Diabetes Prediction Web Application Using Machine Learning

Ashish Dwivedi
Dept. of Computer Engineering,
Armiet,
Maharashtra, India

Jay Patil
Dept. of Computer Engineering,
Armiet,
Maharashtra, India

Prathamesh Pathare
Dept. of Computer Engineering,
Armiet,
Maharashtra, India

Abstract: **Diabetes is a disease caused due to the increase level of blood glucose. Diabetes is a chronic disease with the potential to cause a worldwide health care crisis. According to International Diabetes Federation 382 million people are living with diabetes across the whole world. By 2035, this will be doubled as 592 million. Various traditional methods, based on physical and chemical tests, are available for diagnosing diabetes. However, early prediction of diabetes is quite challenging task for medical practitioners due to complex interdependence on various factors as diabetes affects human organs such as kidney, eye, heart, nerves, foot etc. Data science methods have the potential to benefit other scientific fields by shedding new light on common questions. One such task is to help make predictions on medical data. Machine learning is an emerging scientific field in data science dealing with the ways in which machines learn from experience. The aim of this project is to develop a web application which can perform early prediction of diabetes for a patient with a higher accuracy by combining the results of machine learning technique. This project aims to predict diabetes via supervised machine learning methods Random Forest. This project also aims to propose an effective technique for earlier detection of the diabetes disease.**

Keywords: **Diabetes prediction, Machine learning, Flask, Random Forest (RF).**

I. Introduction

Diabetes mellitus is an endless infection portrayed by hyperglycaemia. It might cause numerous inconveniences. As per the developing bleakness as of late, in 2040, the world's diabetic patients will achieve 642 million, which implies that one of the ten grown-ups later on is experiencing diabetes. There is no uncertainty this disturbing figure needs extraordinary consideration. World Health Organization has assessed 12 million passings happen around the world, consistently because of Heart maladies. A large portion of the passings in the United States and other created nations are expected to cardio vascular maladies. The early visualization of cardiovascular sicknesses can help in settling on choices on way of life changes in high hazard patients and thus decrease the intricacies. This exploration means to pinpoint the most significant/hazard elements of coronary illness just as anticipate the general hazard utilizing calculated relapse. Machine Learning has been connected to numerous parts of medicinal wellbeing. In this project, we utilized Random Forest to anticipate diabetes mellitus.

The dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset.

Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

The datasets consist of several medical predictor (independent) variables and one target (dependent) variable, Outcome. Independent variables include the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

Proposed model is to anticipate diabetes that specialists can be valuable as a model to help foresee diabetes. In this examination, analysed the connection between difficulties in diabetic patients and their properties, for example, blood glucose, circulatory strain, tallness, weight, and haemoglobin and weight record of the patients. The point of this examination is to foresee confusions dependent on their manifestations.

A. Existing System

The healthcare enterprise collects big quantities of healthcare facts which, unluckily, are not “mined” to find out hidden information. Clinical decisions are frequently made based totally on docs‟ intuition and enjoy instead of at the knowledge wealthy statistics hidden in the database. This exercise ends in unwanted biases, errors and excessive medical. The existing process is very slow to give the result. It is very difficult to find diabetes or not.

B. Proposed system & It's Advantages

Diabetes Prediction Using Machine Learning is an internet-primarily based device gaining knowledge of utility, skilled through a Pima Indian dataset. The person inputs its particular clinical information to get the prediction of diabetes. The set of rules will calculate the opportunity of presence of diabetes. Thus, minimizing the price and time required to are expecting the disorder. Format of statistics plays essential element on this software. At the time of uploading the user information utility will take a look at its right record format and if it no longer as consistent with want then ERROR dialog box may be induced. Our device might be implementing the algorithm: Random Forest. The algorithms may be educated the use of the statistics set obtained from University of California, Irvine. 75% of the entries in the statistics set can be used for education and the last 25% for testing the accuracy of the set of rules. Furthermore, a few steps can be taken for optimizing the algorithms thereby enhancing the accuracy.

Advantages:

1. Powerful, flexible, and easy to use.
2. Increased efficiency of doctor.
3. Improved patient satisfaction.
4. Reduce the use of papers.
5. Simple and Quick.
6. More accurate result.

II. Literature Review

In this section, some closely related works are discussed briefly. In most of the research works, Pima Indians Diabetes Dataset (PIDD) have been used by many researchers for diabetes prediction. Various super- vised machine learning algorithms were used to predict diabetes (Kaur & Kumari, 2020) [14]. Radial basis function (RBF) kernel SVM, artificial neu- ral network (ANN), multifactor dimensionality reduction (MDR), linear SVM and k-NN are some of them to mention. Based on p value and odds ratio (OR), Logistic Regression (LR) has been used to recognize the risk factors for diabetes (Maniruzzaman et al., 2020) [17]. Four classifiers have been adopted to predict diabetic patients, such as NB, DT, Adaboost, and RF. Partition protocols like- K2, K5, and K10 were also adopted, repeating these protocols into 20 trails. For the performance measurement of the classifiers, accuracy (ACC) and area under the curve (AUC) were analyzed.

Kopitar et al. (2020) [16] showed a comparison of widely utilized regression models such as Glmnet, RF, XGBoost, LightGBM for predicting type 2 diabetes mellitus. The goal of this work was to examine if innovative machine learning methodologies gave any advantages in early prediction of impaired fast glucose and fasting plasma glucose (FPGL) levels compared to classic regression techniques.

A hybrid model to detect type 2 diabetes was suggested by Albahli (2020) [2] . In order to extract unknown, hidden property from the dataset and to obtain more exact results, we use K-mean clustering, which is followed by the execution of a Random Forest and XGBoost classifier.

Yahyaoui et al. (2019) [27] suggested a Machine Learning Techniques (ML) DSS for anticipating diabetes. They compared traditional machine learning with approaches to the deep learning. The authors applied the classifiers most typically used for a standard machine learning method: SVM and the Random Forest (RF). In contrast, they used a full-scale neural network (CNN) for Deep Learning (DL) to forecast and identify patients who suffer from diabetes.

Zou et al. (2018) [30] predicted diabetes using the decision tree, random forests, and neural network. The dataset is collected from the Luzhou physical exams in China. The PCA was applied to reduce the dimension of the dataset. They selected several ML approaches to execute independent test to verify the universal applicability of method.

Supervised machine learning models which explore data-driven approaches were used to identify patients with diabetes diseases (Dinh et al., 2019) [7]. A complete research was conducted based on the National Health and Nutrition Examination Survey (NHANES) dataset. To develop models for cardiovascular, prediabetes, and diabetes detection, they have used all available feature variables within the data. Using various time frames and set of features within the data, different machine learning models, namely Support Vector Machines, logistic regression, gradient boosting and random forest were evaluated for the classification.

In Choubey et al. (2017) [6] the authors used NBs for the classification on all the attributes. Afterwards GA was used as an attribute selector and NBs used the selected attributes for classification. The experimental results show the performance of this work on PIDD and provide better classification for diagnosis. Three specific supervised machine learning methods are used by Joshi and Chawan (2018) [13] , namely SVM, Logistic regression and ANN. His goal for research was to predict diabetes patients and he has also proposed an effective model for the prior detection of diabetes disease. Rajeswari and Prabhu (2019) [21] focused on machine learning classification algorithms for predicting diabetes disease with more accuracy. Their study in SVM classification algorithm achieved highest accuracy. Various measures have been used to calculate the performance of classification algorithms.

An intelligent model using machine learning practices is developed (Nilashi et al., 2017) [18] to identify diabetes disease. This model is con- structed using approaches like clustering, removal of noise and classification, each of which made use of SOM, PCA and NN, respectively. The adaboost and bagging ensemble techniques are used to detect diabetes

(Perveen et al., 2016) [19]. Along with standalone data mining technique, a base learner is used to identify patients with diabetes mellitus, namely J48 (c4.5) decision tree that makes use of multiple diabetes risk factors. In the Canadian Primary Care Sentinel Surveillance Network, three different ordinal adult groups are selected for classification. Experimental result shows that, the adaboost ensemble method shows better performance than both bagging and standalone J48 decision tree. For diagnosing T2DM, Kazerouni et al. (2020) [15] has taken in consideration four different classification models, namely SVM, K-NN, ANN and LR. A comparison is done among these algorithms to measure the diagnostic power of this algorithms.

A. *Materials and Methods*

1. Random Forest Algorithm: There was a description of the random forest system. He is actually a metal architect, but weka is part of the decision-making tree approach because he has an ad hoc classification, Random Tree. In each cycle of the hauling method, a common training machine for natural trees creates a unique choice matrix and often produces great risk factors. The tree is finally fully cultivated and is not cut. For a fresh dataset the tree is pressed down. The teaching sample is allocated to the tag when the command line node finishes. This operation is known as a Random Forest Production and is elaborated over all forests.

2. Glm In R Logistic Regression: Regression of ordinary lower squares offers linear designs of constant factors. However, a good number of statistics and scientists ' information of concern are not constant and therefore other techniques should be employed to generate helpful predictive models. The glm () control was intended for the performance of generalized linear models for binary results information, count data, probability information, percentage information, etc.

3. Naive Bayes classifier: Naïve Bayes executes Naïve Bayes Simple Probable Naïve classifier. Naïve Bayes is able to use kernel thickness parameter estimation which boost productivity if the hypothesis of normality is largely inaccurate. It utilizes the probability distribution of numerical attributes modeling.

4. Decision tree: There are nodes in each tree. Every node has one output variable connected with it. The corners of the node are the complete feasible node scores. A leaf reflects the valuation depending on the entry numbers provided on the route from the root of the leaf node. Trees begin from a root node always and finish on a leaf. Be noted that at no stage in the process of the nodes, the plants do not fit.

5. Liner SVM: Support Vector Machines (SVM) is used to recognize picture and handwriting patterns in many ranking situations. Medical science has long been using carbohydrate identification support vector machines. Now, there are 2 kinds of problem. One those are linearly separable and the other is nonlinearly separable. For linearly separable problems, SVM uses a linear kernel which classifies dataset among different classes using a linear hyper-plane.

6. RBF kernel SVM: For non-linearly separable problem, SVM uses a RBF kernel which is a non-linear kernel function because no hyper-plane is sufficient enough to accurately classify data.

7. Stratified k-fold Cross Validation: Stratification is the method of reorganizing the information so that every slice represents the entirety. The plates are chosen to nearly equal the average reaction price for all plates. There are many algorithms and approximate that have been utility to bode the feeling assail among patients. But cultivated neural net emerge to be the largest do technique for reins onset soothsaying, and it is a highly powerful drive utility in assortment drudgery, as well as to explain many significant problems namely indication augmentation, identification, and foreboding of foreshadowing and substitute.

- ALGORITHM:

The Random Forest algorithm we've used. Random Forest is a flexible and user-friendly software technique that produces a great result, most of the time without setting super parameters. It is also one of the most common techniques, since it is easy to use and can be utilized for classification and regression. Random Forest is a supervised learning algorithm. It is creating woods and randomizing it somehow. The forest it constructs is a group of decision trees, educated most of the moment by the bagging technique. The general concept of the bagging technique is that the overall outcome is increased by a mixture of training designs. In plain terms: Random tree creates and merges several choice forests to make a more precise and consistent forecast. One major benefit of random forests is that they can be used for ranking as well as for regression issues.

B. Methodology

Fig. 1. Work-flow diagram of the proposal.

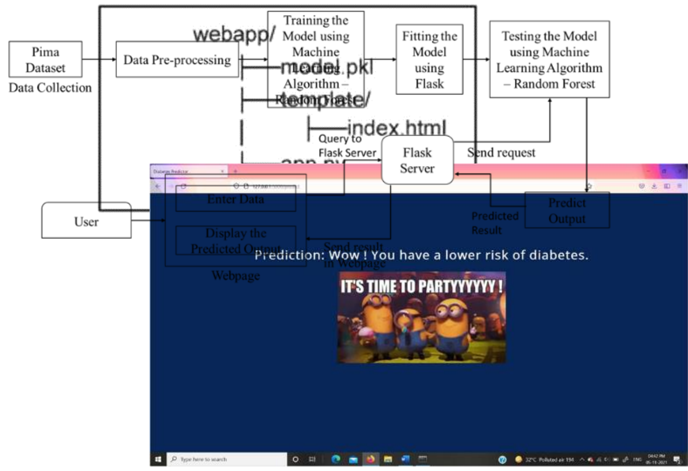
Fig. 1 depicts the proposed framework for diabetes prediction. Firstly, we pre-process datasets. In the pre-processing stage, correlation between attributes of the datasets is analyzed for finding useful features in detecting diabetes. After that, the data is divided into two sets: training and testing. The training set is utilized to develop predictive ML models using a variety of machine learning algorithms. Next, we assess the proposal's performance with respect to different metrics. Finally, the best ML model is deployed in a web application using flask. Following this, we describe the workflow of each part briefly:

1. Data Collection: The datasets were compiled from a wide variety of sources, including diabetes statistics and health characteristics obtained from people around the world and from various health institutes.
2. Data Pre-processing: Several pre-processing techniques are applied on the datasets before feeding these datasets into the machine learning model so that the performance of the model is improved. The pre-processing tasks include removing outliers and dealing with missing values, data standardization, encoding, and so on.
    - Outliers Removal: Attributes' values that are beyond acceptable boundaries and have high variation from the rest of the respective attribute's value might be present in the dataset. Such attributes' value might degrade the machine learning algorithm's performance. To eliminate such outliers, we applied the IQR (Inter-quartile Range) approach.
    - Missing value Handling: To improve model performance, the mean value of each attribute was employed for handling the missing values.
    - Label Encoding: Label encoding is the process of converting the labels of text/categorical values into a numerical format that ML algorithm can interpret.

3. Model Construction and Prediction: To construct the predictive model, 70% of the pre-processed data has been used for training while the remaining 30% data is used for the testing purpose.

4. Web Application Development: To develop a smart web application, we have used the Flask micro-framework and integrated the best model. To predict diabetes, a user is required to submit a form with necessary numbers of diabetes related parameters. The application uploaded in a server predicts the results using the adopted machine learning model. We describe the adopted machine learning algorithms in the following sections.

i. Web Application Development using Flask:


Pima Dataset
Data Collection
Data Pre-processing
Training the Model using Machine Learning Algorithm – Random Forest
Fitting the Model using Flask
Testing the Model using Machine Learning Algorithm – Random Forest
webapp/
model.pkl
template/
index.html
app.py
Query to Flask Server
Flask Server
Send request
User
Enter Data
Display the Predicted Output
Webpage
Send result to Webpage
Predicted Result
Predict Output
IT'S TIME TO PARTYYYYYY !

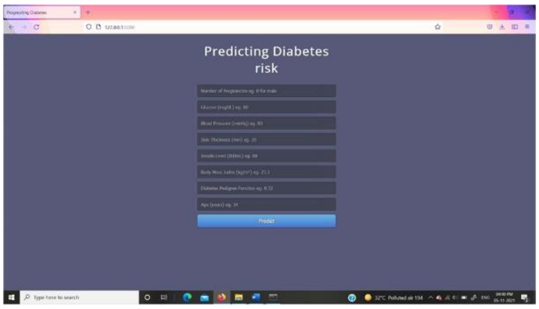
Prediction: Wow ! You have a lower risk of diabetes.

Fig. 2. File structure of the web application.

Flask is a Python-based microweb platform that allows users to add application functionality as if they were built into the framework itself. Fig. 11 shows the basic file structures of the developed application and this development process comprises of four different program modules as follows:

- Model.pkl: This contains the machine learning model to predict diabetes. As Random Forest provide the highest with all the features, we will integrate this as predictive model in the model.pkl file.
- App.py: This package includes Flask APIs that receive Diabetes information through GUI or API calls, compute the predicted value using our model, and return it.
- Template: The HTML form (index.html) in this folder allows the user to enter diabetes information and shows the expected outcome.
- Static: This folder contains the css file which has the styling required for our HTML form.

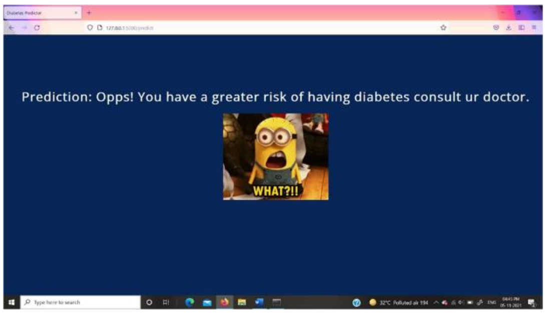
ii. Prediction results of web application:



Predicting Diabetes risk

Number of Pregnancies eg. 0 for male
Glucose (mg/dl) eg. 80
Blood Pressure (mmHg) eg. 80
Skin Thickness (mm) eg. 20
Insulin Level (IU/ml) eg. 80
Body Mass Index (kg/m²) eg. 23.1
Diabetes Pedigree Function eg. 0.52
Age (years) eg. 34

Fig. 3. User input validation in web application.



Prediction: Opps! You have a greater risk of having diabetes consult ur doctor.

Fig. 4. Prediction results in web application.

When a user runs the application, a page will appear as shown in Fig. 3 . The application can check for the valid input for every fields. If the user enters an invalid value for any of the parameters, a warning message is displayed. If the user provides valid information, the appli- cation will predict whether the user has diabetes or not, as illustrated in Fig. 4 .

Conclusion

Diabetes is vital health hassle in human society. Deep studying an rising region of Machine Learning showed a few promising bring about different area of clinical diagnose with excessive accuracy. It continues to be an open area waiting to get applied in Diabetes predication. Some strategies of deep studying has been discussed which may be implemented for Diabetes predication, alongside pioneer machine getting to know algorithms. An analytical assessment has been completed for locating out best available algorithm for clinical dataset. In future our purpose is to carry ahead the work of temporal scientific dataset, wherein dataset varies with time and retraining of dataset is needed. The main aim of this project was to design and implement Diabetes Prediction Web Application Using Machine Learning Methods and it has been achieved successfully. The proposed approach uses classification and ensemble learning method in which Random Forest classifier are used. Individuals who are unsure or simply want a routine check-up may consider this application.

Future Enhancement

The proposed system is Diabetes Prediction Using Machine Learning. We will examine a larger and deeper dataset for Indian patients with additional attributes for improved accuracy, and we will publish the web application on a cloud platform such as AWS that is freely available and can be evaluated by real users.

References

1. Ahamed, K. U., Islam, M., Uddin, A., Akhter, A., Paul, B. k., Yousuf, M.A.,… Moni, M. A., et al. (2021). A deep learning approach using effective pre-processing techniques to detect covid-19 from chest CT-scan and X-ray images. Computers in Biology and Medicine, 139, Article 105014. 10.1016/j.compbiomed.2021.105014.
2. Albahli, S. (2020). Type 2 machine learning: An effective hybrid prediction model for early type 2 diabetes detection. Journal of Medical Imaging and Health Informatics, 10, 1069-1075
3. Brownlee, J. (2016a). A gentle introduction to the gradient boosting algorithm for machine learning.
4. Brownlee, J. (2016b). K-nearest neighbours for machine learning. https:// machinelearningmastery.com/k-nearest-neighbors-for-machine-learning/ Accessed: 2021-03-20
5. Brownlee, J. (2016c). Logistic regression for machine learning. https://www. geeksforgeeks.org/understanding-logistic-regression/ Accessed: 2021-03-20.
6. Choubey, D. K. , Paul, S. , Kumar, S. , & Kumar, S. (2017). Classification of Pima Indian diabetes dataset using naive Bayes with genetic algorithm as an attribute selection. In Proceedings of the international conference on communication and computing system (ICCCS 2016) (pp. 451–455)
7. Dinh, A. , Miertschin, S. , Young, A. , & Mohanty, S. D. (2019). A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Medical Informatics and Decision Making, 19 , 211
8. Gadekallu, T. R. , Khare, N. , Bhattacharya, S. , Singh, S. , Reddy Maddikunta, P. K. , Ra, I. H. , et al. (2020). Early detection of diabetic retinopathy using pca-firefly based deep learning model. Electronics, 9
9. Gandhi, R. (2018). Naive bayes classifier. https://towardsdatascience.com/naive- bayes-classifier-81d512f50a7c Accessed: 2021-03
10. Gulshan, V. , Peng, L. , Coram, M. , Stumpe, M. C. , Wu, D. , Narayanaswamy, A. , et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. Journal of the American Medical Association, 316 , 2402–2410
11. Gupta, S. (2019). Decision tree. https://www.geeksforgeeks.org/decision-tree/ Accessed: 2021-03-20.
12. Haq, A. U. , Li, J. P. , Khan, J. , Memon, M. H. , Nazir, S. , Ahmad, S. , et al. (2020). Intelligent machine learning approach for effective recognition

of diabetes in e-healthcare using clinical data. Sensors, 20 , 2649 .

13. Joshi, T. N. , & Chawan, P. (2018). Diabetes prediction using machine learning techniques. International Journal of Engineering Research and Applications, 8 , 9–13 .
14. Kaur, H., & Kumari, V. (2020). Predictive modelling and analytics for diabetes using a machine learning approach. Applied Computing and Informatics. 10.1016/j.aci.2018.12.004 .
15. Kazerouni, F. , Bayani, A. , Asadi, F. , Saeidi, L. , Parvizi, N. , & Mansoori, Z. (2020). Type2 diabetes mellitus prediction using data mining algorithms based on the long-noncoding rnas expression: A comparison of four data mining approaches. BMC Bioinformatics, 21 , 1–13 .
16. Kopitar, L. , Kocbek, P. , Cilar, L. , Sheikh, A. , & Stiglic, G. (2020). Early detection of type 2 diabetes mellitus using machine learning-based prediction models. Scientific Reports, 10 , 1–12 .
17. Ahammed, B. , & Abedin, M. M. (2020). Classification and prediction of diabetes disease using machine learning paradigm. Health Information Science and Systems, 8 , 1–14 .
18. Nilashi, M. , Ibrahim, O. , Dalvi, M. , Ahmadi, H. , & Shahmoradi, L. (2017). Accuracy improvement for diabetes disease classification: A case on a public medical dataset. Fuzzy Information and Engineering, 9 , 345–357 .
19. Perveen, S. , Shahbaz, M. , Guergachi, A. , & Keshavjee, K. (2016). Performance analysis of data mining classification techniques to predict diabetes. Procedia Computer Science, 82 , 115–121 .
20. Pranto, B. , Mehnaz, S. , Mahid, E. B. , Sadman, I. M. , Rahman, A. , Momen, S. , et al. (2020). Evaluating machine learning methods for predicting diabetes among female patients in Bangladesh. Information, 11 , 374 .
21. Rajeswari, M. , & Prabhu, P. (2019). A review of diabetic prediction using machine learning techniques. International Journal of Engineering and Techniques, 5 , 1–7 .
22. Ray, S. (2017). Understanding support vector machine(svm). https://www. analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example- code/ Accessed: 2021-03-20.
23. Tigga, N. P. , & Garg, S. (2020). Prediction of type 2 diabetes using machine learning classification methods. Procedia Computer Science, 167 , 706–716 .
24. UCI Machine Learning Repository. (1998). Diabetes data set. https://archive.ics.uci.edu/ml/datasets/diabetes Accessed: 2021-03-20.
25. Vinayakumar, R. , Alazab, M. , Soman, K. , Poornachandran, P. , Al-Nemrat, A. , & Venkatraman, S. (2019). Deep learning approach for intelligent intrusion detection system. IEEE Access : Practical Innovations, Open Solutions, 7 , 41525–41550.
26. World Health Organization. (2021). Diabetes . World Health Organization https://www.who.int/news-room/fact-sheets/detail/diabetes Accessed: 2021-04-20 .
27. Yahyaoui, A. , Jamil, A. , Rasheed, J. , & Yesiltepe, M. (2019). A decision support system for diabetes prediction using machine learning and deep learning techniques. In Proceedings of the 1st international informatics and software engineering conference (UBMYK) (pp. 1–4) .
28. Yiu, T. (2019). Understanding random forest. https://towardsdatascience.com/ understanding-random-forest-58381e0602d2 Accessed: 2021-03-20.
29. Yu, W. , Liu, T. , Valdez, R. , Gwinn, M. , & Khoury, M. J. (2010). Application of support vector machine modeling for prediction of common diseases: The case of diabetes and pre-diabetes. BMC Medical Informatics and Decision Making, 10 , 16 .