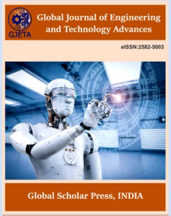
Global Journal of Engineering and Technology Advances

eISSN: 2582-5003

Cross Ref DOI: 10.30574/gjeta

Journal homepage: https://gjeta.com/



(Research Article)

Early diagnosing diabetes using data mining algorithms

Rasha Rokan Ismail *

Department of Computer Science, Diyala University President, Diyala, Iraq.

Global Journal of Engineering and Technology Advances, 2023, 16(02), 106–113

Publication history: Received on 11 June 2023; revised on 17 July 2023; accepted on 20 July 2023

Article DOI: https://doi.org/10.30574/gjeta.2023.16.2.0141

Abstract

Diabetes has become a widespread and long lasting condition that continues to impact an increasing number of individuals across the globe. It is crucial to highlight the significance of accurately identifying, predicting, managing and treating diabetes in order to address this growing concern. Utilizing sophisticated data analysis techniques to examine data relating to diabetes can significantly enhance the early detection and prediction of this ailment, along with its associated complications like low or high blood sugar levels. The findings clearly demonstrate that the decision tree algorithm proves to be the most effective approach in promptly diagnosing diabetes patients and ensuring they receive timely access to suitable treatment options.

Keywords: Diabetes; SVM; TREE; Regression

1. Introduction

Diabetes is a disease where glucose, also known as blood sugar, is not processed correctly, leading to dangerously high blood sugar levels, known medically as hyperglycemia. This condition occurs due to inadequate production of insulin in the body or a weakened response to the insulin produced. While diabetes cannot be cured, it can be effectively managed. It is crucial to be aware of the potential complications associated with this condition, including heart attack, stroke, kidney failure, and nerve damage. Surprisingly, in 2017, 8.8% of the global population had diabetes, and this number is projected to increase to 9.9% by 2045. Diabetes-induced hyperglycemia significantly impacts cardiovascular health [1].

Due to its high occurrence in both adults and children and the significant cost of healthcare and treatment, diabetes has emerged as a major global public health issue. According to [2, 3], over 415 million adults worldwide, accounting for 8.8% of the adult population, had diabetes in 2015, and this number is projected to rise to over 642 million by 2040. Moreover, during this period, the disease has caused nearly 5 million deaths and affected over 500,000 children. Conversely, the financial burden of diabetes globally was approximately USD 673 billion in 2015, with an estimated increase to USD 802 billion by 2040 [3]. The practice of self-monitoring blood glucose (SMBG) through finger stick blood samples has gained widespread acceptance in diabetes management since its introduction three decades ago [4, 5]. Diabetic individuals utilize finger-stick glucose meters as an invasive method to monitor their blood glucose levels, involving piercing their finger's skin three to four times daily. The main objective is to consistently monitor levels of blood sugar and make any necessary adjustments to insulin dosage, diet, and activity in order to maintain healthy levels of blood sugar. However, relying solely on a limited number of self-monitoring blood glucose (SMBG) tests to determine the appropriate amount of insulin can be burdensome and inconvenient, which can result in inaccurate outcomes. Plasma glucose levels can sometimes surpass the range but thanks, to glucose monitoring (CGM) we have made remarkable progress in comprehending the fluctuations, in blood glucose. This valuable data empowers individuals living with diabetes to make informed choices regarding their treatment. CGM employs either a device or system to monitor blood sugar levels offering minimally invasive or entirely non invasive alternatives. Additionally we have both

[*] Corresponding author: Rasha Rokan Ismail

time and retrospective CGM systems each offering advantages depending on the circumstances. A study used the SVM classification method to evaluate six variations.

Advanced and persuasive approaches to effectively manage diabetes [7]. This extensive investigation has provided us with the understanding that the most efficient and prosperous approach to quitting smoking is to completely cease smoking. Additionally, the pioneering technique developed in Ref. [8] utilizes data-driven models to anticipate blood glucose levels.This model incorporates a variety of free-living data, including information on food, activity, and CGM. This approach involves examining the effects of diet, physical activity, and medication on glucose regulation. The methodology incorporates several models, such as the meal model, exercise model, insulin model, and glucose prediction model (SVR), all of which are based on support vector regression. Promising results have been obtained by studying data from seven individuals with type 1 diabetes (CGM, activity insulin, etc.) for 15 and 30 minute predictions [11]. Presenting a study that aims to apply various data mining techniques to healthcare applications by employing different tools for various types of illnesses commonly experienced by many people. In the field of healthcare, algorithms and procedures play a crucial role in diagnosing and predicting illnesses. Nave Bayes, association rule mining, clustering, and classification are the mining techniques used on health data. By utilizing these techniques to predict various diseases, it was demonstrated that the accuracy for predicting cancer reached approximately 97.77% [12].

Taking into consideration the 249 instances of individuals with diabetes and their seven unique qualities, the WEKA tool was utilized on the dataset to implement algorithms such as the Bayes network classifier, J48 Pruned tree, REP tree, and Random forest. This study aimed to determine the prevalence of the condition and raise awareness about the increasing number of people globally who are affected by it. Diabetes mellitus (DM), a group of metabolic disorders, is defined by abnormal insulin production as its primary cause [14]. Insufficient insulin levels lead to hyperglycemia, which is characterized by elevated blood glucose levels and disrupts the metabolism of proteins, lipids, and carbohydrates. DM, a prevalent endocrine disorder, has a global impact on over 200 million individuals and is expected to increase in incidence in the coming years. The Prison Ace (DM) Craftsmanship talks about the distinctive sorts of diabetes, to be specific Sort 1 Diabetes (T1D) and Sort 2 Diabetes (T2D), each having its possess unmistakable clinical introduction. T2D is capable for around 90% of diabetes cases and is characterized by affront resistance. On the opposite, T1D is accepted to happen due to the immune system devastation of the islets of Langerhans, which contain beta cells, whereas T2D is impacted by variables like way of life, physical action, and dietary choices. There are moreover other shapes of diabetes, counting gestational diabetes, endocrine infection, MODY (diabetes of the youthful), diabetes of the infant, mitochondrial diabetes, and diabetes of the infant, which vary in terms of affront discharge and/or onset. Common side effects of diabetes incorporate over the top thirst, noteworthy weight misfortune, and visit urination. Conclusion is based on blood glucose levels (fasting glucose = 7.0 mmol/L) [15]. Information mining, moreover known as information disclosure in databases (KDD), may be a concept that utilizes progressed innovations like counterfeit insights, machine learning, insights, and database frameworks to reveal designs in different datasets [16]. These strategies encourage the recognizable proof and investigation of information designs, making a difference unravel issues such as classification, affiliation, and forecast [17, 18, 19].

The utilization of progressed procedures like machine learning and information mining enormously upgrades the forms of examination, elucidation, and information mining. This is often of most extreme noteworthiness when considering the early discovery of sort 2 diabetes mellitus (T2DM) [20,21]. These progressed procedures have gigantic potential in recognizing and foreseeing illnesses like T2DM, which can have a noteworthy affect on an individual's generally well-being [21,22]. In our inquire about, we present four uncommon calculations for diagnosing diabetes, to be specific choice trees, calculated relapse, bolster vector machines, and arbitrary timberlands.

2. Material and methods

This elegant and persuasive study employed a retrospective approach, utilizing a comprehensive database consisting of 769 individuals who resided during the period of 2019 and 2020. These patients were referred as part of their usual medical care, ensuring a representative sample. To analyze the data, advanced statistical techniques were employed, including the forward stepwise logistic regression analysis (LR), decision tree analysis (DT), Random Forest, and support vector machine. Each of these approaches was assessed by calculating the Receiver Operating Characteristic (ROC) to gauge their effectiveness. Furthermore, the dataset was meticulously divided into two distinct groups: one dedicated to model creation, accounting for 70 percent, and the remaining 30 percent allocated to the validation group. This meticulous division allowed for reliable and robust results.

2.1. Data Analysis

In a state of utmost comfort, the electrocardiograms (ECG) of 769 individuals with diabetes and 769 individuals without diabetes were meticulously observed for a duration of 10 minutes. Employing the esteemed Pan and Tompkins approach, the ECG signals were meticulously processed, yielding a wealth of heart rate time series data. The current methodology ingeniously utilizes a real-time strategy to skillfully identify QRS complexes in the ECG signal by relying on the distinctive characteristics of amplitude, slope, and width. To enhance the sensitivity of detection and minimize the occurrence of false detections caused by unwanted noise, sophisticated techniques such as thresholding operations and digital bandpass filtering are adeptly employed. The remarkable findings are elegantly presented in the captivating pair plot illustrated in Figure 2.

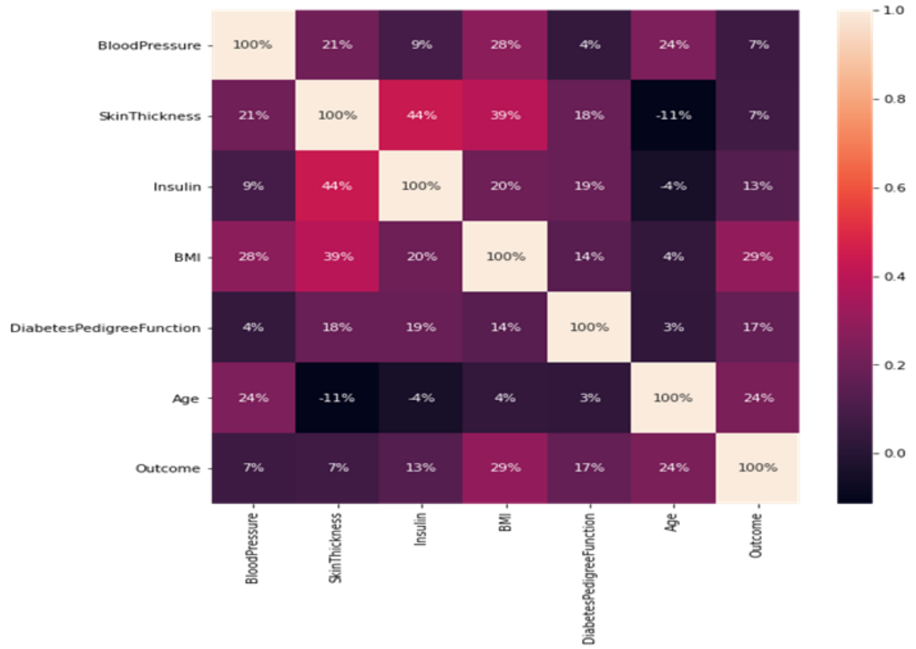


Figure 1 Diabetes correlation matrix

It is worth noting that linear separation is not always feasible in all scenarios. Factors such as noise in observations or nonlinearities in data generation mechanisms can obstruct linear separation. However, SVMs overcome this limitation by allowing a few misclassified data points to exist within the margin while still successfully determining a maximum margin separating hyperplane. Moreover, SVMs have the capability to employ non-linear curves for dividing data into distinct groups. This is accomplished by nonlinearly transforming the input data to a new space, where a linear hyperplane can still be employed for segregation. Upon mapping the separating hyperplane back to the original space, it manifests as a curve, as depicted in Figure 3. Support Vector Machines (SVMs) have gained widespread popularity as an effective method for data modeling. They offer a unique solution to the challenge of high dimensionality while also ensuring accurate generalization. By identifying a hyperplane that maximizes the margin between two classes, SVMs achieve a clear separation. This hyperplane is strategically positioned equidistant from the two classes, optimizing the margin. Furthermore, the SVM algorithm naturally identifies support vectors, which are crucial data points responsible for the solution. Remarkably, even if only the support vectors were considered as inputs, the same hyperplane would still be obtained..

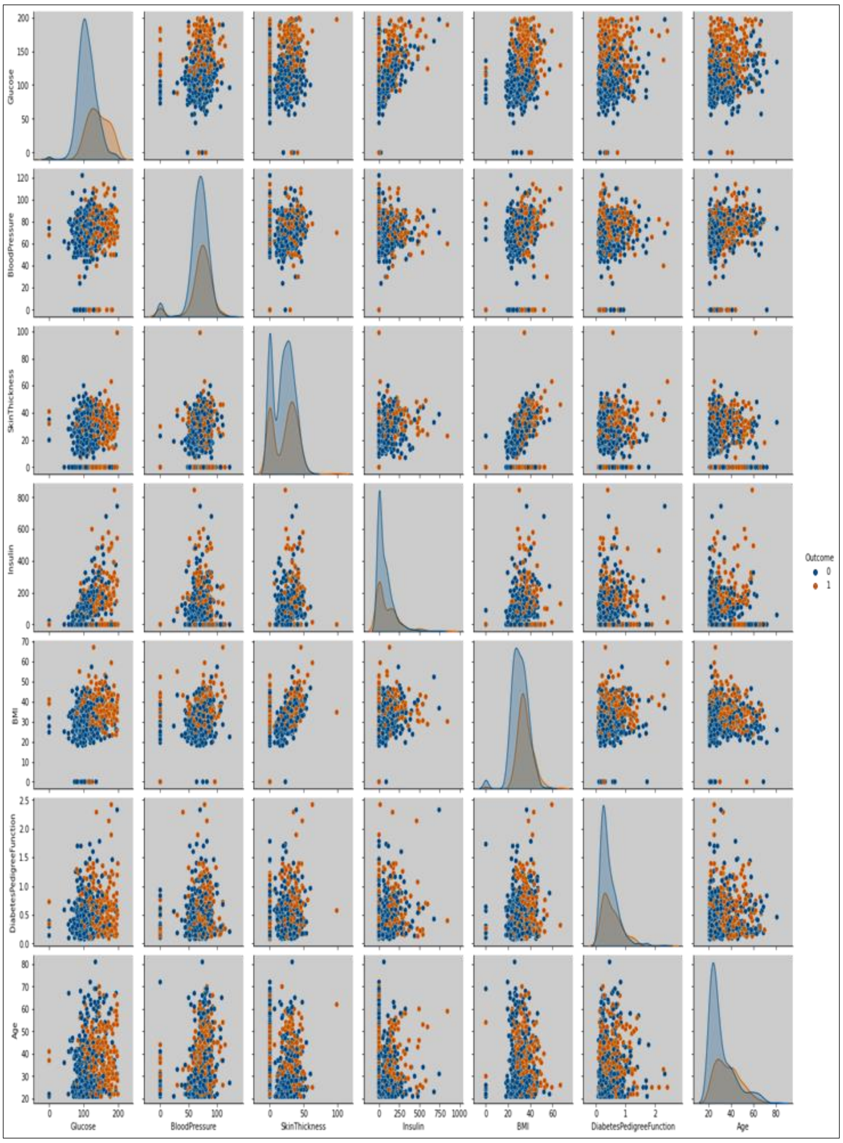


Figure 2 The Pair Plot

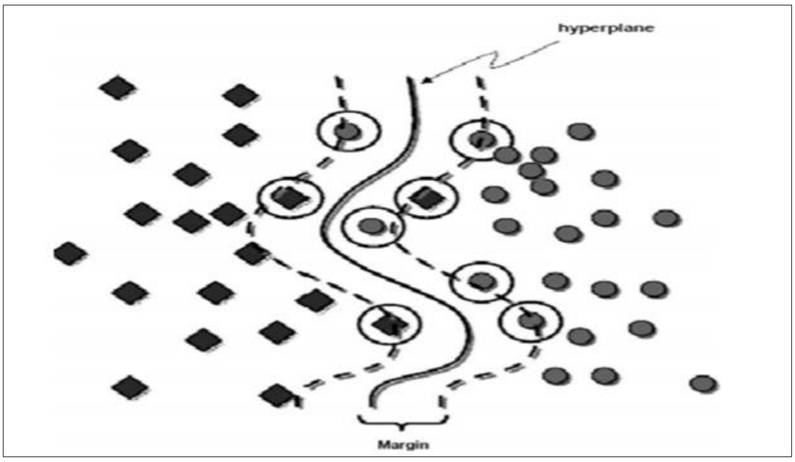


Figure 3 Classification of data using non-linear curves

Logistic Regression, as depicted in Figure 4 through a sigmoid function, embodies an exquisite approach. By elegantly merging input values with carefully assigned weights or coefficients, it unveils its ability to forecast the output value. This remarkable technique focuses on binary variables, gracefully navigating between the values of 0 and 1. Its ultimate objective lies in formulating a refined mathematical equation, capable of delivering a comprehensive score within the refined range of 0 to 1.

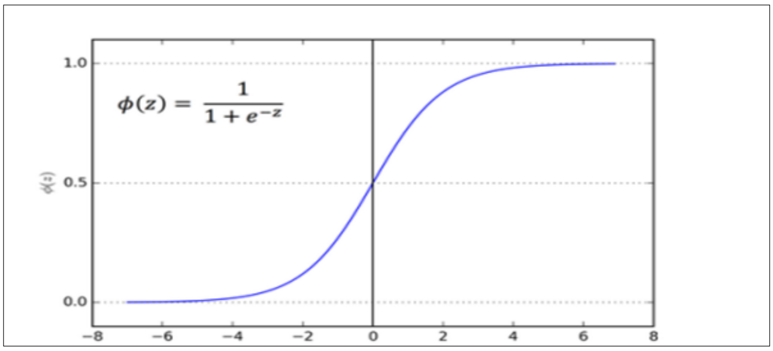


Figure 4 Logistic Regression [19]

Prior to constructing the class that embodies the collective average of the classes or predictions generated by individual trees, random forests, also known as random decision forests, undergo the training process involving a substantial number of decision trees. These forests are utilized in various applications including classification, regression, and other problem-solving scenarios as a method of ensemble learning. The issue of decision trees overfitting their training set is effectively addressed by the implementation of random decision forests. Although they may not be as precise as gradient enhanced trees, random forests consistently outperform individual choice trees. However, it is worth noting that the effectiveness of random forests can be influenced by the quality of the data used. In order to create predictive models for four distinct algorithm types, we have developed the BCR prediction system in Python.

3. Result and Discussion

The results for each algorithm as shown in table 1.

Table 1 Performance of Algorithms

Analysis	**Method**	**Accuracy**	**Sensitivity**
Decision tree	Decision tree	1	1
Decision tree	Random Forest	0.99	0.95
Support Vector Machine	Linear	0.77	66.8

Based on the findings from the analysis presented in Table 1, it is evident that the decision tree algorithm achieved the highest AUC and accuracy values, both reaching a perfect score of 1. Conversely, the logistic regression algorithm exhibited comparatively lower values, with an AUC value of 0.33 and an accuracy value of 0.2933.

As depicted in figure (5, 6), the ROC curve, also known as the receiver operating characteristic curve, visually represents the false positive rate (FPR) of a classification system in an elegant manner. By calculating the ratio of false positives to the total number of test results, the FPR is determined. Within a typical ROC figure, the y axis denotes the true positive rate (TPR), while the x axis signifies the false positive rate (FPR). Leveraging a ROC curve allows us to comprehensively examine the accuracy and bias of a specific classification system, thereby facilitating the evaluation of model performance. Notably, on the ROC plot, we observe that the model's accuracy improves as it progressively moves away from the diagonal line. In the top left-hand corner of the figure, we can witness the ideal scenario where a flawless classifier achieves a TPR of 1 and an FPR of 0, exemplifying its exceptional performance.

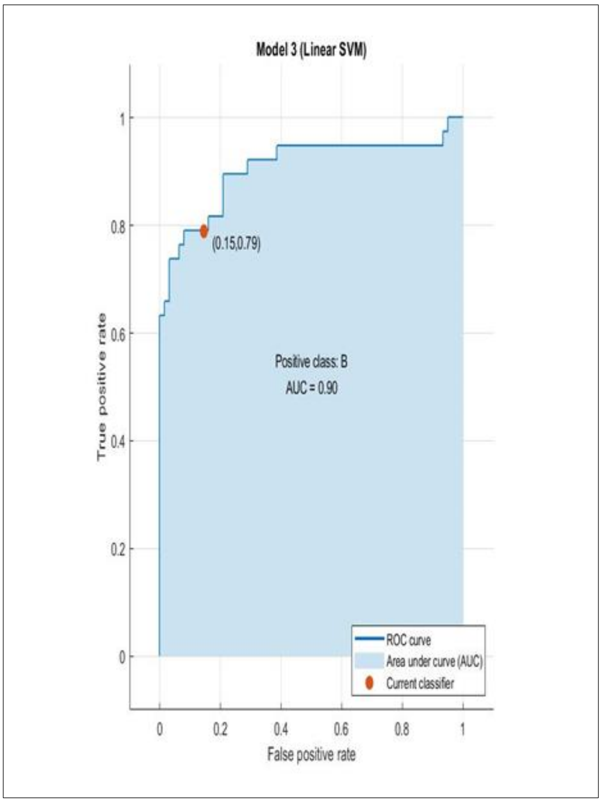


Figure 5 ROC curve for SVM

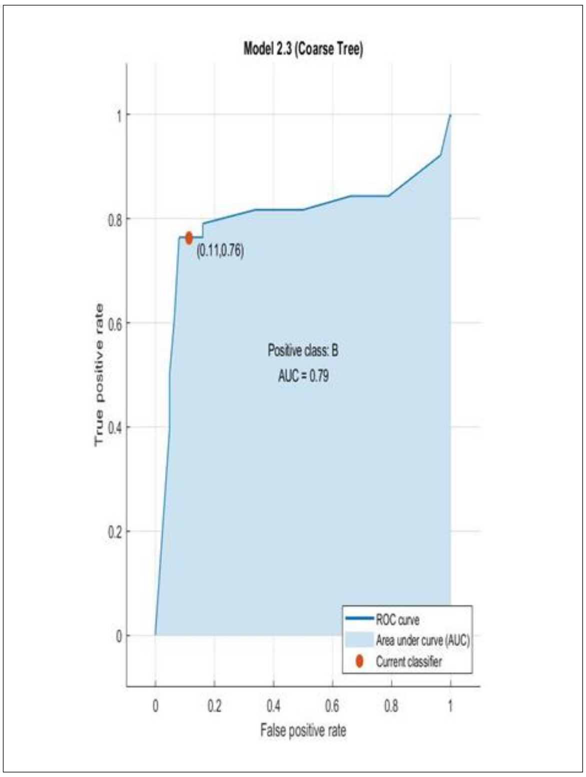


Figure 6 ROC curve for Tree

4. Conclusion

A considerable portion of the global population grapples with diabetes, a chronic ailment for which a cure remains elusive. Inadequate management of diabetes can lead to detrimental health outcomes, underscoring the criticality of timely detection. Notably, diabetes-related nerve impairment poses a significant threat to cardiac well-being. Hence, our current research endeavors to harness the power of advanced deep learning techniques to diagnose diabetes through the intricate analysis of heart rate variability (HRV) data. By leveraging the Random Forest algorithm, we have achieved an unparalleled level of accuracy, reaching an impressive pinnacle of 99%. This remarkable achievement stands as the highest reported threshold in the realm of automated diabetes identification, further affirming the significance of our findings.

Compliance with ethical standards

Statement of ethical approval

The present research work does not contain any studies performed on animals/humans subjects by any of the authors.

Statement of informed consent

Informed consent was obtained from all individual participants included in the study.

References

[1] M.A. Pfeifer, D. Cook, J. Brodsky, D. Tice, A. Reenan, S. Swedine, J.B. Halter, D. Porte, Quantitative evaluation of cardiac parasympathetic activity in normal and diabetic man, Diabetes 31 (4) (1982) 339–345.

[2] C. f. D. Control, G. C. f. D. C. Prevention %J Atlanta, U. D. o. H. Prevention, and H. Services, "National diabetes statistics report, 2020," 2020.

[3] I. D. Federation. IDF DIABETES ATLAS 9th edition 2019. Available: https://diabetesatlas.org/en/

[4] L. Olansky and L. Kennedy, "Finger-Stick Glucose Monitoring: Issues of accuracy and specificity," Diabetes Care, vol. 33, no. 4, pp. 948 – 949, 2010.

[5] J. B. Buse et al., "2019 update to: Management of hyperglycaemia in type 2 diabetes, 2018. A consensus report by the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD)," vol. 63, no. 2, pp. 221-228, 2020.

[6] M. Langendam, Y. M. Luijf, L. Hooft, J. H. DeVries, A. H. Mudde, and R. J. Scholten, "Continuous glucose monitoring systems for type 1 diabetes mellitus," Cochrane Database Syst Rev, vol. 1, 2012.

[7] A. A. Aljumah, M. K. Siddiqui, and M. G. Ahamad, "Application of classification based data mining technique in diabetes care," Journal of applied Sciences, vol. 13, no. 3, 2013.

[8] E. I. Georga, D. I. Fotiadis, and V. C. Protopappas, Glucose prediction in type 1 and type 2 diabetic patients using data driven techniques. INTECH Open Access Publisher, 2011.

[9] Piryonesi S. Madeh; El-Diraby Tamer E. (2020-06-01). "Role of Data Analytics in Infrastructure Asset Management: Overcoming Data Size and Quality Problems". Journal of Transportation Engineering, Part B: Pavements. 146 (2): 04020022. doi:10.1061/JPEODX.0000175.

[10] Piryonesi, S. Madeh; El-Diraby, Tamer E. (2021-02-01). "Using Machine Learning to Examine Impact of Type of Performance Indicator on Flexible Pavement Deterioration Modeling". Journal of Infrastructure Systems. 27 (2): 04021005.

[11] M. Durairaj, V. Ranjani, "Data Mining Applications In Healthcare Sector: A Study", International journal of scientific & technology research volume 2, issue 10, October 2013, ISSN 2277-8616.

[12] P.Yasodha, M. Kannan, "Analysis of a Population of Diabetic Patients Databases in WEKA Tool", International Journal of Scientific & Engineering Research Volume 2, Issue 5, May-2011, ISSN 2229-5518.

[13] Comparative Study on Classic Machine learning Algorithmshttps://towardsdatascience.com/comparative-study-on-classic-machine-learningalgorithms-24f9ff6ab222.

[14] Diagnosis and classification of diabetes mellitus Diabetes Care, 32 (Suppl. 1) (2009), pp. S62-S67.

[15] E.M. Cox, D. ElelmanTest for screening and diagnosis of type 2 diabetes Clin Diabetes, 4 (27) (2009), pp. 132-1384.

[16] Liao M, Liu Q, Li B, Liao W, Xie W, Zhang Y. A group of long non-coding RNAs identified by data mining can predict the prognosis of lung adenocarcinoma. Cancer Sci. 2018;109(12):4033.

[17] Li X, Zhao Z, Gao C, Rao L, Hao P, Jian D, Li W, Tang H, Li M. The diagnostic value of whole blood lncRNA ENST00000550337. 1 for prediabetes and type 2 diabetes mellitus. Exp Clin Endocrinol Diabetes. 2017;125(06):377–83.

[18] Mansoori Z, Ghaedi H, Sadatamini M, Vahabpour R, Rahimipour A, Shanaki M, Kazerouni F. Downregulation of long non-coding RNAs LINC00523 and LINC00994 in type 2 diabetes in an Iranian cohort. Mol Biol Rep. 2018; 45(5):1227–33.

[19] Leti F, DiStefano J. Long non-coding RNAs as diagnostic and therapeutic targets in type 2 diabetes and related complications. Genes. 2017; 8(8):207.

[20] Deshpande S, Thakare V. Data mining system and applications: a review. International Journal of Distributed and Parallel systems (IJDPS). 2010; 1(1):32–44.

[21] Umar Sidiq D, Aaqib SM, Khan RA. Diagnosis of various thyroid ailments using data mining classification techniques. Int J Sci Res Coput Sci Inf Technol. 2019; 5: 131–6.

[22] Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting diabetes mellitus with machine learning techniques. Front Genet. 2018; 9.

[23] A. A. Aljumah, M. K. Siddiqui, and M. G. Ahamad, "Application of classification based data mining technique in diabetes care," Journal of applied Sciences, vol. 13, no. 3, 2013.

[24] E. I. Georga, D. I. Fotiadis, and V. C. Protopappas, Glucose prediction in type 1 and type 2 diabetic patients using data driven techniques. INTECH Open Access Publisher, 2011.