Habib et al.[5], in their current study, we offer a novel machine learning technique as the foundation for our advanced cardiac disease prediction approach. Finding correlation-based features that improve prediction accuracy is its main objective. We utilise the UCI Vascular Cardiovascular Disease Dataset in our work, and we juxtapose our results with those of an earlier investigation. The accuracy of the model we recommended was 85.43%.

Xu Wenxin et al.[6] developed an innovative solution to predict cardiac disease utilizing SVM, decision tree, and ANN models with 87% accuracy.

Shaicy Shaji et al.[7] proposed scheme aims to diagnose various cardiac diseases and to swiftly set preventive measures into place at an affordable cost. The method employs data mining addresses to forecast cardiac problems by feeding information into the Random Forest, SVM, and KNN classification methods. While KNN obtained an accuracy of 83 percent, SVM & random forest model both reached an accuracy of 85 percent.

Saumya Yadav st al.[8] proposed Future potentials can be projected using the prediction analysis technique using the current data collection. For prediction analysis, an earlier SVM classifier is employed in this work. Because the KNN classifier uses the same number of hyper planes as the number of classes, it has a greater accuracy of 83 percent than the SVM classifier.

4. Proposed Architecture

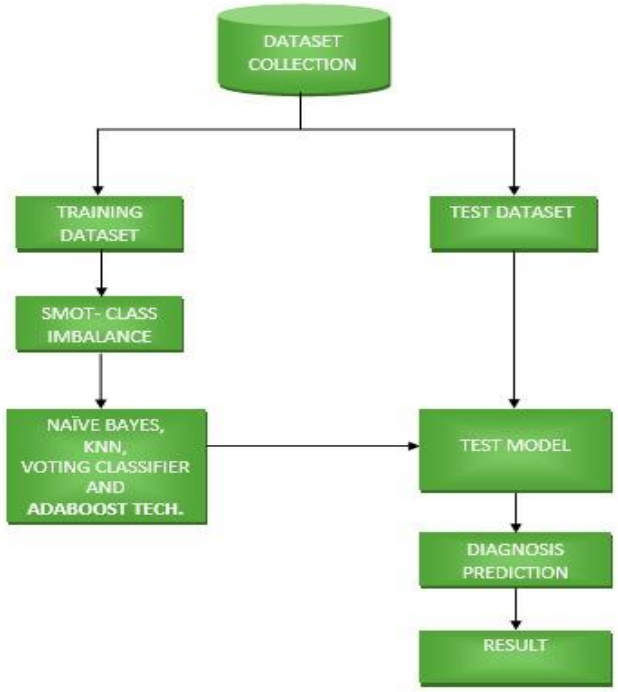

Fig 1. Proposed Architecture

The algorithm in the figure represents a machine learning workflow for diagnosing a condition based on collected datasets. Here's a breakdown of each step:

1. **Dataset Collection**: Data is gathered, likely including both coronary heart disease and lung cancer data, as per your research topic.
2. **Data Splitting**:
    - The collected dataset is split into a **Training Dataset** and a **Test Dataset**.
3. **Training Dataset**:
    - **SMOTE - Class Imbalance**: Synthetic Minority Over-sampling Technique (SMOTE) is used to address class imbalance in the training data. This is especially helpful if the dataset has an uneven distribution between classes (e.g., positive vs. negative diagnoses).
    - **Naïve Bayes, KNN, Voting Classifier, and AdaBoost Techniques**: These algorithms are applied to the balanced training dataset.
        - **Naïve Bayes**: A probabilistic classifier based on Bayes' theorem.
        - **K-Nearest Neighbors (KNN)**: A non-parametric method that classifies data points based on their proximity to neighbors.

- **Voting Classifier**: Combines predictions from multiple models to improve accuracy.
- **AdaBoost**: A boosting algorithm that builds a strong classifier by combining weak classifiers.

4. **Test Dataset**:
   - The test dataset is used to evaluate the trained model.
5. **Model Testing and Prediction**:
   - **Test Model**: The trained model is tested on the test dataset.
   - **Diagnosis Prediction**: The model generates predictions for diagnosis.
   - **Result**: The final diagnostic result is produced based on the predictions.

This flowchart represents a stacked ensemble approach, where multiple classification algorithms are combined to enhance model performance, which aligns with your research focus.

I have applied KNN and AdaBoost on this dataset.

```
from sklearn.neighbors import KNeighborsClassifier
KNN = KNeighborsClassifier(n_neighbors=15)
KNN.fit(x_res,y_res)
KNN.score(x_test,y_test)
```

```
0.6824034334763949
```

```
from sklearn.metrics import classification_report
KNN_Pred=KNN.predict(x_test)
KNNreport = classification_report(y_test, KNN_Pred)
print(KNNreport)
```

```
              precision    recall  f1-score   support

           0       0.66      0.70      0.68       113
           1       0.70      0.67      0.68       120

    accuracy                           0.68       233
   macro avg       0.68      0.68      0.68       233
weighted avg       0.68      0.68      0.68       233
```

Fig 2. Report of KNN (Recall, Precision, F1-score)

Figure 4 displays the results of the KNN classification. With an accuracy of 0.68, the KNN method was used. Both its recall and f1-score are 0.68.

```
from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay
from sklearn.metrics import confusion_matrix
print("confusion matrix for KNN")
cm = confusion_matrix(y_test, KNN_Pred)
disp = ConfusionMatrixDisplay(confusion_matrix=cm)
disp.plot()
```

```
confusion matrix for KNN
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x7920a3bc1fc0>
```

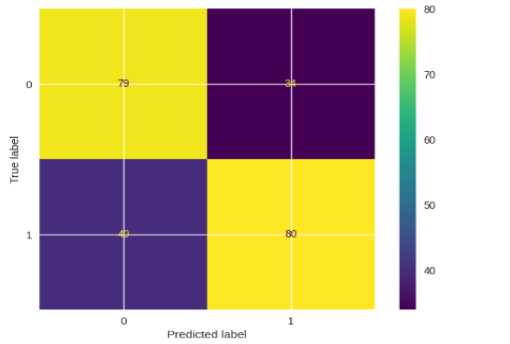

Fig 3. KNN Confusion Matrix

Figure 3 depicts the KNN confusion matrix. The function returns the confusion matrix along with the names of the groups. One way to see the confusion matrix is via the heat map tool.

```
from sklearn.metrics import classification_report
ADB_Pred=ADBclassifier.predict(x_test)
ADBreport = classification_report(y_test, ADB_Pred)
print(ADBreport)
```

```
              precision    recall  f1-score   support

           0       0.91      0.88      0.89        48
           1       0.88      0.91      0.90        47

    accuracy                           0.89        95
   macro avg       0.90      0.89      0.89        95
weighted avg       0.90      0.89      0.89        95
```

Fig 4. Report of ADABOOST Classification (Recall, Precision, F1-score)

Figure 4 illustrates illustrates ADABOOST classification report. ADABOOST algorithm provided the accuracy 0.89. Its recall value is 0.89 and f1-score is 0.89.

```
confusion matrix for ADABoost classifier
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay a
```

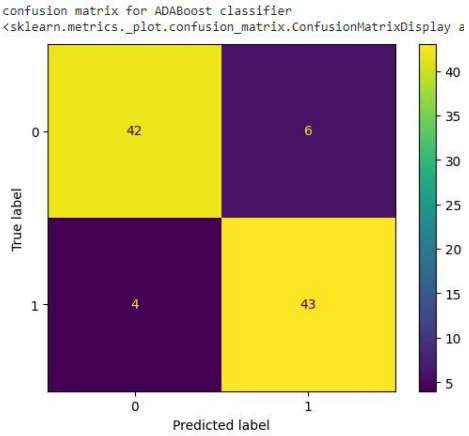

Fig 5. ADABOOST Confusion Matrix

Figure 5 illustrates ADABOOST confusion matrix. The confusion matrix and group names are returned by the function. The heatmap function may be used to visualise the confusion matrix.

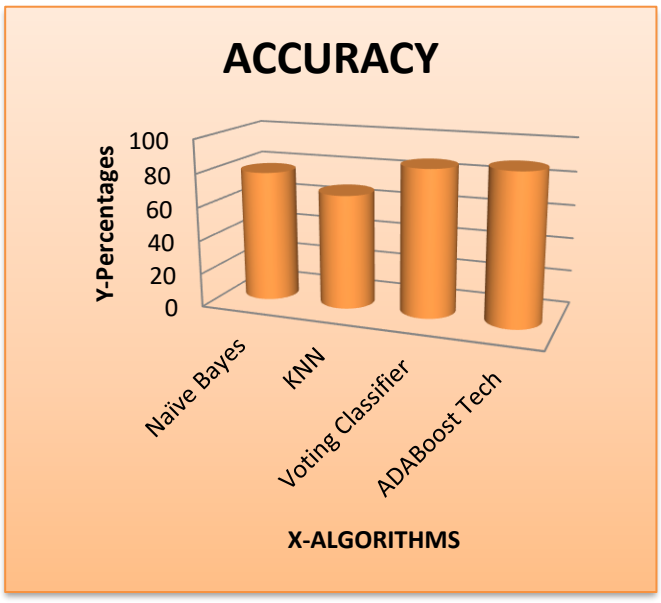

Fig 6. Accuracy Chart of ML algorithms

Table 2. Accuracy Table of ML algorithms

Algorithms	*Accuracy %*
Naïve Bayes	78
KNN	68
VOTING CLASSIFIER	87
ADABOOST CLASSIFIER	89

Figure 6 illustrates Accuracy chart of ML algorithms. Naive Bayes provides 78% accuracy, KNN provides 84% accuracy, Voting Classifier provides 84% accuracy and ADABOOST provides 89% accuracy.

5. Dataset & Model

In this paper, Dataset compiled with characteristics Sex denotes the patient's gender, age signifies the patient's age, trestbps represents the resting blood pressure, cp refers to chest discomfort, and fbs indicates fasting blood sugar, chol signifies cholesterol, thalach is the greatest heart rate attained, whereas restecg signifies the Resting electrocardiogram results suggest one anomaly; oldpeak reflects the ST segment. Depression caused; exang implies exercise-induced conditions. Angina and CA denote the quantity of significant vessels, whereas slope signifies the gradient of maximal exertion ST, pred_attribute, and thal signify the Thalassaemia. The gathered data sample is shown below.

The suggested process has the following benefits. Implemented two machine learning algorithms and a hybrid model. The accuracy of all suggested algorithms was evaluated to identify the optimal model. Implement a hybrid model to optimise the requested task. The execution is conducted using the approaches outlined below.

a. The dataset is sourced from uci.edu.
b. Data visualisation is conducted.
c. The dataset is partitioned into training and testing subsets.
d. NB, KNN, Voting Classifier and AdaBoost models are used for training and analysis.
e. The model is trained.
f. The trained model is evaluated and predictions are made.
g. Obtain a singular input from the user and forecast cardiac disease via a hybrid model.

The Cleveland dataset is taken into account. It is divided into two components: training and testing sets. We allocated 70% of the dataset as training input for the machine learning methods and trained the model. The remaining 30% serves as testing data for heart disease prediction. We KNN, NB, Voting classifier & AdaBoosting. The AdaBoost algorithm is used to predict heart disease using 30% of the test input, with the predicted values then displayed and assessed for accuracy.

6. Conclusion & Future Work

Cardiovascular disease is a prevalent life-threatening condition globally. The evolving lifestyle and insufficient physical activity provide a greater risk to health conditions. Numerous diagnostic techniques exist within the medical business. Nonetheless, for precision, machine learning is regarded as the optimal option. The suggested project uses a Jupiter Python application for predicting cardiac disease. The suggested approach employs a model of AdaBoost Algorithm for heart disease prediction. The Cleveland database is used for this investigation.

Important roles played by deep learning algorithms in healthcare applications. Therefore, it is possible to get better results when using deep learning algorithms to forecast cardiac disease. Furthermore, we are keen on recognizing it as a multi-class condition in order to determine the severity of the sickness.

7. References

Author contributions

Hardik Prajapati: Role of Primary Author (a) Feasibility Study (b) Requirement Analysis from Health Stakeholder (c) Requirement Gathering from Hospital (d) Planning (e) Designing (f) Implementation / Coding [Python / Google Colab] (g) Testing/Validation

Dr. Dushyantsinh Rathod: Guidance about research problem, Documentations, Deadline Maintenance

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Shankar, V., Kumar, V., Devagade, U. *et al.* Heart Disease Prediction Using CNN Algorithm. *SN COMPUT. SCI.* **1**, 170 (2020). https://doi.org/10.1007/s42979-020-0097-6.

[2] Vinayaka, S., Gupta, P.K. (2020). Heart Disease Prediction System Using Classification Algorithms. In: Singh, M., Gupta, P., Tyagi, V., Flusser, J., Ören, T., Valentino, G. (eds) Advances in Computing and Data Sciences. ICACDS 2020. Communications in Computer and Information Science, vol 1244. Springer, Singapore. https://doi.org/10.1007/978-981-15-6634-9_36.

[3] Abdeldjouad, F.Z., Brahami, M., Matta, N. (2020). A Hybrid Approach for Heart Disease Diagnosis and Prediction Using Machine Learning Techniques. In: Jmaiel, M., Mokhtari, M., Abdulrazak, B., Aloulou, H., Kallel, S. (eds) The Impact of Digital Technologies on Public Health in Developed and Developing Countries. ICOST 2020. Lecture Notes in Computer Science(), vol 12157. Springer, Cham. https://doi.org/10.1007/978-3-030-51517-1_26 Rakesh A. Singh and R. Kumar, "Heart Disease Prediction Using Machine Learning Algorithms," 2020 International Conference on Electrical and Electronics Engineering (ICE3), Gorakhpur, India, 2020, pp. 452-457, doi: 10.1109/ICE348803.2020.9122958.

[4] Rakesh A. Singh and R. Kumar, "Heart Disease Prediction Using Machine Learning Algorithms," 2020 International Conference on Electrical and Electronics Engineering (ICE3), Gorakhpur, India, 2020, pp. 452-457, doi: 10.1109/ICE348803.2020.9122958.

[5] M. A. Alim, S. Habib, Y. Farooq and A. Rafay, "Robust Heart Disease Prediction: A Novel Approach based on Significant Feature and Ensemble learning Model," 2020 3rd International Conference on

Computing, Mathematics and Engineering Technologies (iCoMET), Sukkur, Pakistan, 2020, pp. 1-5, doi: 10.1109/iCoMET48670.2020.9074135. keywords: {Diseases;Heart;Forestry;Machine learning algorithms;Vegetation;Machine learning;Correlation;Machine Learning;Stratified KFold;Random Forest and ROC}

[6] Wenxin, Xu. (2020). Heart Disease Prediction Model Based on Model Ensemble. 195-199. 10.1109/ICAIBD49809.2020.9137483 .

[7] Alex P, M., & Shaji, S.P. (2019). Predictionand Diagnosis of Heart Disease Patients using Data Mining Technique. 2019 International Conference on Communication and Signal Processing (ICCSP), 0848-0852.

[8] Chakarverti, Mohini & Yadav, Saumya & Rajan, Rajiv. (2019). Classification Technique for Heart Disease Prediction in Data Mining. 1578-1582. 10.1109/ICICICT46008.2019.8993191.

[9] Kumar, Abhishek & Kumar, Pardeep & Srivastava, Ashutosh & V D, Ambeth Kumar & Vengatesan, K. & Singhal, Achintya. (2020). Comparative Analysis of Data Mining Techniques to Predict Heart Disease for Diabetic Patients. 10.1007/978-981-15-6634-9_46.

[10] Harika, Navya & Swamy, Sita & Nilima, Nilima. (2021). Artificial Intelligence-Based Ensemble Model for Rapid Prediction of Heart Disease. SN Computer Science. 2. 10.1007/s42979-021-00829-9.

[11] Joshi, S., Nair, M.K. (2021). A Risk Assessment Model for Patients Suffering from Coronary Heart Disease Using a Novel Feature Selection Algorithm and Learning Classifiers. In: Chiplunkar, N.N., Fukao, T. (eds) Advances in Artificial Intelligence and Data Engineering. AIDE 2019. Advances in Intelligent Systems and Computing, vol 1133. Springer, Singapore. https://doi.org/10.1007/978-981-15-3514-7_20

[12] Prakash, Jothi & Karthikeyan, N.. (2021). Enhanced Evolutionary Feature Selection and Ensemble Method for Cardiovascular Disease Prediction. Interdisciplinary Sciences: Computational Life Sciences. 13. 10.1007/s12539-021-00430-x.

[13] Alim, Muhammad & Habib, Shamsheela & Farooq, Yumna & Jungsher, Abdul. (2020). Robust Heart Disease Prediction: A Novel Approach based on Significant Feature and Ensemble learning Model. 1-5. 10.1109/iCoMET48670.2020.9074135.

[14] Kumar, Abhishek & Kumar, Pardeep & Srivastava, Ashutosh & V D, Ambeth Kumar & Vengatesan, K. & Singhal, Achintya. (2020). Comparative Analysis of Data Mining Techniques to Predict Heart Disease for Diabetic Patients. 10.1007/978-981-15-6634-9_46.

[15] Anbarasi, M., Anupriya, E., & Iyengar, N.C. (2010). ENHANCED PREDICTION OF HEART DISEASE WITH FEATURE SUBSET SELECTION USING GENETIC ALGORITHM.

[16] Singh, Poornima & Singh, Sanjay & Pandi Jain, Gayatri. (2018). Effective heart disease prediction system using data mining techniques. International Journal of Nanomedicine. 13. 121-124. 10.2147/IJN.S124998.

[17] Kamaraj, K.Gomathi & Priyaa, D.Shanmuga. (2016). Multi Disease Prediction using Data Mining Techniques. International Journal of System and Software Engineering.

[18] Miranda, Eka & Irwansyah, Edy & Amelga, Alowisius & Maribondang, Marco & Salim, Mulyadi. (2016). Detection of Cardiovascular Disease Risk's Level for Adults Using Naive Bayes Classifier. Healthcare Informatics Research. 22. 196. 10.4258/hir.2016.22.3.196.

[19] Agarap, Abien Fred. (2018). Deep Learning using Rectified Linear Units (ReLU). 10.48550/arXiv.1803.08375.

[20] Masilamani, Anbarasi & ANUPRIYA, & Iyenger, N Ch Sriman Narayana. (2010). Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm. International Journal of Engineering Science and Technology. 2.

[21] Cherian, V., Bindu, M.S.: Heart disease prediction using Na¨ıve Bayes algorithm and laplace smoothing technique. Int. J. Comput. Sci. Trends Technol. 5(2), 68–73(2017)

[22] Dulhare, U.N.: Prediction system for heart disease using Naive Bayes and particle swarm optimization. Biomed. Res. (India) (2018). https://doi.org/10.4066/biomedicalresearch.29-18-620

[23] Enriko, I.K.A., Suryanegara, M., Gunawan, D.: Heart disease prediction system using k-Nearest neighbor algorithm with simplified patient's health parameters. J. Telecommun. Electron. Comput. Eng. 8(12), 59–65 (2016)

[24] Gupta, P., Maharaj, B.T., Malekian, R.: A novel and secure iot based cloud centric architecture to perform predictive analysis of users activities in sustainable healthcentres. Multimed. Tools Appl. 76(18), 18489–18512 (2017)

[25] Gupta, P., Tyagi, V., Singh, S.: Predictive Computing and Information Security.Springer, Heidelberg (2017). https://doi.org/10.1007/978-981-10-5107-4