Diabetes Disease Prediction Using Machine Learning Algorithms

Prashant Kumbharkar[1*], Deepak Mane[2], Santosh Borde[3], Sunil Sangve[4]

Abstract

Diabetes is one of the most awful diseases in the world which has no method to treat it after a set stage. Over 422 million individuals in the world are diagnosed with diabetes and many more are at danger. Thus, timely detection and medication is crucial to reduce diabetes and its accompanying health consequences. In this study a system is developed for diabetes diseases prediction and classification using Machine Learning (ML) approaches. The dataset is taken from KM Hospital and Research Centre, Pune, Sahyadri Hospital Pune and Research Centre and Data. Four independent ML algorithms Logistic Regression, Naïve Bayes, Support Vector Machine and Decision Tree are employed and analyzed the model using multiple quantitative criteria. The purpose of this framework is to discover diabetes early and to save money and time of a patient using several machine learning methods.

KeyWords: Diabetes, Prediction, Classification, Machine Learning, Model Evaluation, Logistic Regression, Naive Bayes, Support Vector Machine, Decision Tree.

DOI Number: 10.14704/NQ.2022.20.12.NQ77198 **NeuroQuantology2022;20(12): 2225-2231**

Introduction

Diabetes is rapidly becoming one of the most widespread diseases in India. “According to the World Health Organization, India has roughly 31.7 million diabetes patients in the year 2000, and this number is expected to rise to 79.4 million by the year 2030. The statistics collected by the WHO on diabetes patients in India are shown in Figure 1. In India, there is an urgent need to implement disease prevention and control measures”.

Machine learning algorithms are mathematical approaches that are highly effective in evaluating enormous amounts of data and recommending specific actions on the basis of that data. These algorithms are sometimes referred to as "deep learning algorithms." These algorithms are helpful for assessing a data collection and making predictions about the values that should be entered next. ML algorithms are being used by a multitude of researchers [1] [5][6] for the purpose of illness forecasting and management. The findings of machine learning algorithms indicated that they performed very well in the prediction of various illnesses. It is vital to employ machine learning algorithms in order to investigate their potential for diabetes prediction. This will allow the appropriate preventative measures to be implemented in the event that diabetes develops.

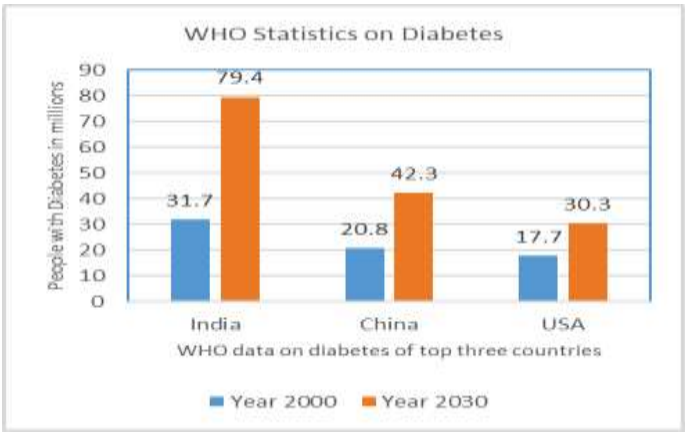

Fig 1: WHO report on Diabetes

Corresponding author: Prashant Kumbharkar
Address: [1]Professor, Computer Engineering JSPM'S RajarshiShahu College of Engineering,Pune, [2,3,4]Professor, Computer Science Department, JSPM'S RajarshiShahu College of Engineering, Pune

E-mail: [1]pbk.rscoe@gmail.com, [2]dtmane@gmail.com, [3]santoshborde@yahoo.com, [4]sunilsangve@gmail.com

Literature Review

This article's goal is to provide an explanation of a function that belongs to an algorithm and classification ensemble that belongs to a machine learning approach. The J48 decision tree was applied in order to identify hypertension in patients, despite the fact that the patients may or may not have had diabetes. The following are some of the risk factors for diabetes that were taken into consideration: Based on the outcomes of the research, it seems that the Ad boost Machine Learning ensemble strategy is superior to bagging and a J48 decision tree in terms of efficiency. The primary objective behind the development of a glucose prediction model was to establish whether or not a patient will be at risk for developing diabetes at a certain age. In the field of machine learning, the conceptual model is put to considerable use, and the implementation of it makes extensive use of decision trees to address and solve problems. The results that were seen are accurate because the method that was developed is effective in diagnosing diabetes episodes at a given age with greater accuracy by employing a Prediction Model.

The consequently, the findings that were observed. This is a result of the approach that was created. Before being utilized for estimation, the support vector was put through its paces by way of testing and analysis of the dataset. The investigation into disease used the compilation of data on high blood pressure that was gathered from the database at UCI.

Because of the way the software was written, we were able to collect the most reliable data possible. It is possible to achieve very high levels of dependability by reducing the amount of time needed to generate a dataset. This may be done in a number of ways. Utilizing a data sampling model for hypertension that is comprised of two sub-modules for diabetes prediction may often end up being rather pricey. In the first section of our analysis, we make use of an artificial neural network, and in the second half of our study, we follow the same approach. The patient's blood sugar levels when they are fasting are included into decision trees, which are then used to estimate the effects that hypertension has on patients' overall health. In addition to this, it may be included into the system that is used to classify the risk of acquiring diabetes.

The DT, ANN, and Naive Bayes are three of the most well-known classification algorithms that may be used for machine learning. The author employed these algorithms so that they could complete this work. It is possible, via the use of methods such as classifiers, to increase the dependability of the models that are being constructed. According to the conclusions of the research, the algorithm referred to as Random Forest yields the best outcomes when compared to all of the other algorithms that were investigated for this project. Machine learning is a vital way for predicting the results of a wide range of medical datasets, including data relevant to hypertension. These results may be predicted more accurately using this technology. The goal of this research is to use machine learning to develop a framework within a medical device that employs significant factors that are mostly related with the disease in order to produce accurate predictions regarding hypertension. Despite the fact that data is one of the most significant factors to take into consideration, the classification process is wholly reliant on the information it provides. When data are gathered from a range of sources in a raw format, there is the chance of a number of changes, some of which the model may not be able to account for. There is also the risk that the model may not be able to account for all of the possible changes. As a result of this, preprocessing is proposed as a strategy for getting rid of all the variations and producing an accurate collection of data. This is because of the fact that it can.

K.VijiyaKumar et al., Create a more precise method for early diabetes prediction using the Random Forest algorithm inside a machine learning framework. The prediction of diabetes was proposed as an application for this system. Findings showed that the prediction system could correctly, effectively, and most significantly, quickly forecast diabetes as a disease, and the proposed model provided the best results.

NonsoNnamoko et al., They employed five common classifiers for the ensembles and a meta-classifier to aggregate their outputs in their presentation of an ensemble supervised learning approach for predicting diabetes onset. The goal of this technique was to foretell who will get diabetes. The results are presented and compared to other studies that have used the same data set published in the academic literature. Predictions of diabetes onset are improved using the proposed method.

Tejas N. Joshi et al., Previously, Diabetes Prediction was offered. The goal of this study is to use

supervised machine learning methods to predict diabetes utilizing the support vector machine (SVM), logistic regression, and an artificial neural network (ANN). This research provides support for an approach that has been shown to be successful and feasible in detecting diabetes disease at an earlier stage.

Deeraj Shetty et al, The suggested method for predicting diabetes illness uses data mining to build an Intelligent Diabetes Disease Prediction System that provides an analysis of diabetes disease by making use of a database of diabetes patients. They suggest the usage of algorithms such as Bayesian and KNN (K-Nearest Neighbor) in this system to use on a database of diabetes patients and evaluate them by taking numerous diabetes-related characteristics into consideration for the purpose of making a prediction about diabetes illness.

Muhammad Azeem Sarwar et al., Healthcare cost prediction using machine learning techniques for diabetes In all, six distinct machine learning algorithms were used. Different algorithms are compared and contrasted in terms of their efficiency and accuracy. This study's extensive exploration of machine learning techniques yielded the discovery of the algorithm most successful in diabetes prediction. Using the right classifier on the dataset at hand, researchers are becoming more interested in the topic of Diabetes Prediction in order to teach a computer to decide whether or not a patient has diabetes. Based on prior study efforts, it has been concluded that the classification process has not been greatly improved. Therefore, a system is required since diabetes prediction is a fundamental subject in computing, and this is so that it can deal with the difficulties that have been uncovered via prior study.

Proposed System

In this part, the suggested Framework is related with the diabetes prediction, in which the author makes use of the Data base of persons who have diabetes, as seen in Figure 2.

The author begins by obtaining a database containing information on individuals who have diabetes. Next, the author preprocesses the data by using many distinct classification algorithms, such as NB, SVM, and DT. After that, the author conducts performance on a variety of measures by using the analysis-based precision, and ultimately, the author obtains the result for forecasting diabetes.

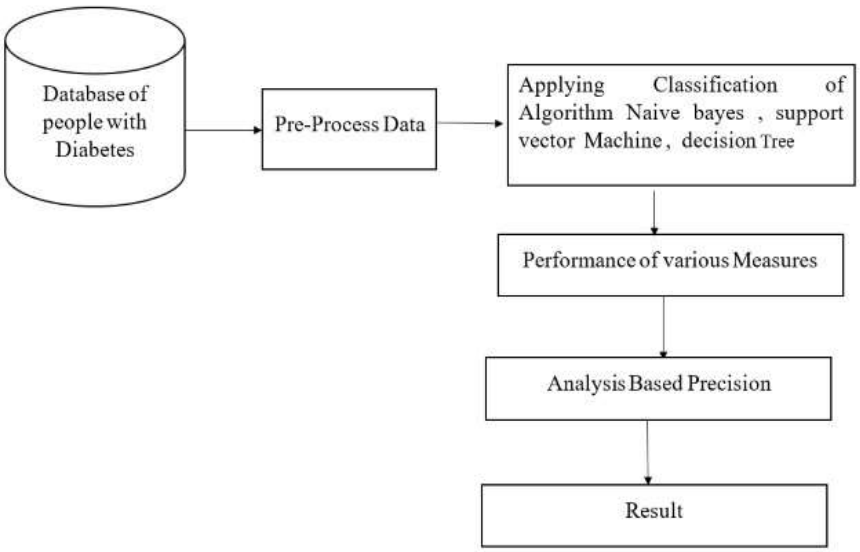

Fig 2: Proposed Frame Work of Diabetes Prediction

Several lines of study on diabetes illness have emerged throughout the years. In the past, a number of researchers carried out a variety of investigations in hospitals and other medical facilities. Some of those researchers used their income to treat diabetic patients, as the treatment of diabetic patients was a set of problems that were prevalent in the early days of research and could only be carried out in hospitals, as opposed to making use of alternative approaches such as machine learning.

The Machine Learning Approach

Our Data Set Will Have Its Design Generated by an Algorithm Based on the Problem Machine Learning will be used to construct the design of our data set. The author of this work has made use of many different methods, including the NB Algorithm, SVM, and DT.

SVM

It is a kind of machine learning called supervised learning. It is a sample of the in-class instruction that was received. SVM is used to locate the straight line that provides the best conclusion between the two classes; nevertheless, this line should not be positioned closer to the data set of the other class. A group of straight lines that are distant from the data set should be picked in each group. The areas that are closest to the support vector machine (SVM) should be positioned at the classify margin. The WEKA Software is used to do an evaluation of the researcher's precision. Growing the gap between the two potential outcomes is one of the steps that the SVM takes to determine the appropriate distance matrix.

Naïve Bayes Algorithm

The data set is categorized using the Naive Bayes Algorithm in accordance with Bayes' rule of probability. It is an approach for optimization based on the assumption that all of the qualities are unconnected to one another and exists in isolation from one another. Additionally, the classification of one feature within a class that does not influence the status of another feature within the same class is specified by this element. It's an algorithm for supervised machine learning, if you were wondering. The categorization is carried out based on the likelihood of an equation that has been presented. It works very well for the knowledge that faces the difficulty of balancing out lacking value and missing information.

Decision Tree Algorithm

In order to categorize the data set, the DT Algorithm, which is a classification algorithm, makes use of the notion of ensemble learning. It is a method of Machine Learning that is supervised, and it is used to solve the classification issue. When used to research, a decision tree's purpose is to make predictions about certain classes by looking at one particular data set utilizing DT. It recognizes and diagnoses conditions based on information from both the node and the internode. The decision-making process is carried out by the tree's roots in a number of different ways, depending on the qualities. The categorization may be seen in the leaf nodes, which may have two or more branches depending on the kind of tree. At each stage, the decision tree determines which traits have the greatest potential for knowledge growth and then uses them to pick the nodes of each tree.

Data set description

This study is being carried out as part of an experiment being carried out by WEKA. WEKA is a Specific Software that was developed by the University of Waikato that Involves the Selection of the Machine Learning Approach for the Data Classification Process Such as Classification, Regression, Clustering, etc. The fact that WEKA may be adapted to suit a user's particular requirements is one of its most important advantages. The major objective of this study is to determine whether or not it is possible to diagnose diabetes in patients by making use of the WEKA tools and the (PIDD) Pima India Diabetes Data Set database.

India Diabetes Data set

An approach that has been recommended and put into action with regard to the diabetes data collection known as PIDD, which the authors obtained from the UCI Repository [31]. The clinical information of 767 different female patients is included in the data collection. It also comprises the numerical values of seven qualities, with one class being given the value '0' if the test for diabetes illness turns out to be negative, and another class being given the value '1' if the test reveals that diabetic disease is present in the patient. It defined the description of the data set by using tables 4 and 5, which represented the explanation of the characteristics.

Result Analysis

Table 1 show that the effectiveness of the algorithm in classifying data varies greatly depending on the measure that is being used. According to Table 1, NB has the best accuracy compared to other categorization models. This can be shown by looking at the results. The Machine Learning technique, in comparison to previous categorization models, is capable of producing more accurate predictions about the danger of insulin. The comparison of the Naive Bayes algorithm, the Support Vector Machine, and the

Decision tree revealed that the Naive Bayes method produces the superior outcome. In this research, three different classification algorithms are utilized to develop a framework that helps in the early-stage prediction of diabetic utilizing key characteristics connected to this disease. This framework is one of the main contributions of this study. The algorithms used are called SVM, Naive Bayes, and DT. The dataset of PIDD that was collected from the UCI database is utilized as the basis for these methods.

Table 1: Performance and Classifications of algorithms

Algorithms	Precision	Recall	Accuracy
Naïve Bayes	0.757	0.761	74.28
Support Vector Machine	0.422	0.649	63.10
Decision Tree	0.733	0.736	71.81

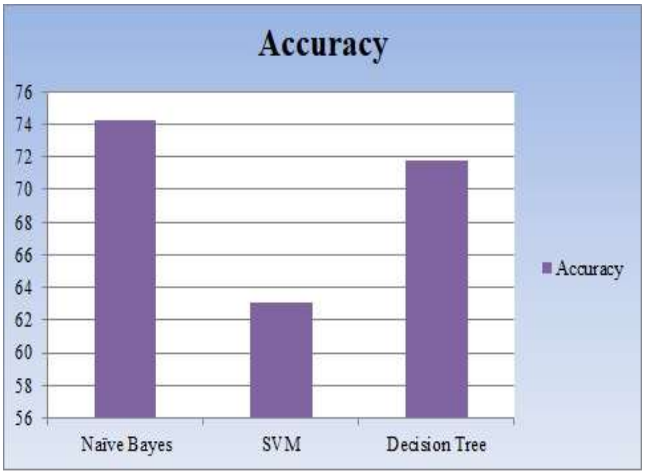

Fig 3: Accuracy

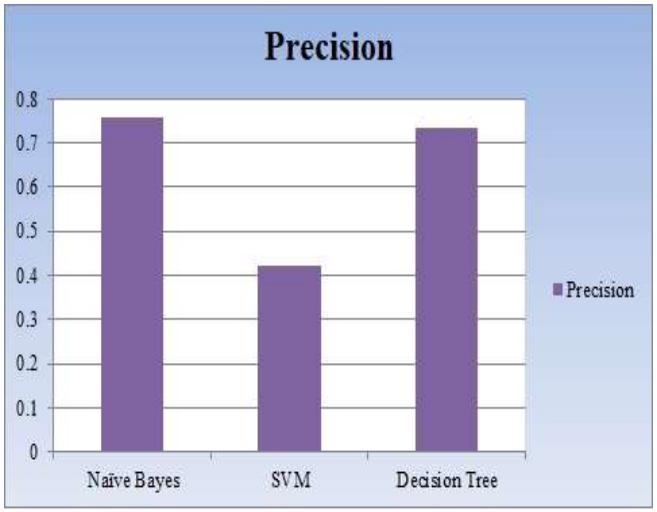

Fig 4: Precision

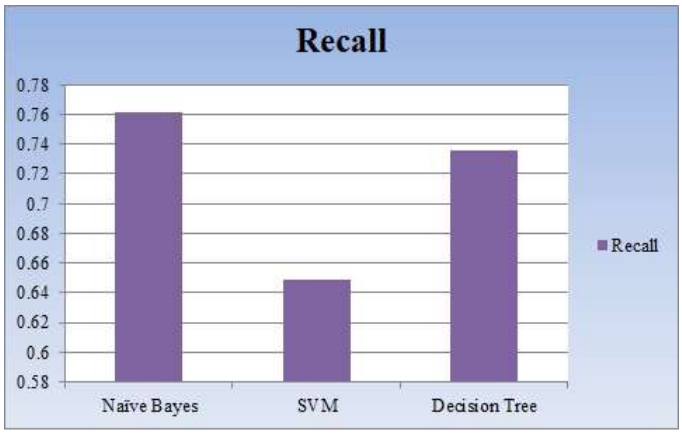

Fig 5: Recall

Conclusion

Researchers working in the area of human health care development face a potentially difficult issue in the form of a significant research difficulty in the early phase diagnosis of diabetes. Diabetes is a real illness, and recognizing its symptoms in its early stages may be difficult at times. This study applies machine learning classification methods to construct a model that is capable of overcoming all issues and is beneficial in the early prediction of diabetes illness. Specifically, the research focuses on the early detection of diabetes disease. In the course of this study, methodical attempts are being made to create a Framework that is capable of diabetes prediction. As part of this study, three distinct categorization algorithms based on the Machine Learning technique will be dissected and evaluated using a variety of metrics. During the course of the trials, the PIDD will be used. The experimental findings, when analyzed using the NB classification algorithm, demonstrate that a developed approach may achieve an accuracy of

74.28 percent, which is sufficient. In the future, the established framework, in conjunction with the many Machine Learning classifiers that were utilized, may be put to use in order to discover or diagnose additional illnesses. The research into diabetes might be expanded and enhanced, in addition to a number of other Machine Learning methodologies. The authors also have plans to conduct classification for other algorithms with missing data.

References

J. Neelaveni and M. S. G. Devasana, "Alzheimer Disease Prediction using Machine Learning Algorithms," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020, pp. 101-104.

V. K. Yarasuri, G. K. Indukuri and A. K. Nair, "Prediction of Hepatitis Disease Using Machine Learning T echnique," 2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (ISMAC), Palladam, India, 2019, pp. 265-269.

M. P. N. M. Wickramasinghe, D. M. Perera and K. A. D. C. P. Kahandawaarachchi, "Dietary prediction for patients with Chronic Kidney Disease (CKD) by considering blood potassium level using machine learning algorithms," 2017 IEEE Life Sciences Conference (LSC), Sydney, NSW, 2017, pp. 300-303.

Maurya, R. Wable, R. Shinde, S. John, R. Jadhav and R. Dakshayani, "Chronic Kidney Disease Prediction and Recommendation of Suitable Diet Plan by using Machine Learning," 2019 International Conference on Nascent Technologies in Engineering (ICNTE), Navi Mumbai, India, 2019, pp. 1-4.

V. Vats, L. Zhang, S. Chatterjee, S. Ahmed, E. Enziama and K. T epe,"A Comparative Analysis of Unsupervised Machine Techniques for Liver Disease Prediction," 2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Louisville, KY, USA, 2018, pp. 486-489.

Gavhane, G. Kokkula, I. Pandya and K. Devadkar, "Prediction of Heart Disease Using Machine Learning," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, 2018, pp. 1275-1278.

S. K. J. and G. S., "Prediction of Heart Disease Using Machine Learning Algorithms.," 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), CHENNAI, India, 2019, pp. 1-5.

R. J. P. Princy, S. Parthasarathy, P. S. Hency Jose, A. Raj Lakshminarayanan and S. Jeganathan, "Prediction of Cardiac Disease using Supervised Machine Learning Algorithms," 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2020, pp. 570-575.

R. Atallah and A. Al-Mousa, "Heart Disease Detection Using Machine Learning Majority Voting Ensemble Method," 2019 2nd International Conference on new T rends in Computing Sciences (ICTCS), Amman, Jordan, 2019, pp. 1-6.

S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019.

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.

Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia and J. Gutierrez, "A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease," 2017 IEEE Symposium on Computers and Communications (ISCC), Heraklion, 2017, pp. 204-207.

Mir and S. N. Dhage, "Diabetes Disease Prediction Using Machine Learning on Big Data of Healthcare," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-6.

M. Patil, V. B. Lobo, P. Puranik, A. Pawaskar, A. Pai and R. Mishra, "A Proposed Model for Lifestyle Disease Prediction Using Support Vector Machine," 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), angalore, 2018, pp. 1-6.

S. R. Alty, S. C. Millasseau, P. J. Chowienczyc and A. Jakobsson, "Cardiovascular disease prediction using support vector machines," 2003 46th Midwest Symposium on Circuits and Systems, Cairo, 2003, pp. 376-379 Vol. 1.

S. Kaur and S. Kalra, "Disease prediction using hybrid K-means and support vector machine," 2016 1st India International Conference on Information Processing (IICIP), Delhi, 2016, pp. 1 -6.

R. S. Raj, D. S. Sanjay, M. Kusuma and S. Sampath, "Comparison of Support Vector Machine and Naïve Bayes Classifiers for Predicting Diabetes," 2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE), Bangalore, India, 2019, pp. 41 -45.

Kaveeshwar SA, Cornwall J. The current state of diabetes mellitus in India. Australas Med J. 2014;7(1):45-48. Published 2014 Jan 31. doi:10.4066/AMJ.2013.1979

Han Wu, Shengqi Yang, Zhangqin Huang, Jian He, XiaoyiWang"Type 2 diabetes mellitus prediction model based on data mining"

Mayo Clinic Q and A: Childhood diabetes, March 31, 2021, 06:00 p.m. CDT

Science Saturday: Could regenerative medicine provide a new approach to diabetes care?, Nov. 28, 2020, 12:00 p.m. CDT

Iancu, I., Mota, M., and Iancu, E. (2008). "Method for the analysing of blood glucose dynamics in diabetes mellitus patients," in Proceedings of the 2008 IEEE International Conference on Automation, Quality and Testing, Robotics, Cluj-Napoca. doi: 10.1109/AQTR.2008.4588883

Cox, M. E., and Edelman, D. (2009). Tests for screening and diagnosis of type 2 diabetes. Clin. Diabetes 27, 132–138. doi: 10.2337/diaclin.27.4.132

American Diabetes Association (2012). Diagnosis and classification of diabetes mellitus. Diabetes Care 35(Suppl. 1), S64– S71. doi: 10.2337/dc12-s064

Lee, B. J., and Kim, J. Y. (2016). Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning. IEEE J. Biomed. Health Inform. 20, 39–46. doi:10.1109/JBHI.2015.2396520

Alghamdi, M., Al-Mallah, M., Keteyian, S., Brawner, C., Ehrman, J., and Sakr, S. (2017). Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: the henry ford exercise testing (FIT) project. PLoS One 12:e0179805. doi: 10.1371/journal.pone.0179805

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., and Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Comput. Struct. Biotechnol. J. 15, 104–116. doi: 10.1016/j.csbj.2016.12.005

Rashid, Ahlam (2020), "Diabetes Dataset", Mendeley Data, V1, doi: 10.17632/wj9rwkp9c2.1

Chao-Ying Joanne Peng, Kuk Lida Lee, Gary M. Ingersoll, "An Introduction to Logistic Regression", Indiana University-Bloomington, September 2002

J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," Journal of Machine Learning Research, vol. 13, pp. 281–305, 2012.