Save Nature to Survive The Bioscan 19(1): 39-43, 2024
AN INTERNATIONAL QUARTERLY JOURNAL OF LIFE SCIENCES
www.thebioscan.com

# COMPARATIVE STUDY OF MACHINE LEARNING ALGORITHMS FOR DIABETES RISK PREDICTION: A FOCUS ON K-NN, SVM, DECISION TREE, AND RANDOM FOREST.

**S ADITI APURVA***
IIIT Ranchi, Ranchi - 834004
e-mail: sapurva.2023dr103@iiitranchi.ac.in

**KEYWORDS**
Machine Learning
Diabetes
Support Vector Machine
k-Nearest Neighbours

**Received on :**
01.01.2024

**Accepted on :**
26.03.2024

***Corresponding author**

## ABSTRACT

The global health situation is significant due to diabetes mellitus, a chronic metabolic disorder. One of the chronic diseases known as chronic metabolic ailment is caused by persistently elevated blood sugar levels. It is believed to be among the deadliest illnesses worldwide. If an accurate early prognosis is available, the severity and risk factors of diabetes can be greatly reduced. The early diagnosis of diabetes can be aided by algorithms for machine learning. Early identification can help diabetes patients reduce their health risks. The results can be beneficial for doctors, patients, and family members of patients. It is essential to estimate the patient's state upon entry in order to allocate resources correctly in healthcare settings with limited resources. Medical diagnosis accuracy is increased and costs are decreased using machine learning approaches. This research paper presents a comprehensive comparative study of machine learning algorithms, namely k-Nearest Neighbours (k-NN), Support Vector Machine (SVM), Decision Tree, and Random Forest, to identify the most effective model for diabetes risk prediction. All four algorithms' results are assessed using a range of metrics, including recall, F-measure, accuracy, and precision. The number of correctly and wrongly classified cases is used to calculate accuracy. The results demonstrate that when compared to other algorithms, Random Forest performs with the greatest accuracy of 99.03%. preceded by Decision Tree with an accuracy of 95% preceded by SVM with an accuracy of 90% and K-Nearest Neighbour with an accuracy of 89%.

## INTRODUCTION

Millions of people worldwide suffer with diabetes mellitus, a common and chronic illness. According to the International Diabetes Federation (IDF), diabetes mellitus had an overall prevalence of 366 million in 2011 and was predicted to increase to 552 million by 2030. The glucose level in an critically ill patient must be maintained at 140–180 mg/dL (7.8–10.0 mmol/L) via continuous intravenous insulin infusion (Alam,2014)A key component of efficient management and preventive treatment is early detection and risk prediction. We provide a thorough comparative examination of four machine learning algorithms in this study: Random Forest, Decision Tree, Support Vector Machines (SVM), and k-Nearest Neighbours (k-NN), with an emphasis on how well they predict the risk of diabetes. For model training and assessment, a broad dataset of demographic, clinical, and lifestyle variables is used (Kavakiotis,2017; Swapna,2018; Deepti,2018; Sneha,2019; Pujianto,2019; Abdulhadi,2021;). SVM seeks to identify the best hyperplane for classification, whereas k-NN analyses patterns in feature space by taking instances' closeness into account. In contrast, Decision Tree and Random Forest utilize hierarchical tree structures to effectively represent the relationships present in the data. Accuracy, precision, recall, and F1-score are among the evaluation measures that offer a thorough grasp of each algorithm's predictive power. In order to shed light on the variables influencing diabetes risk prediction, the study additionally examines the models' interpretability and feature importance. The results show subtle differences in the algorithms' performance: SVM performs well in high-dimensional feature spaces, whereas k-NN performs better in some situations since it is sensitive to local patterns. Interpretability is aided by decision trees and random forests, which demonstrate the capacity to capture intricate relationships and offer insights into feature relevance.

In order to find the best configurations for increased predictive accuracy, the study also investigates how hyperparameter adjustment affects the algorithms' performance. The results add to the expanding corpus of research on diabetes risk prediction and provide physicians and researchers with useful information for choosing appropriate machine learning models depending on certain data attributes and goals.

As machine learning advances, this comparative study offers a useful manual for utilizing several algorithms in diabetes risk assessment, supporting a data-driven strategy for early intervention and tailored healthcare.

### Support Vector Machine

It is a type of supervised learning which is applied to regression and classification problems. SVMs can capture intricate correlations in the data and are especially useful in high-dimensional spaces (Cortes,1995; ilvanciuc,2005; https://scikit-learn.org/stable/modules/svm.html).

K-Nearest Neighbours: Sophisticated and reliable, the K-Nearest Neighbours (KNN) algorithm is a machine learning technique used to solve regression and classification issues. KNN uses its K nearest neighbours in the training dataset to predict the label or value of a new data point by leveraging the similarity notion (Friedman,1975; https://www.geeksforgeeks.org/k-nearest-neighbours/).

**Random Forest**

It is an ensemble learning algorithm, has emerged as a powerful and versatile tool in the realm of machine learning. Comprising a collection of decision trees, this algorithm excels in both classification and regression tasks, demonstrating resilience to overfitting and a remarkable capacity to handle diverse datasets (Breiman,2001;Benbelkacem ,2019;https://towardsdatascience.com/understanding-random-forest-58381e0602d2).

Decision Tree: It is a versatile and interpretable machine learning models widely employed in both classification and regression tasks. This paper provides an overview of decision trees, emphasizing their fundamental concepts, construction, and applications. Decision trees recursively partition the input space based on feature attributes, forming a tree-like structure where each node represents a decision point. The splitting process is guided by metrics such as Gini impurity or information gain, aiming to maximize homogeneity within resulting subsets (Argentiero,1982;J arullah,2011; Vijay an,2015; https://www.geeksforgeeks.org/decision-tree/).

## MATERIALS AND METHODS

### Data collection and data pre-processing

Patients' medical, demographic, and diabetes status—positive or negative—are all included in the diabetes_pre diction_dataset .csv file (https://www.kaggle.com/code/therealsampat/early-stage-diabetes-prediction). It includes a number of variables, including blood glucose level, age, gender, body mass index (BMI), blood pressure, heart disease, smoking history, and HbA1c level. Using the Dataset, machine learning models that predict a patient's risk of developing diabetes based on their medical history and demographic information can be built [4]. Split the dataset into training, validation, and testing sets to assess the model's performance. Handle missing values, outliers, and duplicates. Normalize or standardize numerical features. Encode categorical variables. Visualization and analysis of the distribution of features. Identification of the correlations and patterns between the features. Identification and selection of relevant features that contributes to diabetes risk prediction

Table 1: Features in dataset

| Sl.No | Features |
|---|---|
| 1 | Gender |
| 2 | Age |
| 3 | Hypertension |
| 4 | Heart Disease |
| 5 | Smoking Habit |
| 6 | BMI |
| 7 | HbA1c |
| 8 | Blood Sugar |
| 9 | Diabetes |

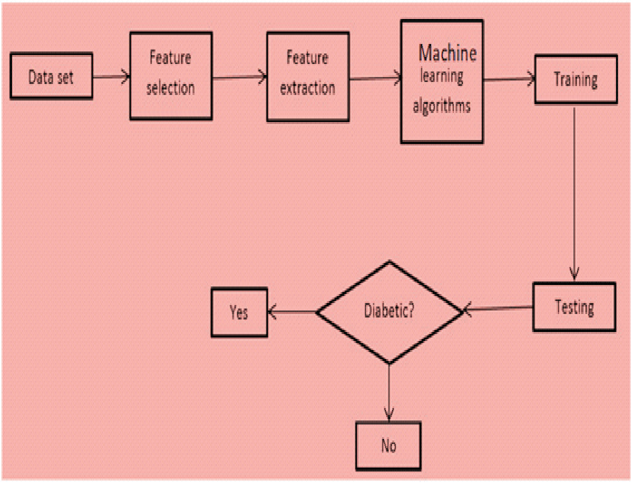

**Fig 1. Workflow of the experiment**

**Experiment Framework**

The architecture of the experiment was divided into four components

1.)Collection of datasets

2.)Classification of dataset and splitting dataset into train and test dataset

3.)Passing data into the model

4.)Obtaining output

**Model Evaluation**

Assessed the model performance using metrics such as accuracy, precision, recall, F1-score.

Implemented cross-validation method to ensure robust performance assessment.

## RESULTS AND DISCUSSION

The result of training Random Forest, Decision Tree, SVM and KNN model on a dataset of diabetes risk prediction for the number of epochs depends on various factors, including the quality and size of dataset, the preprocessing steps applied to the data, the learning rate, and other hyperparameters.

For the experiment all four selected models have been trained separately, and predictions have been generated for the validation data-set, based on which the classification reports

**Table 2: Accuracy Obtained by using different models**

| Models | Accuracy Obtained |
|---|---|
| Random Forest | 99.03 % |
| Decision Tree | 95.00 % |
| K Nearest Neighbour | 89.00% |
| SVM | 90.00% |

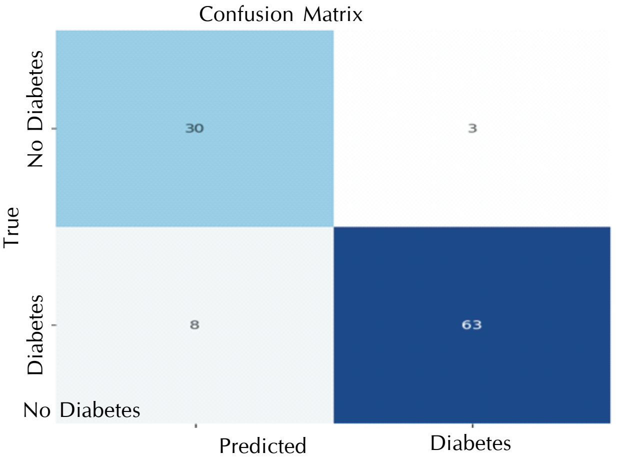

**Fig 2: Confusion Matrix of KNN for diabetes risk prediction**

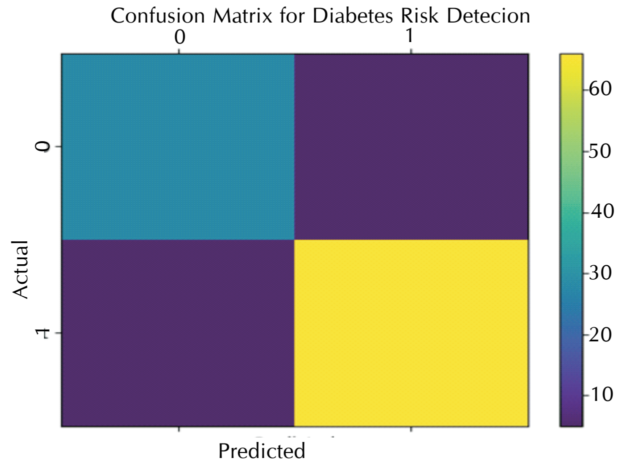

**Fig 3: Confusion Matrix of SVM for diabetes risk prediction**

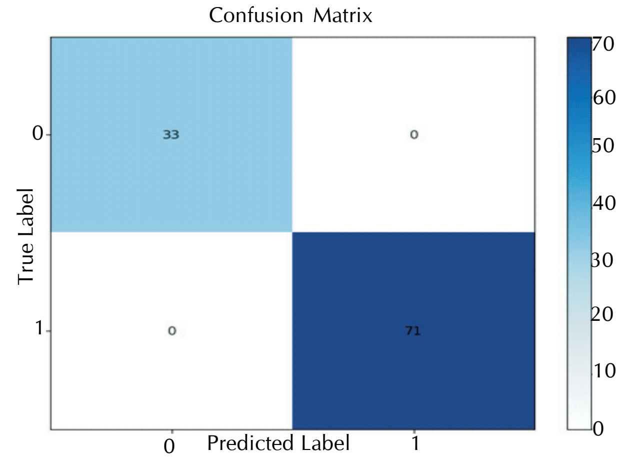

**Fig 4: Confusion Matrix of Random Forest for diabetes risk prediction**

for each model have been generated.

Necessary libraries and dependencies have been imported the for all the selected models for the experiment. The dataset of diabetes risk prediction has been loaded and pre-processed using suitable techniques for the models. The data have been organized into training and validation sets.

The factors affecting the performance of the selected models Random Forest, Decision Tree, K-Nearest Neighbour and SVM prediction for the diabetes risk prediction are dependent on various factors, including the dataset size, the complexity of the problem, and the availability of labelled data.

Accuracy is a common evaluation metric for classification models, providing an overall measure of how well the model correctly classifies the data.

Formula: (TP + TN) / (TP + TN + FP + FN)

Precision measures the accuracy of positive predictions. It calculates the ratio of correctly predicted positive observations to the total predicted positives.

Formula: TP / (TP + FP)

Recall measures the ability of the model to capture all the relevant instances. It calculates the ratio of correctly predicted positive observations to the total actual positives.

Formula: TP / (TP + FN)

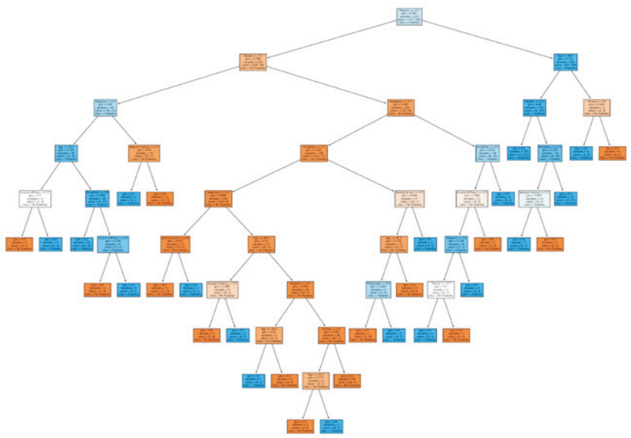

**Fig 5: Decision Tree for diabetes risk prediction**

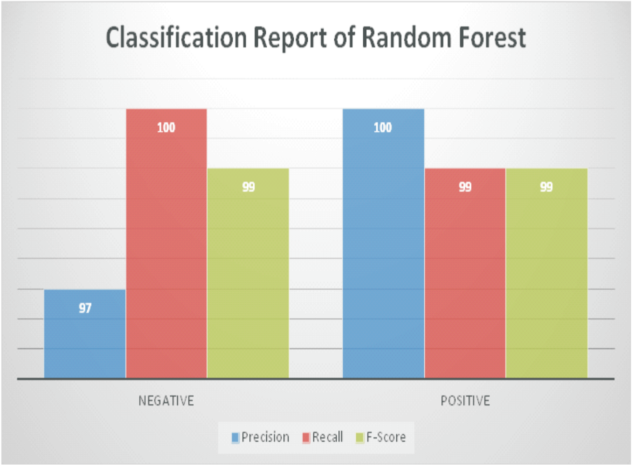

**Chart 1: Classification Report of Random Forest**

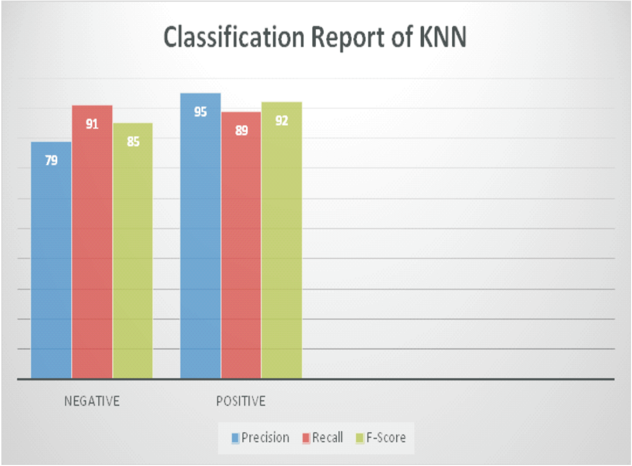

**Chart 2: Classification Report of KNN**

F-Score is the harmonic mean of precision and recall. It provides a balanced measure of a model's performance.

Formula**:** 2 * (Precision * Recall) / (Precision + Recall)

Where:

**•TP (True Positives)**:

The number of samples correctly predicted as positive (correctly classified as belonging to the positive class).

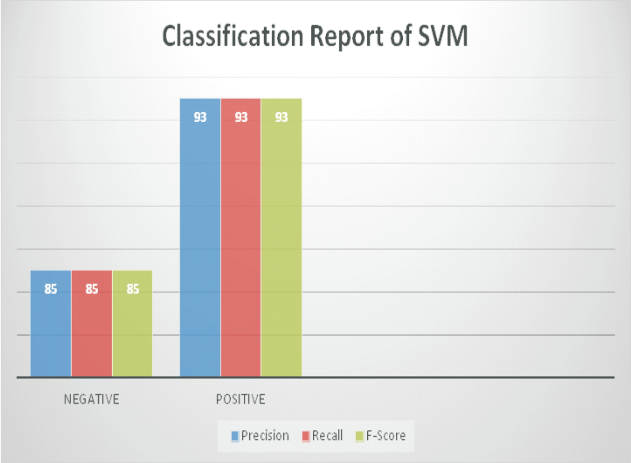

**Chart 3: Classification report of SVM**

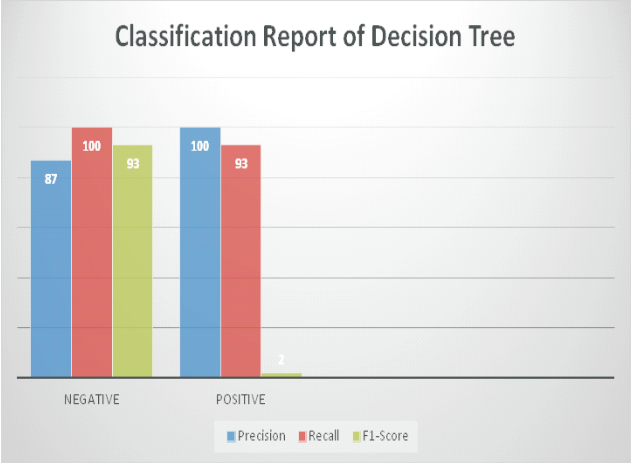

**Chart 4: Classification report of decision tree**

•**TN (True Negatives)**:
The number of samples correctly predicted as negative (correctly classified as not belonging to the positive class).

•**FP (False Positives)**:
The number of samples incorrectly predicted as positive (misclassified as belonging to the positive class when they do not).

•**FN (False Negatives)**:
The number of samples incorrectly predicted as negative (misclassified as not belonging to the positive class when they do).

Since for diabetes disease risk prediction the dataset used from Kaggle (https://www.kaggle.com/code/therealsampat/early-stage-diabetes-prediction) had comma separated values and hence the Random Forest Classifier found to outperform and provide a well decisive result. The results demonstrate that when compared to other algorithms, Random Forest performs with the greatest accuracy of 99.03%. preceded by Decision Tree with an accuracy of 95% preceded by SVM with an accuracy of 90% and K-Nearest Neighbour with an accuracy of 89%. Through the result of excellent accuracy and generalization was obtained which in-turn be helpful for medical practitioners (Sneha,2019; Swapna,2018; Kavakiotis,2017;Deepti,2018;Abdulhadi,2021; Pujianto,2019).

## CONCLUSION

One of the greatest tools for using classification and prediction techniques is machine learning. In order to compare the results on the following metrics: Accuracy, Recall, F1-Score, Precision, we used a range of machine learning algorithms in this work, including SVM, Decision Tree, k-nearest neighbour and Random Forest on the diabetes risk prediction Dataset. According to the experiment's findings, the random forest classifier is the greatest with the accuracy of 99.03% preceded by Decision Tree with an accuracy of 95% preceded by SVM with an accuracy of 90% and K-Nearest Neighbour with an accuracy of 89%. Due to its shown accuracy in detection, efficacy in therapeutic application, and cost-effectiveness, machine learning has been included into medical diagnosis systems.

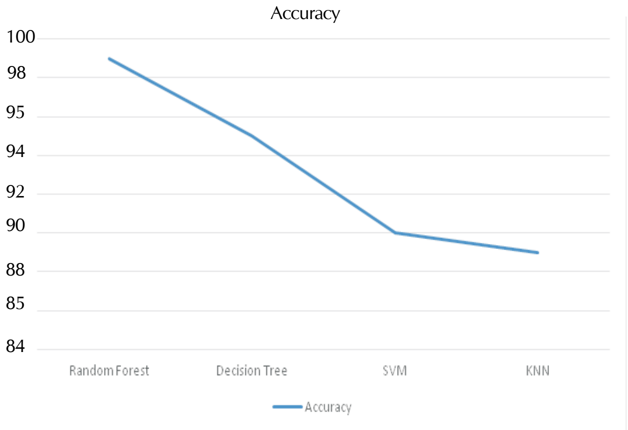

**Chart 5: Accuracy depicted in graph of different models for predicting diabetes risk**

The general consensus among academics, medical professionals, and industry participants is that artificial intelligence has the ability to change the current state of late medicine and detection as a result of human mistake (Sneha,2019; Swapna,2018; Kavakiotis,2017; Deepti,2018; Abdulhadi,2021; Pujianto,2019). Medical detection systems can be built with efficiency and dependability because of automation. This is made possible in large part by machine learning and its potent classification and prediction models.

## REFERENCES

**Argentiero, P., Chin, R. and Beaudet, P.1982**. An automated approach to the design of decision tree classifiers. IEEE Trans. *Pattern Anal. Mach. Intell.* **1:** 51–57 .

**A. Al Jarullah. 2011.** Decision tree discovery for the diagnosis of type II diabetes. *2011 International Conference on Innovations in Information Technology*, Abu Dhabi, United Arab Emirates, pp. 303-307, doi: 10.1109/INNOVATIONS.2011.5893838.

**Breiman, L. 2001.** Random forests. *Mach. Learn***. 45(1):** 5–32

https://www.kaggle.com/code/therealsampat/early-stage-diabetes-prediction

ilvanciuc, O.Support Vector Machine [internet].2005. Available from: http://www.support-vectormachines.org/SVM_review.html.

https://scikit-learn.org/stable/modules/svm.html

https://www.geeksforgeeks.org/k-nearest-neighbours/

https://towardsdatascience.com/understanding-random-forest-58381e0602d2

https://www.geeksforgeeks.org/decision-tree/

**Cortes, C. and Vapnik, V. 1995.** Support-vector networks. *Mach. Learn*. **20(3):** 273–297.

**Deepti Sisodia and Dilip Singh Sisodia**. **2018**.Prediction of Diabetes using Classification Algorithms,Procedia Computer Science,Volume 132.pp. 1578-1585,ISSN 1877-0509, https://doi.org/10.1016/j.procs.2018.05.122.

**Friedman, J. H., Baskett, F. and Shustek, L.J**. **1975.**An algorithm for finding nearest neighbors. IEEE Trans. Comput. **100(10):** 1000–1006.

**Ioannis Kavakiotis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas and Ioanna Chouvarda. 2017**. Machine Learning and Data Mining Methods in Diabetes Research.*Com putational and Structural Biotechnology J*,Volume 15.pp.104-116, ISSN 2001-0370, https://doi.org/10.1016/j.csbj.2016.12.005.

**N. Abdulhadi and A. Al-Mousa. 2021.** Diabetes Detection Using Machine Learning Classification Methods, *2021 International Conference on Information Technology (ICIT)*, Amman, Jordan, pp. 350-354. doi: 10.1109/ICIT52682.2021.9491788.

**Pujianto, U., Setiawan, A. L., Rosyid, H. A. and Salah, A.M.M. 2019.** Comparison of naïve Bayes algorithm and decision tree C4. 5 for hospital readmission diabetes patients using hba1c measurement. *Knowledge Engineering and Data Science*. **2(2):** 58-71.

**Sneha, N. and Gangil, T. 2019** Analysis of diabetes mellitus for early prediction using optimal features selection. *J Big Data* 6. 13 . https://doi.org/10.1186/s40537-019-0175-6

**Swapna G., Vinayakumar R. and Soman K.P. 2018.** Diabetes detection using deep learning algorithms,ICT Express, Volume 4, Issue 4. pp. 243-246, ISSN 2405-9595. https://doi.org/10.1016/j.icte.2018.10.005.

**S. Benbelkacem and B. Atmani. 2019 .** Random Forests for Diabetes Diagnosis, International Conference on Computer and Information Sciences (ICCIS), Sakaka, Saudi Arabia, 2019. pp. 1-4. doi: 10.1109/ICCISci.2019.8716405.

**Uazman Alam., Omar Asghar., Shazli Azmi and Rayaz A. Malik.** 2014. Chapter 15 - General aspects of diabetes mellitus,Editor(s): Douglas W. Zochodne, Rayaz A. Malik,Handbook of Clinical Neurology, Elsevier,Volume 126.pp. 211-222,ISSN 0072-9752,ISBN 9780444534804, https://doi.org/10.1016/B978-0-444-53480-4.00015-1.

**V. V. Vijayan and C. Anjali. 2015**. Prediction and diagnosis of diabetes mellitus — A machine learning approach, *2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*, Trivandrum, India, pp. 122-127. doi: 10.1109/RAICS.2015.7488400.