Case Report
Malays. j. med. biol. res.

Automatic Diagnosis of Diabetes Using Machine Learning: A Review

Takudzwa Fadziso[1], Harshini Priya Adusumalli[2*]

[1]Institute of Lifelong Learning and Development Studies, Chinhoyi University of Technology, **ZIMBABWE**
[2]Software Developer, iMinds Technology systems, Inc., 1145 Bower Hill Rd, Pittsburgh, PA, **USA**

[*]Email for Correspondence: harshinipa.gs@gmail.com

ABSTRACT

The health sector, like the other sectors, contains a large amount of data that should be used to better understand and treat the various ailments that are prevalent. For example, diabetes is a condition that is becoming more prevalent but that may be managed if discovered at an early stage. The algorithms of machine learning (ML) can be utilized for this purpose. We have examined the various machine learning methods and the attributes that can be utilized to train these algorithms for the purpose of detecting diabetic complications.

Keywords: Machine Learning, Diabetes, Diabetes Detection, Blood Sugar

Manuscript Received: 25 August 2020 - Revised: 14 October 2020 - Accepted: 28 October 2020

INTRODUCTION

The health sector needs help to solve different issues. A health care system can help this sector to increase its performance and help the people in solving their health related issues. Diabetes is one of the dangerous diseases that people face now a days. This disease very common among the people even the child can also be its victim. This disease is so dangerous that it can lead to the death. This disease is detected through the blood test of a person. A person is considered a diabetes patient if he contains sugar more than the normal value in his blood. According to the researchers (Lukmanto & Irwansyah, 2015) there are two major types of the diabetes.

Type 1

This type of diabetes occurs because insulin is not produced by the liver of a patient. Insulin is used for absorbing the glucose contained in the blood of a person. When there is no insulin, the glucose will be increased in the person's body and it will make him a Type one patient of diabetes. The children can also be the patient of this type of diabetes. The patients of this type feel more thirsty, nerves problems etc. Insulin therapy is a mechanism that is commonly adopted for treating this kind of patient.

Type 2

The patients of this type of diabetes are mostly the persons who are older than 40 years. The lack of exercise can be the major cause for this type. The patients of this type of diabetes are more than the type 1. Almost 90% of diabetes patients are the patients of type 2.

The number of diabetes patients is increasing with the passage of time. According to researchers (Atlas, 2015; Kaul et al., 2013) the patients of diabetes will increase 55% in 2035 and there will be one death because of diabetes in every six to ten seconds. Resistant to insulin can be the cause of diabetes mellitus and this a complex kind of debilitating disease (Beloufa et al., 2013; Thirugnanam et al., 2012; Varma et al., 2014). The discovery of the anti-diabetic drug is a great challenge of the current time (Sakurai et al., 2002). There is a huge amount of data available that is related to this disease

and its patients. It is necessary to utilize this data for the wellbeing of diabetes patients. Various sectors now a days using machine learning (ML) to solve the different issues (Rahman et al., 2019). Researchers (Fadziso et al., 2018) used the algorithms of ML for predicting the stock market price. These algorithms can also be used in the health care sector. For example, researchers (Finkelstein & Cheol Jeong, 2017) used these algorithms for predicting the asthma at an initial stage. Researchers (Kavakiotis et al., 2017) said a huge amount of data is collected related to the patients in the health care sector. This data must be used in an efficient way. They used this data set for the detection of diabetes.

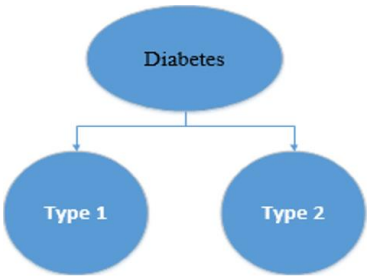

Figure 1: Classification of Diabetes

The purpose of writing this literature review is to discuss the importance of early prediction of diabetes. The importance of machine learning in this regard will also be discussed. We will discuss the different attributes and machine learning algorithms that can be used for detecting the diabetes. The literature review is formulated as: section two contains the Methodology, section three shows the Research Questions, search process is presented in section four, result and discussion is section five, and lastly conclusion is given in the section six.

METHODOLOGY

The SLR (systematic literature review) methodology is used for writing this literature review. This is a widely used methodology for this purpose. The sub parts of the methodology are discussed below.

Inclusion and Exclusion: There are a huge number of papers that are available on different databases such as the IEEE, ACM etc. It is necessary to have a criteria for including the suitable papers. The criteria that we followed for including the papers was based on the language and complete availability of the paper i.e. the papers that were completely available and their language was English, were included and rest of the papers were excluded.

Quality Assessment: For assessing the quality of the papers, the content of the papers were considered. A paper that presents a good content was considered as good quality paper.

RESEARCH QUESTIONS (RQ)

We have created two research questions for this literature review. A detailed discussion will be made on these RQ's in the result and discussion section.

RQ 1: Why machine learning is required for the early detection of diabetes and what are the different attributes that can be used for this purpose?

RQ 2: What are the different machine learning algorithms that can be used for the prediction or detection of diabetes?

SEARCH PROCESS

We followed a search process for collecting the highly relevant papers on the topic. Our search process was based on different steps. In the first step, the papers with not relevant titles were excluded. Second step was excluding the papers whose abstract was not relevant. In the third step, the rest of papers were studied completely and based on the complete content we excluded the papers whose content was not suitable to our topic.

RESULT AND DISCUSSION

Diabetes is one of the major diseases that is faced by different countries of the world. World health organization (WHO) is also doing research for finding the best solution for this disease (Alberti & Zimmet, 1998). In 1970, WHO created criteria and a classification system for the diabetes, with the help of National Diabetes Data Group. With the advancement in technology, it is necessary to make use of technology in the health sector to resolve the issue of diabetes. Obviously, this cannot completely resolve the issue but early detection of the diabetes will help in dealing with diabetes. Machine learning (ML) provides a different algorithms that are useful in the prediction cases. The major aim of the ML is to make the computers powerful to learn from their experience and make predictions (Wilson & Keil,

2001; Sun & applications, 2013; Kaur et al., 2017). Machine learning can be applied in different domains such as search engine, traffic management, gaming, email filtering, disease prediction (Jordan & Mitchell, 2015; Sattigeri et al., 2014; Li & Arandjelovic, 2017). Now a days, ML is using for different purposes in the health sector. These include the diagnosis of liver disease, risk assessment, cancer classification (Libbrecht & Noble, 2015; Kourou et al., 2015). SVM algorithm of machine learning used by the researcher for the diagnosis of liver disease (Hashem & Mabrouk, 2014).

The algorithms of ML are typically categorized into three categories (Russell & Norvig, 2002). These are the supervised, unsupervised, and reinforcement learning. Researchers (Bagherzadeh-Khiabani et al., 2016) developed a logistic model for the prediction of diabetes. They tried to find the most suitable predictive attributes for this purpose. The more accurate predictive attributes help in achieving the good accuracy. Researchers (Georga et al., 2015; Robnik-Šikonja & Kononenko, 2003; Adusumalli, 2018) used the random forest for finding the best diabetes predictive attributes. ML is used by the researchers for the classification of diabetes (Roychowdhury et al., 2013).

In the healthcare sector, the development of any diagnosis software requires the software should also be able to perform the prediction. Diabetes is a disease that can affect the entire body of the patient (Kaur et al., 2012). Researchers (Priya & Aruna, 2013; Pasupuleti et al., 2019) used the ML algorithms for detecting the eye disease that is caused by the diabetes. It is necessary to predict or detect these kinds of diseases at an early stage.

RQ 1 : "Why machine learning is required for the early detection of diabetes and what are the different attributes that can be used for this purpose?"

In almost every sector, a huge amount of data is available. This data is collected electronically so it is easy to apply the ML algorithms for different purposes. The method of keeping the records manually is almost finished in each sector. Because the use of databases and computers makes it easy to manage and collect the data. In this way, each sector has a huge volume of data and ML can be applied to this data for performing the different tasks such as the prediction or detection. In the health sector, one of the major issues is the early detection of the diabetes so that doctors can start the treatment early. For this purpose the ML algorithms can be trained with the data set available in the hospitals. Researchers (Pasupuleti, 2017) used the ML algorithms for the early detection of the Type 2 diabetes. This is the most common type of the diabetes as the 90% of diabetes patients contain this type of diabetes. Various researchers used the ML for detecting the diabetes at an initial stage to prevent the different issues that can be caused by the diabetes (Gittens et al., 2014; Choubey et al., 2016; Wu et al., 2018). It is necessary to predict the diabetes early so that the treatment can be started. For this purpose, it is necessary to check the medical record, health information of the persons and it is very time consuming if it is performed in a manual way (Rubaiat et al., 2018). A large amount of data is difficult to handle through the traditional methods (Nguyen et al., 2015; Adusumalli, 2019; Sanz Delgado et al., 2013). That's why it is necessary to use the ML for this purpose. But the challenging task in the training of ML algorithms is the use of correct features (Wang et al., 2015). The use of most accurate features for the training purpose can be beneficial in terms of less time consumption and more prediction accuracy (Guyon & Elisseeff, 2003; Frank & Hall, 2011; Sideris et al., 2016). Different features or attributes are given below that are used by the researchers for the training and detection of diabetes.

Table 1: Attributes for diabetes detection

Polydipsia	Polyphagia
Polyuria	Age
2-hour postprandial blood glucose level	Fasting blood glucose level

RQ 2: What are the different machine learning algorithms that can be used for the prediction or detection of diabetes?

Machine learning has made the different things easy in the various sectors. The issue of diabetes detection and prediction can be handled through the algorithms of ML. Researchers (Lee et al., 2015) used the ML for the detection of different risk factors that are associated with the type 2 of diabetes. Researchers (Aslam, Zhu, & Nandi, 2013; Pasupuleti & Amin, 2018) used the genetic programming for the classification of diabetes. The models that help in the prediction or assessment of risk are popular in the health sector (Oh et al., 2016). Researcher (Pasupuleti & Adusumalli, 2018) used the ML approaches for the detection of diabetes. Various researchers used the ML for detecting the diabetes (Habibi et al., 2015; Razavian et al., 2015; Meng et al., 2013; Anderson et al., 2016; Anderson et al., 2016; Ozcift et al., 2011; Adusumalli & Pasupuleti, 2017). Some of the most popular ML algorithms that are used for detecting the diabetes are given below.

Logistic Regression: Researchers (Devi et al., 2016) used this algorithm for predicting the diabetes. They modified this algorithm for increasing the accuracy. The previous accuracy was 79% and they improved it to 90.4%.

Random Forest: Random forest is used by the researchers (Pasupuleti, 2018) for detecting the type of diabetes. Researchers (Benbelkacem & Atmani, 2019) also used this algorithm and compared it with other methods of ML. On the basis of their results, they announced this algorithm as a better algorithm for diabetes detection.

SVM: Researchers (Kumari et al., 2013) said the number of diabetes patients is increasing continuously. It is necessary to use an automated way for detecting the diabetes. They used this algorithm for this purpose. Researchers (Carrera et al., 2017) used this algorithm for early detection of the eye disease caused by the diabetes.

Naïve Bayes: This algorithm is used by the researchers (Adusumalli, 2017) as a novel approach for predicting the diabetes at an initial stage. They developed the web interface to show the prediction on the basis of different given values such as the age, insulin level etc.

The diagram given below shows the steps that can be followed to get a prediction result from any of these ML algorithms.

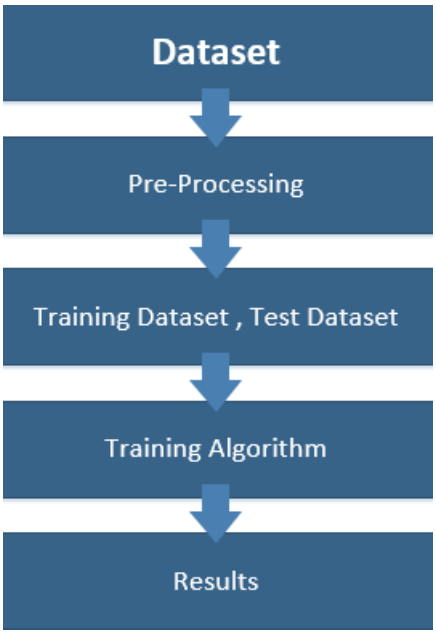

Figure 2: Algorithm Training

CONCLUSION

Diabetes is a disease that is continuously increasing in developed and developing countries. The efficient solution to deal with this disease is need of the time. The health care sector can use the ML for the detection of this disease early so that the doctors can help the patient to recover from this disease. This disease can also be the reason for different other diseases that can also be detected through the ML. In the modern time, it is necessary to use modern tools to deal with the increasing disease.

REFERENCES

Adusumalli, H. P. (2017). Software Application Development to Backing the Legitimacy of Digital Annals: Use of the Diplomatic Archives. *ABC Journal of Advanced Research*, *6*(2), 121-126. https://doi.org/10.18034/abcjar.v6i2.618

Adusumalli, H. P. (2018). Digitization in Agriculture: A Timely Challenge for Ecological Perspectives. *Asia Pacific Journal of Energy and Environment*, *5*(2), 97-102. https://doi.org/10.18034/apjee.v5i2.619

Adusumalli, H. P. (2019). Expansion of Machine Learning Employment in Engineering Learning: A Review of Selected Literature. *International Journal of Reciprocal Symmetry and Physical Sciences*, *6*, 15–19. Retrieved from https://upright.pub/index.php/ijrsps/article/view/65

Adusumalli, H. P., & Pasupuleti, M. B. (2017). Applications and Practices of Big Data for Development. *Asian Business Review*, *7*(3), 111-116. https://doi.org/10.18034/abr.v7i3.597

Alberti, K. G. M. M., & Zimmet, P. Z. J. D. m. (1998). Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus. Provisional report of a WHO consultation. *15*(7), 539-553.

Anderson, A. E., Kerr, W. T., Thames, A., Li, T., Xiao, J., & Cohen, M. S. J. J. o. b. i. (2016). Electronic health record phenotyping improves detection and screening of type 2 diabetes in the general United States population: a cross-sectional, unselected, retrospective study. *60*, 162-168.

Anderson, J. P., Parikh, J. R., Shenfeld, D. K., Ivanov, V., Marks, C., Church, B. W., . . . technology. (2016). Reverse engineering and evaluation of prediction models for progression to type 2 diabetes: an application of machine learning using electronic health records. *10*(1), 6-18.

Aslam, M. W., Zhu, Z., & Nandi, A. K. J. E. S. w. A. (2013). Feature generation using genetic programming with comparative partner selection for diabetes classification. *40*(13), 5402-5412.

Atlas, D. J. I. D. A., 7th edn. Brussels, Belgium: International Diabetes Federation. (2015). International diabetes federation.

Bagherzadeh-Khiabani, F., Ramezankhani, A., Azizi, F., Hadaegh, F., Steyerberg, E. W., & Khalili, D. J. J. o. c. e. (2016). A tutorial on variable selection for clinical prediction models: feature selection methods in data mining could improve the results. *71*, 76-85.

Beloufa, F., Chikh, M. A. J. C. m., & biomedicine, p. i. (2013). Design of fuzzy classifier for diabetes disease using Modified Artificial Bee Colony algorithm. *112*(1), 92-103.

Benbelkacem, S., & Atmani, B. (2019). *Random forests for diabetes diagnosis.* Paper presented at the 2019 International Conference on Computer and Information Sciences (ICCIS).

Carrera, E. V., González, A., & Carrera, R. (2017). *Automated detection of diabetic retinopathy using SVM.* Paper presented at the 2017 IEEE XXIV international conference on electronics, electrical engineering and computing (INTERCON).

Choubey, D. K., Paul, S. J. I. J. o. I. S., & Applications. (2016). GA_MLP NN: a hybrid intelligent system for diabetes disease diagnosis. *8*(1), 49.

Devi, M. N., alias Balamurugan, A., Kris, M. R. J. I. J. o. S., & Technology. (2016). Developing a modified logistic regression model for diabetes mellitus and identifying the0 important factors of type II DM. *9*(4), 1-8.

Fadziso, T., Adusumalli, H. P., & Pasupuleti, M. B. (2018). Cloud of Things and Interworking IoT Platform: Strategy and Execution Overviews. *Asian Journal of Applied Science and Engineering*, *7*, 85–92. Retrieved from https://upright.pub/index.php/ajase/article/view/63

Finkelstein, J., & cheol Jeong, I. J. A. o. t. N. Y. A. o. S. (2017). Machine learning approaches to personalize early prediction of asthma exacerbations. *1387*(1), 153.

Frank, E., & Hall, M. A. (2011). *Data mining: practical machine learning tools and techniques*: Morgan Kaufmann.

Georga, E. I., Protopappas, V. C., Polyzos, D., Fotiadis, D. I. J. M., engineering, b., & computing. (2015). Evaluation of short-term predictors of glucose concentration in type 1 diabetes combining feature ranking with regression models. *53*(12), 1305-1318.

Gittens, M., King, R., Gittens, C., & Als, A. (2014). *Post-diagnosis management of diabetes through a mobile health consultation application.* Paper presented at the 2014 IEEE 16th International Conference on e-Health Networking, Applications and Services (Healthcom).

Guyon, I., & Elisseeff, A. J. J. o. m. l. r. (2003). An introduction to variable and feature selection. *3*(Mar), 1157-1182.

Habibi, S., Ahmadi, M., & Alizadeh, S. J. G. j. o. h. s. (2015). Type 2 diabetes mellitus screening and risk factors using decision tree: results of data mining. *7*(5), 304.

Hashem, E. M., & Mabrouk, M. S. J. A. J. o. I. S. (2014). A study of support vector machine algorithm for liver disease diagnosis. *4*(1), 9-14.

Jordan, M. I., & Mitchell, T. M. J. S. (2015). Machine learning: Trends, perspectives, and prospects. *349*(6245), 255-260.

Kaul, K., Tarr, J. M., Ahmad, S. I., Kohner, E. M., & Chibber, R. J. D. (2013). Introduction to diabetes mellitus. 1-11.

Kaur, H., Chauhan, R., & Ahmed, Z. J. B. H. S. R. (2012). Role of data mining in establishing strategic policies for the efficient management of healthcare system–a case study from Washington DC area using retrospective discharge data. *12*(1), 1-2.

Kaur, H., Lechman, E., & Marszk, A. J. T. D. W. E. (2017). Catalyzing development through ICT adoption. *4*.

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., Chouvarda, I. J. C., & journal, s. b. (2017). Machine learning and data mining methods in diabetes research. *15*, 104-116.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., Fotiadis, D. I. J. C., & journal, s. b. (2015). Machine learning applications in cancer prognosis and prediction. *13*, 8-17.

Kumari, V. A., Chitra, R. J. I. J. o. E. R., & Applications. (2013). Classification of diabetes disease using support vector machine. *3*(2), 1797-1801.

Lee, B. J., Kim, J. Y. J. I. j. o. b., & informatics, h. (2015). Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning. *20*(1), 39-46.

Li, J., & Arandjelovic, O. (2017). *Glycaemic index prediction: a pilot study of data linkage challenges and the application of machine learning.* Paper presented at the 2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI).

Libbrecht, M. W., & Noble, W. S. J. N. R. G. (2015). Machine learning applications in genetics and genomics. *16*(6), 321-332.

Lukmanto, R. B., & Irwansyah, E. J. P. C. S. (2015). The early detection of diabetes mellitus (DM) using fuzzy hierarchical model. *59*, 312-319.

Meng, X.-H., Huang, Y.-X., Rao, D.-P., Zhang, Q., & Liu, Q. J. T. K. j. o. m. s. (2013). Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. *29*(2), 93-99.

Nguyen, T., Khosravi, A., Creighton, D., & Nahavandi, S. J. E. S. w. A. (2015). Classification of healthcare data using genetic fuzzy logic system and wavelets. *42*(4), 2184-2197.

Oh, W., Kim, E., Castro, M. R., Caraballo, P. J., Kumar, V., Steinbach, M. S., & Simon, G. J. J. B. d. (2016). Type 2 diabetes mellitus trajectories and associated risks. *4*(1), 25-30.

Ozcift, A., Gulten, A. J. C. m., & biomedicine, p. i. (2011). Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms. *104*(3), 443-451.

Pasupuleti, M. B. (2017). AMI Data for Decision Makers and the Use of Data Analytics Approach. *Asia Pacific Journal of Energy and Environment*, *4*(2), 65-70. https://doi.org/10.18034/apjee.v4i2.623

Pasupuleti, M. B. (2018). The Application of Machine Learning Techniques in Software Project Management- An Examination. *ABC Journal of Advanced Research*, *7*(2), 113-122. https://doi.org/10.18034/abcjar.v7i2.626

Pasupuleti, M. B., & Adusumalli, H. P. (2018). Digital Transformation of the High-Technology Manufacturing: An Overview of Main Blockades. *American Journal of Trade and Policy*, *5*(3), 139-142. https://doi.org/10.18034/ajtp.v5i3.599

Pasupuleti, M. B., & Amin, R. (2018). Word Embedding with ConvNet-Bi Directional LSTM Techniques: A Review of Related Literature. *International Journal of Reciprocal Symmetry and Physical Sciences*, *5*, 9–13. Retrieved from https://upright.pub/index.php/ijrsps/article/view/64

Pasupuleti, M. B., Miah, M. S., & Adusumalli, H. P. (2019). IoT for Future Technology Augmentation: A Radical Approach. *Engineering International*, *7*(2), 105-116. https://doi.org/10.18034/ei.v7i2.601

Priya, R., & Aruna, P. J. I. J. o. s. c. (2013). Diagnosis of diabetic retinopathy using machine learning techniques. *3*(4), 563-575.

Rahman, M. M., Pasupuleti, M. B., & Adusumalli, H. P. (2019). Advanced Metering Infrastructure Data: Overviews for the Big Data Framework. *ABC Research Alert*, *7*(3), 159-168. https://doi.org/10.18034/abcra.v7i3.602

Razavian, N., Blecker, S., Schmidt, A. M., Smith-McLallen, A., Nigam, S., & Sontag, D. J. B. D. (2015). Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. *3*(4), 277-287.

Robnik-Šikonja, M., & Kononenko, I. J. M. l. (2003). Theoretical and empirical analysis of ReliefF and RReliefF. *53*(1), 23-69.

Roychowdhury, S., Koozekanani, D. D., Parhi, K. K. J. I. j. o. b., & informatics, h. (2013). DREAM: diabetic retinopathy analysis using machine learning. *18*(5), 1717-1728.

Rubaiat, S. Y., Rahman, M. M., & Hasan, M. K. (2018). *Important feature selection & accuracy comparisons of different machine learning models for early diabetes detection.* Paper presented at the 2018 International Conference on Innovation in Engineering and Technology (ICIET).

Russell, S., & Norvig, P. (2002). Artificial intelligence: a modern approach.

Sakurai, H., Kojima, Y., Yoshikawa, Y., Kawabe, K., & Yasui, H. J. C. C. R. (2002). Antidiabetic vanadium (IV) and zinc (II) complexes. *226*(1-2), 187-198.

Sanz Delgado, J. A., Galar Idoate, M., Jurío Munárriz, A., Brugos Larumbe, A., Pagola Barrio, M., & Bustince Sola, H. J. A. S. C. (2013). Medical diagnosis of cardiovascular diseases using an interval-valued fuzzy rule-based classification system.

Sattigeri, P., Thiagarajan, J. J., Shah, M., Ramamurthy, K. N., & Spanias, A. (2014). *A scalable feature learning and tag prediction framework for natural environment sounds.* Paper presented at the 2014 48th Asilomar Conference on Signals, Systems and Computers.

Sideris, C., Pourhomayoun, M., Kalantarian, H., Sarrafzadeh, M. J. C. i. b., & medicine. (2016). A flexible data-driven comorbidity feature extraction framework. *73*, 165-172.

Sun, S. J. N. c., & applications. (2013). A survey of multi-view machine learning. *23*(7), 2031-2038.

Thirugnanam, M., Kumar, P., Srivatsan, S. V., & Nerlesh, C. J. P. e. (2012). Improving the prediction rate of diabetes diagnosis using fuzzy, neural network, case based (FNC) approach. *38*, 1709-1718.

Varma, K. V., Rao, A. A., Lakshmi, T. S. M., Rao, P. N. J. C., & Engineering, E. (2014). A computational intelligence approach for a better diagnosis of diabetic patients. *40*(5), 1758-1765.

Wang, K.-J., Adrian, A. M., Chen, K.-H., & Wang, K.-M. J. J. o. b. i. (2015). An improved electromagnetism-like mechanism algorithm and its application to the prediction of diabetes mellitus. *54*, 220-229.

Wilson, R. A., & Keil, F. C. (2001). *The MIT encyclopedia of the cognitive sciences*: MIT press.

Wu, H., Yang, S., Huang, Z., He, J., & Wang, X. J. I. i. M. U. (2018). Type 2 diabetes mellitus prediction model based on data mining. *10*, 100-107.

--0--