Contemporary Mathematics
http://ojs.wiserpub.com/index.php/CM/

Research Article

Ensemble Technique for Diabetic Precision Medicine Classification

Badugu Sobhanbabu*, **K. F. Bharati**

Department of Computer Science and Engineering, Jawaharlal Nehru Technological University, Anantapur, 515002, Andhra Pradesh, India
E-mail: sobhanbabugec2015@gmail.com

Received: 24 September 2023; **Revised:** 15 January 2024; **Accepted:** 16 January 2024

Abstract: It is precisely 100 years after giving an insulin shot to a human, which created a revolution in diabetes treatment. Since the fate of diabetes patients has changed in humankind. In this regard, the exact dosage and same set of medicine for diabetes patients may not fit. Thus, there is a requirement for Precision medicine which can change the treatment of diabetes. There is a requirement for building automated intelligent systems to recommend precision medicine that can help practitioners. This paper discusses precision medicine, which has a complete schema for deriving Precision medicine from Big data. In our proposed schema, the component Intelligent Precision Medicine Engine, a new phase is added to filter the non-diabetic patient records to be processed by the Recommender engine, which would reduce the computational energies. With this object, a multi-layered bagging technique is used to classify with the best result nearing 96% with the UCI machine learning dataset for diabetes. Three layers of classification multi-models with majority voting are designed, and results are discussed.

Keywords: ensemble technique, precision medicine, diabetic classification big data analytics, intelligent systems

MSC: 68T05, 68T10, 62P10

1. Introduction

There are many benefits to big data in the health sector as well. A high-level medicine recommender system such as Precision medicine will require large health datasets, analytical ability, and technological support. This big data analytics helps predict more accurate treatment and prevention mechanisms for any disease, according to MedlinePlus Precision medicine is a new approach to disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person. A precision medicine approach is the opposite of a one-size-fits-all approach to medicine. Every individual differs due to their internal resistance and habits. In certain areas, such as blood fusion, the treatment depends on the patient's blood group. By providing personalized medicine, our technical approach also alters the medication schema. Precision Medicine and Big Data healthcare applications are presented in the research [1–5].

Diabetes is one of the most concerning problems, along with cancer or human immunodeficiency viruses (HIV) in the real world. It gets worse for the patient, as they are not even permitted for certain kinds of surgeries. The high Glucose in the blood leads to early deterioration of the Kidneys, which will impact the other organs quickly. The higher level of glucose directly causes Glaucoma, blindness, and alzhemers Alzheimers, such an important effect dedication that is

DOI: https://doi.org/10.37256/cm.6120253699

not uniform for all. There are several levels of Diabetes [6, 7], which are classified by standard organizations like the American Association of Diabetes [7].

1.1 *Precision medicine in diabetes*

In order to prevent long-term morbidity and mortality, clinicians must make the right diagnosis and use the right treatment. It is important to understand the impact of disease heterogeneity on potential treatment options for specific subtypes of diabetes in order to facilitate precision medicine in diabetes [8].

The high glucose in the blood leads to early deterioration of the Kidneys, which will impact the other organs quickly. The higher level of glucose directly causes Glaucoma, blindness and alzhemers too. Being such an important problem, the medication is not uniform for all. There are several levels of Diabetes [7, 9], which are classified by standard organizations like the American Association of Diabetes [9].

1.2 *Motivation*

To create an automated system to recommend precision medicine for diabetic patients new generation data analysis is required. As its been 100 years since insulin medicine is discoverred, this is a right time for this research on diabetic precision medicine. The centennial milestone since the discovery of insulin marks a pivotal moment to advance diabetic precision medicine. Leveraging contemporary data analysis and computational capabilities, this research aims to automate precise medicine recommendations for diabetic patients. The timely integration of new analytical methods and frameworks holds significant promise for practitioners. The paper's primary objective is to furnish an efficient classification system and timely drug recommendations for enhanced diabetic care.

In this paper organizes the Section 2: Literature Survey: Explores existing research in diabetic classification and recommendation, establishing a foundation for the proposed advancements. Section 3: Proposed Precision Medicine Schema: a novel schema designed for diabetic precision medicine, outlining the framework that supports effective classification and drug recommendation. Section 4: Experimentation and Results Analysis: Details the experimental setup and provides an analysis of the results obtained from implementing the proposed precision medicine schema. Section 5: Conclusion and Future Works: Summarizes key findings, conclusions drawn from the research, and outlines potential directions for future work in the domain of diabetic precision medicine.

2. Literature survey

Abhaya et al. [10] , and Lambay et al. [11] proposed the intelligent drug recommender using Big data analytics using distributed file systems, this paper also discuss about the data mangement in hethe althcare domain too. Nanehkaran et al. [12] presented a recommender system for chronic diseases using big data. Portha et al. [8] explains the significance of future road maps for diabetes durg recommendation. Their insightful directions helps to create new frame works for this diabetes medicine. Hulsen Tim et al. [13] highlighted the methods to derive precision medicine from the Big data.

Gou [9] presented a multidimensional analysis of big data in healthcare. Data privacy and security were highlighted in their research work. Healthcare data is sensitive for patients under treatment, so security is one of the most challenging aspects. As mentioned by Abid [2], and Haseem [14], there are a couple of Blockchain implementations for Precision medicine as well. Anil et al. [15] suggest evaluating them. The Big Data Healthcare management systems, their progress, and technological advancements were presented by Wang, Lidong. Additionally, Wang discussed Cloud and Stream processing in the Big Data healthcare domain.

Granda Morales et al. [16] has proposed an interesting Precision medicine system using clustering and collaborative filtering using dig data. It is highly useful when we deal with teh big data.

Saloni kumari [17] has also proposed diabetes classification using soft vote classifeir, but our approach differs from this approach. Hasan et al. [18] has proposed an empirical method to classify the diabetes, but our approach is not an empirical one.

Zheng et al. [18] the related work encompasses the historical evolution of deep learning, emphasizing the significance of image classification in computer vision. It explores the simulation of the human brain in neural networks and surveys advancements in activation functions. Relevant studies on the impact of activation functions on neural network performance are reviewed, along with analyses of the development status and performance of existing functions. The literature also includes proposed activation function designs and comparative studies on popular deep learning architectures. Additionally, it considers benchmark datasets like MNIST, CIFAR10/100, and ImageNet, while exploring state-of-the-art approaches in image classification for contextual understanding of the proposed improved activation function.

Zheng et al. (2017) in the realm of computer vision, fine-grained image classification poses challenges, particularly in discerning subclasses like different dog breeds. To address this, our approach integrates artificial features with deep convolutional activation features and employs support vector machines (SVM) based on feature importance. We utilize the bilinear neural network model for deep feature extraction, combining it with artificial features. The bilinear form facilitates gradient computation and end-to-end training. Subsequently, multi-kernel SVMs, based on weighted features, are trained for image classification. Experiments on FGVC-Aircraft and Stanford Dogs databases demonstrate the efficacy of our strategy, achieving 83.8% and 66.1% accuracy, respectively.

3. Methodology

A schema to arrive at Precision medicine from the Healthcare Big data is presented below. In our previous works, a Map-Reduced based dimensionality reduction in the first phasee and the Second phase a Recommender system for Personalized drug recommendation is executed. The proposed schema is presented in the Figure 1.

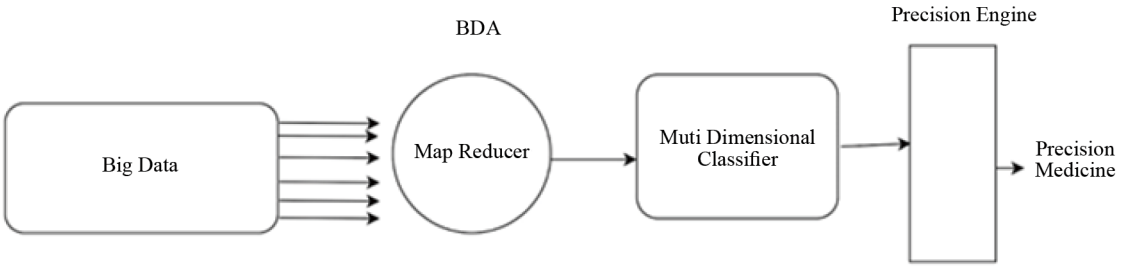

Figure 1. Proposed schema for diabetes precision medicine

The Precision Engine part is what being focused in this paper. A direct drug recommendation was proposed in the previous work. But it was observed that there are considerable amount of records which does not have diabetes. Executing a Recommender system before analyzing the diabetes was taking more energies in terms of computations. To overcome this limitation and Optimize the Interoperable engine, a new strategy (Figure 2) is proposed in this research work.

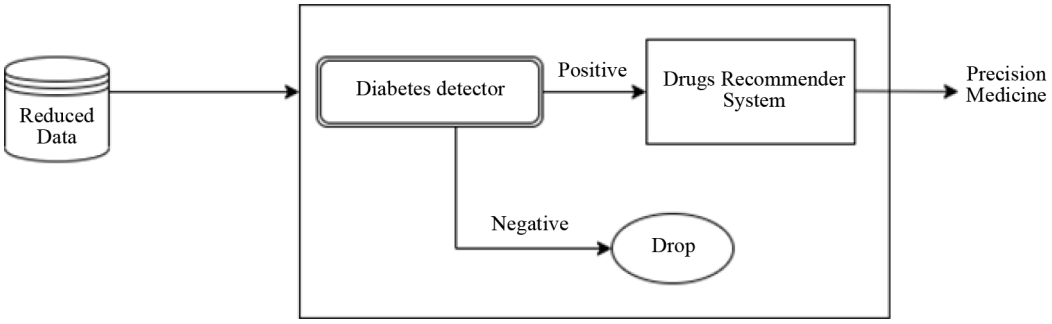

Figure 2. Refined intelligent precision engine

The objective of this attempt to optimize the Intelligent Precision medicine engine will reduce the complexity of the system.

There were attempts to classify diabetes using several machine learning techniques. But in this approach, the novelty of this research work is to propose an assembled based classification of diabetes. As the "Ensemble" learning techniques use multiple algorithms together to solve a classification problem. The accuracy is expected to be high compared to any other general traditional machine learning model in this work our initial assumption as per our previous works reflected in the Figure 1. From big data we have got a reduced data set to the intelligent precision engine. The data got reduced from 100,000 plus records to 18,566 records with 50 plus features. As we have an ample amount of records and an ample number of features, we can proceed for feature selection and row selection as well, which is a necessary condition for ensemble learning. In ensemble learning, we do proceed with splitting the data into multiple parts and multiple machine learning models. In this phase, we take the advantage of each type of machine learning classification algorithm in different layers. With which the true positive rate can be much higher than a single traditional model. As a standard ensemble pattern, we proceed with the majority voting of classification results from each model and then conclude the classification. The complexity of this system can be a little high compared to a single traditional system, but when we are more focused on the true positive and false negative cases, this approach can be more useful. In our experimentation, we have designed in such a way that three different layers with three data sets of 6,000 plus records in each data set is considered. The first layer data set one has 6,120 records in it and we have considered 14 features for classification. In the second layer data set, we have considered 6,220 records with the same 14 features. The layer third data set also have 6,220 records with the same 14 features that resulted from the Dimensionality reduction step.

Now each data set is independently implemented with a Decision Tree, K nearest neighbor and SVM. In layer one, the data set one has experimented with three classification algorithms and the majority of output of these three will be noted as resultant from layer one. In the layer two we have used Dataset 2 again and the same 3 algorithms and majority of the results of these three algorithms is noted as result. The third layer also has the same three algorithms, and the majority result of these three algorithm models is noted as the result in this third layer. Now in a holistic manner the majority of the results from layer one, layer two, layer three is noted as a final classification result. It may be a complex model in other fields but in the healthcare domain the. The precision of the system is very typical in taking the decisions its bearable. As we are proceeding with the drug recommender system for diabetes patients, this complex system. May be much more helpful in properly and accurately classifying a person whether they are having diabetes or not.

3.1 *Dataset description*

The dataset considered for this experimentation is obtained from UCI Machine Learning Repository [19]. There are 101,766 records with the diabetes data including following features. It includes 24 medicines that are given for diabetes and changes in the prescribed medicine too.

['encounter_id',
'patient_nbr',
'race',
'gender',
'age',
'weight',
'admission_type_id',
'dischange_disposition_id
'admission source id',
time_in_hospital',
'payer_code',
'medical specialty',
'num_lab_procedures',
'num procedures'
'num medications',
'number_outpatient',
'number_emergency' ,
'number_inpatient',
'diag_1',
'diag_2',
'diag_3',
'number diagnoses',
'max_glu_serum',
'AlCresult',
'metformin',
'repaglinide',
'metformin',
'repaglinide',
'nateglinide',
'chlorpropamide',
'glimepiride',
'acetohexamide',
'glipizide',
'glyburide',
'tolbutamide',
'pioglitazone',
'rosiglitazone',
'acarbose',
'miglitol',
'troglitazone',
'tolazamide',
'examide',
'citoglipton',
'insulin',
'glyburide-metformin',
glipizide-metformin',
'glimepiride-pioglitazone',
metformin-rosiglitazone',
'metformin-pioglitazone',
'change',
'diabetesMed',
'readmitted']

From the above 50 features, every feature may not be required for our analysis. There are some features like Patient id, which we remove from our initial computation.

From the first phase of our work we have got a reduced dimension of the given data. I.e 18,566 records are obtained after performing Map-Reduce on the given data. There are 18 attributes selected out of 50 attributes available in the dataset. They are derived from the Collaborative filtering used for dimensionality reduction. They are age, Gender, Weight, 'max_glu_serum', 'metformin', 'repaglinide', 'chlorpropamide', 'citoglipton', 'insulin', 'glyburide-metformin', 'glipizide-metformin', 'metformin-rosiglitazone', 'metformin-pioglitazone', "A1Cresult"-'diabetesMed' is the classification label.

- A uniform distribution of the dataset size to each layer is considered.
- The variation in the records of each set yield a better accuracy and decisive power.
- It's a bagging technique to have parallel execution of the multiple algorithms.
- The training set is given to each layer and 3 models are built at each layer.
- Over all, there will be 9 models built on the given training sets.

3.2 *Evaluation criteria*

- For any new entry, the same is given to each layer independently.
- At each layer, it is evaluated with the three models built in the training phase.
- The Final result of each later is computed with the majority voting of model 1, model 2, model 3.
- Over all result is obtained gain with the majority voting of Results of each layer.

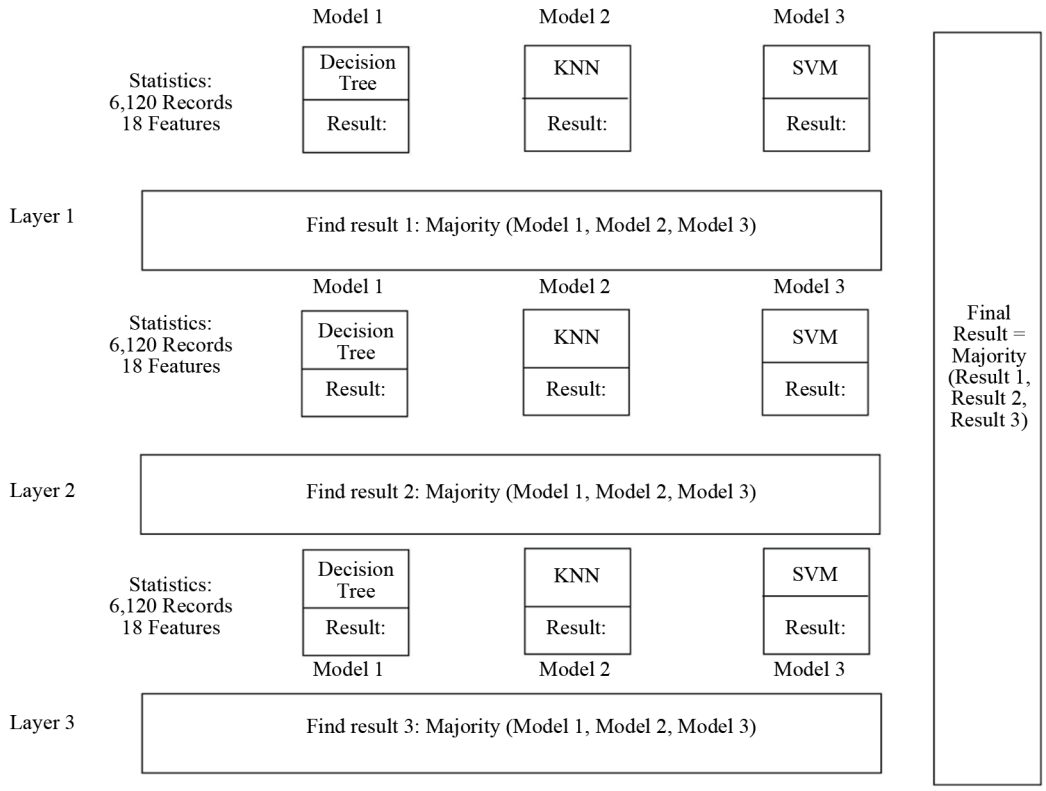

Figure 3. Proposed ensemble learning based intelligent precision

4. Experimentation and results

The experimental setup for this experimentation is Google Colab. The refined dataset from UCI Machine learning repository for Diabetes [19] is derived from our first phase of the system design mentioned in Figure 1.

- For the Decision tree, the entropy is the parameter used.
- For K Nearest Neighbor the k value used is 5.

4.1 *Result discussion*

The following table represents the Accuracy score, Precision, Recall and F1-Score of the diabetes classification.

Table 1. Results of each layer in diabetes classification in intelligent precision engine

		Accuracy	Precision	Recall	F1_Score
Layer1	Decision tree	0.83	0.85	0.87	0.86
	K-Nearest Neighbor (KNN)	0.73	0.73	0.72	0.71
	SVM	0.68	0.67	0.68	0.65
Layer 2	Decision tree	0.81	0.84	0.81	0.80
	KNN	0.70	0.72	0.71	0.71
	SVM	0.69	0.73	0.70	0.71
Layer 3	Decision tree	0.86	0.86	0.86	0.85
	KNN	0.75	0.74	0.74	0.7
	SVM	0.70	0.71	0.72	0.72

4.2 *Observations*

The analysis of the experimentation is as follows:

From the result Table 1, it is evident that in layer one, the decision tree has performed better than the other two algorithms. The decision tree has got an accuracy score of 0.83. Whereas K-Nearest Neighbor with an Accuracy score of 0.73 and the SVM with an accuracy score of 0.68 using the first part of the data set. In layer one, we could conclude that the decision tree has performed better than the remaining two algorithms. However, the result of each model is considered for evaluation. Now, as per the schema design, we will proceed with the majority vote on these three algorithms.

The second layer also has the same algorithms. The accuracy was noted to be 0.81 for the Decision Tree algorithm. The Precision is noted to be 0.84 that is significant for a healthcare-related problem. Just like in the first table, the K-Nearest Neighbor (KNN) stood be second, and SVM result was very near to the KNN. We can assume that the result of KNN and SVM are almost the same concerning layer 2. But Decision Tree has performed better than these two algorithms.

Layer three had better results with the decision tree algorithm. With this data set, we can observe that the three methods have performed better than the remaining two layers. The data in this layer might have more decisive power features. The Decision tree algorithms has got 0.86 accuracy score. The holistic result will be better than any individual models.

The following Table 2 represents the holistic results of all these layers together.

Table 2. Holistic results of all these layers together

Accuracy_Score	Precision	Recall	F1_Score
0.96	0.96	0.91	0.93

• We have achieved an accuracy of 96% using our proposed schema, which is far better than any remaining Classification research work performed on this data set yet.

• The precision we have got is 0.96 and the Recall 0.96 and F1 score is 0.93.

• The observed performance is obviously better than all three layers.

• The accuracy plays an important role in medical healthcare data-related research problems, especially when we want to provide personalized medicine or Precision medicine to a patient.

• The proper classification of precision and recall is highly required.

• This is our approach has justified the objective with a decent performance.

5. Conclusions and future work

To build a complete system for the diabetic precision medicine different phases of analytics need to be done. In this work, we have completed the classification part with a decent accuracy score. This will reduce the burden on the Hybrid Drug recommender system, which unnecessarily execute for the non diabetic patients. Any extension of this work can adopt other advanced algorithms including deep learning. Efforts for reducing the complexity of the schema can be an appreciatable contribution.

Conflict of interest

Authors declare that no conflict of interest.

References

[1] Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA*. 2018; 319(13): 1317-1318. Available from: https://doi.org/10.1001/jama.2017.18391.

[2] Prosperi M, Min JS, Bian J, Modave F. Big data hurdles in precision medicine and precision public health. *BMC Medical Informatics and Decision Making*. 2018; 18: 139. Available from: https://doi.org/10.1186/s12911-018-0719-2.

[3] Skylar JS, Bakris GL, Bonifacio E, Darsow T, Eckel RH, Groop L, et al. Differentiation of diabetes by pathophysiology, natural history, and prognosis. *Diabetes*. 2017; 66(2): 241-255. Available from: https://doi.org/10.2337/db16-0806.

[4] American Diabetes Association. Classification and diagnosis of diabetes mellitus: standards of medical care in diabetes. *Diabetes Care*. 2018; 41(Suppl 1): S13-S27. Available from: https://doi.org/10.2337/dc18-S002.

[5] De Silva D, Burstein F, Jelinek HF, Stranieri A. Addressing the complexities of big data analytics in healthcare. *Australasian Journal of Information Systems*. 2015; 19: S99-S115. Available from: https://doi.org/10.3127/ajis.v19i0.1183.

[6] Solis-Herrera C, Triplitt C, Reasner C, DeFronzo RA, Cersosimo E. Classification of diabetes mellitus. In: Feingold KR, Anawalt B, Boyce A. (eds.) *Endotext*. South Dartmouth (MA): MDText.com, Inc.; 2000.

[7] Agarwal A, Pritchard D, Gullett L, Garner Amanti K, Gustavsen G. A quantitative framework for measuring personalized medicine integration into US healthcare delivery organizations. *Journal of Personalized Medicine*. 2021; 11(3): 196. Available from: https://doi.org/10.3390/jpm11030196.

[8] Portha B, Bowman P, Flanagan SE, Hattersley AT. Future roadmaps for precision medicine applied to diabetes: Rising to the challenge of heterogeneity. *Journal of Diabetes Research*. 2018; 2018(1): 3061620. Available from: https://doi.org/10.1155/2018/3061620.

[9] American Diabetes Association. Diagnosis and classification of diabetes mellitus. *Diabetes Care*. 2010; 33(Suppl 1): S62-S69. Available from: https://doi.org/10.2337/dc10-S062.

[10] Sahoo A, Mallik S, Pradhan C, Mishra B, Barik R, Das H. Intelligence-based health recommendation system using big data analytics. *Health Information Science and Systems*. 2019; 19: 227-246. Available from: https://doi.org/10.1016/B978-0-12-818146-1.00009-X.

[11] Lambay M, Sheik P. Big data analytics for healthcare recommendation systems. In: *2020 International Conference on System, Computation, Automation and Networking (ICSCAN)*. Pondicherry, India: IEEE; 2020. p.3-4. Available from: https://doi.org/10.1109/ICSCAN49426.2020.9262304.

[12] Nanehkaran YA, Zhu L, Chen J, Qiu Z, Yuan X, Navaei YD, et al. Diagnosis of chronic diseases based on patients' health records in IoT healthcare using the recommender system. *Wireless communications and Mobile computing*. 2022; 2022(1): 5663001. Available from: https://doi.org/10.1155/2022/5663001.

[13] Hulsen T, Jamuar SS, Moody AR, Karnes JH, Varga O, Hedensted S, et al. From big data to precision medicine. *Frontiers in Medicine*. 2019; 6: 34. Available from: https://doi.org/10.3389/fmed.2019.00034.

[14] Hashim F, Harous S. Precision medicine blockchained: A review. In: *International Conference on computation, Automation and Knowledge Management (ICCAKM)*. 2nd ed. Dubai, United Arab Emirates: IEEE; 2021. p.48-52. Available from: https://doi.org/10.1109/ICCAKM50778.2021.9357760.

[15] Anil GR, Moiz SA. Blockchain enabled smart learning environment framework. In: Satapathy S, Raju K, Shyamala K, Krishna D, Favorskaya M. (eds.) *Advances in Decision Sciences, Image Processing, Security and computer Vision*. Basel Switzerland: Springer International Publishing; 2020. p.728-740. Available from: https://doi.org/10.1007/978-3-030-24318-0_83.

[16] Granda Morales LF, Valdiviezo-Diaz P, Reátegui R, Barba-Guaman L. Drug recommendation system for diabetes using a collaborative filtering and clustering approach: Development and performance evaluation. *Journal of Medical Internet Research*. 2022; 24(7): e37233. Available from: https://doi.org/10.2196/37233.

[17] Sivashankari R, Sudha M, Hasan MK, Saeed RA, Alsuhibany SA, Abdel-Khalek S. An empirical model to predict the diabetic positive using stacked ensemble approach. *Frontiers in Public Health*. 2021; 9: 792124. Available from: https://doi.org/10.3389/fpubh.2021.792124.

[18] Zheng Q, Yang M, Tian X, Wang X, Wang D. Rethinking the role of activation functions in deep convolutional neural networks for image classification. *Engineering Letters*. 2020; 28(1): 1-13.

[19] Dua D, Graff C. *UCI Machine Learning Repository*. Irvine (CA): University of California, School of Information and Computer Science; 2019.