Transactions on Engineering and Computer Science

Research| Vol 3 Iss 2

Type 2 Diabetes Prediction Using Machine Learning Confidence Weighted Based Ensemble Methods

Roland Babayev*

Colorado Technical University, USA

*Corresponding author: Babayev R, Colorado Technical University, USA; E-mail: rolandhb37[AT]gmail.com

Received: May 25, 2022; Accepted: July 01, 2022; Published: July 12, 2022

Abstract

In 2021, the International Diabetes Foundation estimated that 537 million people suffered from diabetes and that number is expected to rise to over 783 million by the year 2045. Type 2 diabetes (T2D) makes up 90% of diabetes cases. Machine learning research has been conducted to predict T2D risk to curb individual onset through intervention. Many of these machine learning studies involved the exclusive use of individual models such as Random Forest, Logistic Regression, and Neural Networks. The purpose of this study is to highlight contemporary research in the application of machine learning techniques for T2D prediction and to propose an alternative ensemble-based approach which uses confidence weighted voting of different base model classifiers. The ensemble methodology used in this study generated an average AUC metric of .846 over five different iterations during testing compared to individual models which produced AUC metric values ranging from .716 to .845. The ensemble performed better than a Bayesian Network, Neural Network, and Logistical Regression model. It performed marginally better than a Support Vector Machine. This highlights that "confidence weighted voting" based ensemble approach holds promise in producing improved predictive performance for type 2 diabetes prediction. Researchers and practitioners alike may achieve better predictive performance by applying such methods and should therefore seek to incorporate confidence weighted based ensembles into their standard machine learning procedures for type 2 diabetes prediction. Future research can seek to explore different statistical methods of using the confidence levels of models which compromise an ensemble in making final predictions besides the voting mechanism used in this study.

Keywords: ***Machine learning; Type 2 diabetes; Ensemble; T2D; Risk prediction; Stacking; Data mining; blending.***

1. Introduction

Type 2 diabetes (T2D) is the most prevalent form of diabetes in the world [1]. Over 90% of those who have diabetes around the globe have type 2 diabetes. This type of diabetes is characterized by either insulin resistance or insulin

Citation: Babayev R. Type 2 diabetes prediction using machine learning confidence weighted based ensemble methods. Trans Eng Comput Sci. 2022;3(2):127.

deficiency, or both, and can occur at any age [1]. Insulin resistance and/or insulin deficiency compromises the body's ability to manage blood glucose levels and therefore can lead to serious health complications over time, including cardiovascular disease, kidney failure, or diabetic nephropathy, blindness, diabetic neuropathy, diabetic retinopathy, and more [1]. Individuals with diabetes are 2 to 3 times more likely to develop cardiovascular disease and one third of those with diabetes suffer from some form of diabetic retinopathy, or eye disease. Pregnant women with diabetes who do not manage their blood glucose levels optimally risk complications during delivery due to excess weight that can be developed by the child. Not managing blood glucose levels during pregnancy can lead to organ failure in the fetus as well as an increased risk of the child developing diabetes, hypertension, kidney disease and/or obesity [1]. The International Diabetes Foundation estimated that in 2021, 537 million people lived with diabetes worldwide. By 2045 the number of people with diabetes is expected to rise to 783 million. Diabetes resulted in 6.7 million deaths in 2021 and 1 in 2 people with diabetes were undiagnosed [1].

To that end, predicting diabetes risk is of extreme significance as early intervention can lead to a reduction in the overall risk of type 2 diabetes onset [2]. In a clinical trial known as the Diabetes Prevention Program conducted by the National Institute of Health, modest lifestyle changes which led to weight reductions of 5 to 7 percent decreased the onset of type 2 diabetes by as much as 58% [2]. Obesity, defined as having a body mass index (BMI) of 30 or above, is a risk factor for T2D onset [2]. Multiple studies have been conducted in the area of type 1 and 2 diabetes prediction using machine learning techniques [3]-[5]. The purpose of this study shall be to highlight contemporary research conducted in the area of the application of machine learning techniques for T2D risk prediction and to propose another methodology which leverages the notion of confidence weighted voting based ensembles, where ensembles involve the use of predictions from multiple, base, individual machine learning models as inputs into a final, meta-classifier type model [6]. The proposed methodology of this study shall be demonstrated by using the Pima Indians Diabetes Dataset [7]. In addition to using a confidence weighted voting-based ensemble for T2D prediction, the proposed algorithm will also assess, by means of Area Under the Curve (AUC), which individual base machine learning models to use as inputs into the ensemble technique.

The problem to be addressed in the study is the lack of the use of confidence weighted based ensemble machine learning methods in the domain of type 2 diabetes prediction. The current study will seek to demonstrate that confidence weighted based machine learning models can achieve better predictive performance than individual models for type 2 diabetes prediction. Following the introduction section, the paper is structured according to these sections: The related works, problem statement, hypothesis statement, purpose statement, research questions, materials and methods, data preparation, base/ensemble model building, results, and conclusion.

2. Related Works

Sisodia and Sisodia [7] used various individual machine learning model classification techniques to predict the risk of type 2 diabetes against the Pima Indians Diabetes Dataset (PIDD) which is also being used as a part of this study. The models used included Decision Tree, SVM and Naive Bayes [7]. The best performing model in their study was Naïve Bayes, with an accuracy of 76.30%. Sisodia and Sisodia [7] also measured model performance using F-measure, precision, and recall.

Talaei-Khoei and Wilson [8] studied T2D risk prediction in the context of short, medium, and long-term onset development. The individual models developed included Neural Networks, Support Vector Machines, Decision Tress and Logistic Regression [8]. Model performance was compared to assess which models were most effective depending on whether the prediction objective was short term (i.e. 1 year), medium term (i.e. 3 years), or long term (i.e. 8 years) disease development [8]. The research uncovered that for short- and medium-term onset predictions, Logistic Regression performed best, while for long term predictions, Support Vector Machines were the most effective [8].

Wang et al. [9] sought to predict T2D using individual models including Artificial Neural Networks (ANN) and compared the results to a Multivariate Logistic Regression model (MLR). The data of 6480 rural adult subjects was used for their study. The measures of performance used in their study included Area Under the Curve (AUC) which is a mechanism used to compare predictive performance of different models [9]. The ANN model generated an AUC of .891 while the MLR generated an AUC of .744, indicating that the ANN had superior predictive performance against the same dataset [9].

Various machine learning models have been developed to predict diabetes from data generated by non-invasive diagnostic techniques. Nirala et al. [5] used the signals generated by photoplethysmogram (PPG) tests administered to subjects in order to predict T2D. The test involved attaching sensors to the big toes of patients in order to measure the volume of blood in vessels within the area [5]. Stiffness in blood vessel walls is a symptom of diabetes which can be used as a predictive attribute [5]. The data points derived from these tests included the slope of the PPG signal, the area on the curve of the signal, and more [5]. By using a hybrid feature selection algorithm and a Support Vector Machine model, the researchers achieved an accuracy of over 97%. The number of patients in the study included 83 patients with T2D and 58 who did not [5].

Samant and Agarwal [10] used scans of the iris in patients with and without diabetes in order to demonstrate that machine learning methods can be applied to such scans for T2D prediction. Iridology is a traditional, alternative Chinese medicine approach which correlates segments of the iris with different organs of the human body [10]. The section of the iris studied by Samant and Agarwal [10] is correlated with the pancreas. In iridology, the color, tissue weakness, patterns, and other features of the iris are indicative of compromised pancreatic health [10]. The study used scans of the eyes of 338 patients and was able to achieve an accuracy score of 89.63% using a Random Forest model.

López et al. [11] used machine learning models to predict T2D from genetic data, namely, single nucleotide polymorphisms (SNP's). The most predictive model was a Random Forest algorithm which achieved an AUC of .890 [11]. The clinical health data of 677 patients was used in the study, along with 96 SNP's for each patient, to create a hybrid dataset of traditional clinical health information and genetic data [11].

A study by Zhu et al. [12] leveraged Principal Components Analysis (PCA) to reduce the feature space in a Pima Indians Diabetes dataset in addition to a K-Means clustering algorithm. The approach is therefore in contrast to the prior studies in that a feature reduction method was used in conjunction with an unsupervised machine learning clustering methodology for diabetes grouping. Finally, a Logistic Regression model was applied to the grouped data for ultimate prediction. Kaur and Kumari [13], also analyzed the Pima Indians Diabetes dataset by applying various models such as Support Vector Machines, Neural Networks, K-Nearest Neighbor, and more in an effort to compare model results. Unlike Zhu et al. [12] who used PCA, Kaur and Kumari [13] used embedded feature selection methods to reduce the feature space in an effort to improve model performance. Similar to Zhu et al. [12], Wu et al. [14] sought to combine unsupervised methods such as K-Means clustering and supervised methods, namely Logistical Regression, for diabetes prediction. Wu et al. [14] optimized the model for diabetes prediction by using multiple diabetes datasets to improve generalizability.

In summary, prior studies have leveraged various individual machine learning models as well as ensemble approaches for the prediction of type 2 diabetes. Additionally, feature engineering and dimension reduction methods have been utilized to improve predictive performance. Such methods, however, have not taken into consideration the confidence of model predictions or methods of using confidence values associated with different model predictions in making final predictive decisions. The aim of the current study will be to address this current gap in the literature.

3. Problem Statement

The machine learning methods used for T2D prediction may not achieve optimal performance levels partially because they do not employ ensemble techniques combined with an assessment of the best base models to use in the ensemble algorithm.

4. Hypothesis Statement

By using an ensemble machine learning technique which leverages a confidence weighted voting method against the individual base model predictions and prediction confidence scores, the overall predictive performance of T2D systems can be improved compared to the use of the individual base models alone.

5. Purpose Statement

The purpose of this study was not to compare the performance metrics achieved in this effort to those achieved in prior studies. The purpose was rather to demonstrate, through a methodological process and the use of a single T2D dataset, that an ensemble approach which uses confidence weighted voting can be more effective than the application of isolated, individual base models alone. Thus, the contention is that the same approach can be used by other researchers to improve their own results even further.

6. Subject Matter Areas

The purpose of this study is to apply machine learning techniques to the prediction of T2D. The focus of this research shall be on studying the applicability of a confidence weighted voting-based ensemble technique for T2D prediction on the Pima Indians Diabetes Dataset (PIDD). An ensemble combines multiple classification models for greater predictive efficacy rather than relying on a single machine learning model [6]. In essence, a suite of individual models are trained against the data and their separate predictions are subsequently used together as input features into a meta-classifier for improved predictive performance [6]. Murphree et al. [6] sought to predict which patients would be most likely to fail primary or secondary treatment on the diabetes drug metformin within one year of treatment. The ensemble approach executed by Murphree et al. [6] involved training a multitude of base models, including Random Forest, Linear Logistic Regression, Tree-based models, Stochastic Gradient Boosting [gbm], Neural Networks and more to predict metformin failure. The researchers used a total of 20 base models whose predictions were then fed as features into meta-classifier models, including Gradient Boosting Models and Logistic Regression [6]. The study included data from 12,147 patients and resulted in AUC values between .58 and .75 [6].

The ensemble methodology is used across many different domains outside the realm of disease prediction as well [15]. Li et al. [15] used stacking ensembles in order to combine Gradient Boosting Decision Tree (GBDT), XGBoost, and LightGBM models to help identify phishing web pages. The system created by the researchers leveraged a two-layer approach, in which the three models were first trained individually [15]. After being trained individually, the predictions of the models and the original predictors of the training dataset were used as features to train a second layer composed of the same three individual models [15]. Finally, a GBDT model was used as a meta-classifier to predict phishing web pages.

7. Research Questions

The proposed technique in this study is to predict T2D against a Pima Indians Diabetes Dataset (PIDD) using individual base model classifier predictions and prediction confidence scores. The base model classifier predictions and predictions confidence scores will be analyzed using confidence weighted voting applied by an ensemble classifier to vote on the final prediction of a record. The research question posed as part of this study is: Will predictive performance improve for T2D by using an ensemble approach that applies confidence weighted voting relative to individual base model classifiers?

8. Materials and Methods

In order to prove the hypothesis and answer the research questions posed in the earlier sections of this paper, the following research design was utilized. IBM SPSS Modeler was leveraged as the modelling tool because it provides ease and transparency into the flow of data and associated analytics applied throughout a predictive modelling processing stream. The first phase of study involved the building of base models which would be evaluated across metrics including accuracy and Area Under the Curve (AUC). These same models would then be used as components of the ensemble classifier. The performance measure in terms of AUC of each base model was then compared to the performance of the overall ensemble.

8.1 Data preparation

The Pima Indians Diabetes Dataset (PIDD) retrieved from the UCI learning repository was used as a source of data and consisted of the features listed in Table 1. The PIDD dataset was analyzed using the SPSS data audit node to ensure that there are no missing data values for each of the attributes. The number of records in the dataset is 768. The data analysis results indicated that no data was missing, as can be seen in Figure-1. Furthermore, the distribution of data is depicted below in Figure-2, where 500 records are associated with T2D and 268 are non-diabetic. The data attributes were then all standardized using z-score standardization to ensure that none of the features have an undue influence in the model building process as a result of their scaling [16].

Complete fields (%): 100% Complete records (%): 100%

Field	Measurement	% Complete	Valid Records	Null Value
Pregnancies	Continuous	100	768	0
Glucose	Continuous	100	768	0
BloodPressu...	Continuous	100	768	0
SkinThickness	Continuous	100	768	0
Insulin	Continuous	100	768	0
BMI	Continuous	100	768	0
DiabetesPed...	Continuous	100	768	0
Age	Continuous	100	768	0
Outcome	Nominal	100	768	0

Fig. 1. Attribute datatypes and completeness.

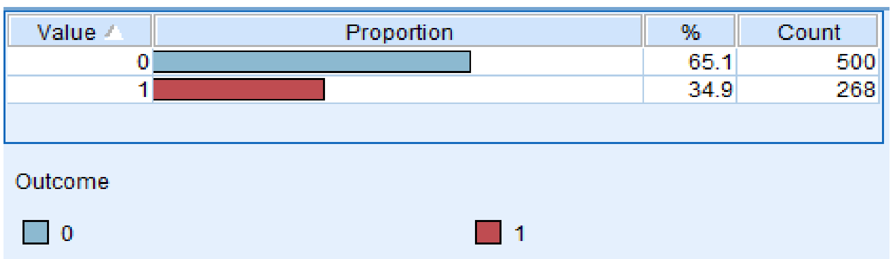

Fig. 2. Dataset distribution.

The formula below was applied to each attribute to generate new, standardized versions of the attributes. In the below formula, the variable A_z represents the z-score standardized value. The variable A represents the original attribute value while A_m represents the mean across all the values of the attribute. The difference between these variables is divided by A_{SD} which represents the standard deviation of the attribute values. The formula used is provided in Eq. (1).

$$A_z = \frac{A - A_m}{A_{SD}} \tag{1}$$

Table 1: Pima Dataset Data Attributes.

Attribute
Number of times pregnant
Plasma glucose concentration
Diastolic blood pressure (mm Hg)
Skin fold thickness (mm)
2-hour serum insulin (mu U/ml)
BMI (weight in kg/(height in m)^2)
Diabetes pedigree function
Age in years
Class '0' or '1'

8.2 Base model building

During the base model building exercise, the z-score representations of the attributes were used as inputs into multiple base models for training purposes. The PIDD dataset was split randomly using a 70/30 partition ratio, in which 70 percent of the data was used for training and the remainder was used as a holdout dataset for testing. The process involved randomly drawing, with replacement, 5 separate training and test datasets, each having a 70/30 split. The base model classifiers were therefore trained and tested over 5 iterations using the generated datasets. A "confidence weighted voting" ensemble of these based models was therefore tested and compared to the base classifiers over 5 iterations as well.

8.3 Ensemble model

The base models used in the study were Logistic Regression, Support Vector Machine (SVM), Artificial Neural Network (ANN), and Bayesian Network. The purpose of the "confidence weighted voting" ensemble model was to use trained and tested base model predictions as attributes for a combined, meta-classifier. The ensemble model was configured to leverage confidence-weighted voting to make final classifications based on the predictions and the prediction confidence scores of the base models against the test datasets. Assuming the problem domain is binary class based (i.e., 'Y' for presence of T2D and 'N' for absence of T2D), then the below depicts an example of confidence weighted voting.

1. Let *LP, NP, BP, and SP* represent the respective Logistic Regression, Neural Network, Bayesian Network, and Support Vector Machine predictions for record *i.*
2. Let *LC, NC, BC*, and *SC* represent the confidence values between 0 and 1 calculated by IBM SPSS Modeler of a Logistic Regression, Neural Network, Bayesian Network, and Support Vector Machine predictions, respectively, for record *i*.
3. If *LP* and *BP* predict 'Y' and *NP* and *SP* predict 'N':
   - Then *LC*(Y) + *BC*(Y) = The confidence vote for class 'Y' of record *i, or CV(Y)*
   - Then *NC*(N) + *SC*(N) = The confidence vote for class 'N' of record *i, or CV(N)*
4. If *CV*(Y) > *CV*(N), then final prediction for record *i* = *CV*(Y)
   - elseif *CV*(N) > *CY*(Y), then *CV*(N)
   - else, select prediction of max(*LC, BC, NC, SC*)

The final ensemble results were compared to the AUC values generated by the base models themselves against the same test dataset. This process of training base model classifiers, testing them, running an ensemble model using confidence weighted voting, and comparing the AUC based metric results was repeated over 5 iterations leveraging bootstrap drawn datasets, as described in the base model building section. Table 2 presents the AUC values resulting from the 5 iterations of base and ensemble model runs.

Table 2: AUC Results of Ensemble vs. Base Models.

Model	AUC1	AUC2	AUC3	AUC4	AUC5	**Average AUC**
Logistic Regression	.840	.861	.799	.838	.842	.836
Support Vector Machine	.850	.864	.813	.844	.855	.845
Artificial Neural Network	.811	.857	.835	.827	.850	.836
Bayesian Network	.634	.726	.753	.709	.756	.716
Ensemble	**.842**	**.861**	**.826**	**.849**	**.850**	**.846**

9. Results

9.1 Base model results

The results of the base model comparisons can be found in Fig. 3. The models in Fig. 3 are sorted based on performance using the AUC value. The best performing model was Support Vector Machine with an accuracy of 75.472% and an AUC of .850. This model is followed in terms of performance by the Logistic Regression which produces an accuracy of 77.358 % and an AUC value of .840. The Neural Network produces an accuracy of 75.000% and an AUC of .811. The fourth model in terms of AUC was the Bayesian Network, with a value of .634.

Model	Build Time (mins)	Max Profit	Max Profit Occurs in (%)	Lift{Top 30%}	Overall Accuracy (%)	No. Fields Used	Area Under Curve
SVM 1	< 1	160.0	18	1.936	75.472	8	0.850
Logistic regression 1	< 1	150.0	23	1.893	77.358	8	0.840
Neural Net 1	< 1	125.0	27	1.893	75.0	8	0.811
Bayesian Network 1	< 1	77.000	15	1.296	48.113	8	0.634

Fig. 1. Base model results against test dataset.

9.2 Ensemble vs. Base Model Results

The metric comparison used to compare base model performance and ensemble performance was AUC and can be found in Table 2. As can be seen in Table 2, the ensemble produces an average AUC value over 5 iterations which is higher than any of the individual base models alone, although the improvement is negligible compared to certain base models, namely the Support Vector Machine. The ensemble method resulted in slightly higher improvements compared to the Logistic Regression, Bayesian Network, and the Artificial Neural Network Models. This demonstrates that ensemble approaches leveraging confidence weighted voting algorithms on the predictions and confidence scores of base model classifiers can be used to improve predictive performance of T2D compared to individual models alone.

Prior research has either leveraged individual models or conventional ensemble approaches to predict type 2 diabetes. Prior research has also leveraged feature engineering and dimensionality reduction methods to help improve predictive performance. The results of the current study demonstrate that confidence weighted based ensemble approaches can achieve better predictive performance than certain individual models alone. Confidence weighted based ensemble approaches should therefore be included in the process of assessing which predictive models best predict type 2 diabetes in practice and research.

10. Conclusion

The contention posed during the course of this article was by leveraging the ensemble machine learning technique of confidence weighted voting on base model predictions, the predictive performance of T2D systems can be improved compared to the use of individual models. The results of this study demonstrate that this is the case, although the ensemble outperformed individual base models to different extents. Base models were constructed, evaluated based on AUC, and then implemented as inputs for use in an ensemble model. The ensemble proved to have better performance in terms of average AUC over 5 iterations compared to the base models themselves. The purpose of this study was not to compare the performance metrics achieved in this effort to those achieved in prior studies. The purpose was to demonstrate that a “confidence weighted voting” ensemble approach can be more effective than the application of isolated, individual base models. Thus, based on the results of this study, the contention is also that the same approach can be used by other researchers to improve their results even further. Beyond improving the predictive performance using an “confidence weighted voting” ensemble approach, performance could possibly have been further improved by using the concept of layering which repeats the method discussed in this study. The predictive results of the first “confidence weighted voting” ensemble layer could be used as predictors in yet another, secondary layer ensemble to improve predictive performance even further. Future research can explore in more detail the potential performance improvement that can be achieved by using multiple layers. Studies which have applied individual model methods in type 1 and 2 diabetes research may seek to explore whether their results would be improved using “confidence weighted voting” ensemble methods as well.

In conclusion, the implications of the study include changes to how future machine learning model processes can be structured for type m2 diabetes prediction. Machine learning efforts, both in practice and research, should consider using confidence weighted approaches when applying ensemble methods to make predictions. The weighted approaches, such as the one applied in this study, should take the confidence values of the individual model predictions which compose the ensemble into consideration when determining the final model prediction. A 'weighted' approach based on confidence values is in contrast, for example, to ensemble approaches where all models within a machine learning ensemble are given equal weight in making the final prediction. Researchers and practitioners alike may achieve better predictive performance by applying such methods and should therefore seek to incorporate confidence weighted based ensembles into their standard machine learning protocols and procedures for type 2 diabetes prediction.

Future research can also seek to explore different ways of using the confidence values of models which compromise an ensemble in making final predictions. The current study was limited to a single method of using model confidence levels, namely comparing the totals of the confidence level values for each prediction type. Opportunities exist to leverage different algorithms for confidence level aggregation and comparison, including using statistical methods to aggregate and compare confidence values. The current study was also limited to a handful of machine learning models which were used to develop the confidence weighted based ensemble. The study did not compare the performance of different ensembles composed of different underlying model types. Such a study would shed more light on how changing the underlying models of a confidence weighted based ensemble can change the ultimate predictive performance of the ensemble.

REFERENCES

1. International Diabetes Foundation. What is diabetes? 2021. [Online]. Available: https://www.idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html
2. American Diabetes Association. Preventing diabetes with good nutrition. 2021. [Online]. Available: http://www.diabetes.org/advocacy/advocacy-priorities/prevention/preventing-diabetes-nutrition.html
3. Kandhasamy JP and Balamurali S. Performance analysis of classifier models to predict diabetes mellitus. Procedia Comput Sci. 2015;47:45-51. [Online]. Available: https://doi.org/10.1016/j.procs.2015.03.182
4. Kavakiotis I, Tsave O, Salifoglou A, et al. Machine Learning and Data Mining Methods in Diabetes Research. Computat Struct Biotechnol J. 2017;15:104-116. [Online]. Available: https://doi.org/10.1016/j.csbj.2016.12.005
5. Nirala N, Periyasamy R, Singh BK, et al. Detection of type-2 diabetes using characteristics of toe photoplethysmogram by applying support vector machine. Biocybern Biomed Eng. 2019;39(1):38-51. [Online]. Available: https://doi.org/10.1016/j.bbe.2018.09.007
6. Murphree DH, Arabmakki E, Ngufor C, et al. Stacked classifiers for individualized prediction of glycemic control following initiation of metformin therapy in type 2 diabetes. Comput Biol Med. 2018;103:109-115. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2018.10.017
7. Sisodia D and Sisodia DS. Prediction of diabetes using classification algorithms. Proc Comput Sci. 2018;132(2018):1578-1585.
8. Talaei-Khoei A and Wilson JM. Identifying people at risk of developing type 2 diabetes: A comparison of predictive analytics techniques and predictor variables. Int J Med Informat. 2018;119:22-38. [Online]. Available: https://doi.org/10.1016/j.ijmedinf.2018.08.008
9. Wang C, Li L, Wang L, et al. Evaluating the risk of type 2 diabetes mellitus using artificial neural network: An effective classification approach. Diabetes Res Clin Pract. 2013;100(1):111-118. [Online]. Available: https://doi.org/10.1016/j.diabres.2013.01.023
10. Samant P and Agarwal R. Machine learning techniques for medical diagnosis of diabetes using iris images. Comput Methods Programs Biomed. 2018;157:121-128. [Online]. Available: https://doi.org/10.1016/j.cmpb.2018.01.004
11. López B, Torrent-Fontbona F, Viñas R., et al. Single nucleotide polymorphism relevance learning with random forests for type 2 diabetes risk prediction. Artif Intell Med. 2018;85:43-49. [Online]. Available: https://doi.org/10.1016/j.artmed.2017.09.005
12. Zhu C, Idemudia CU, and Feng W. Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. Inform Med Unlocked. 2019;2019:100179. [Online]. Available: https://doi.org/10.1016/j.imu.2019.100179
13. Kaur H, and Kumari V. Predictive modelling and analytics for diabetes using a machine learning approach. Appl Comput Informat. 2018. [Online]. Available: https://doi.org/10.1016/j.aci.2018.12.004
14. Wu H, Yang S, Huang Z, et al. Type 2 diabetes mellitus prediction model based on data mining. Inform Med Unlocked. 2018;10:100-107. [Online]. Available: https://doi.org/10.1016/j.imu.2017.12.006

15. Li Y, Yang Z, Chen X, et al. A stacking model using URL and HTML features for phishing webpage detection. Future Gener Comput Syst. 2019;94:27-39. [Online]. Available: https://doi.org/10.1016/j.future.2018.11.004
16. Larose DT and Larose CD. Data mining and predictive analytics, 2nd ed., Wiley; 2015, 824p.

Roland Babayev received the Master of Science and Doctorate degrees in Computer Science from Colorado Technical University. He has also worked in the technology sector for over two decades. His research focus has been the application of machine learning methods for disease prediction. Additionally, his research interests include the application of neural networks, including graph neural networks and other graph-based machine learning methods, for use cases such as fraud detection and social network analysis. His work has been in the financial technology domain where he has focused on the design and development of surveillance monitoring systems which apply algorithms to detect illicit capital market manipulation attempts through various trading strategies.

Citation: Babayev R. Type 2 diabetes prediction using machine learning confidence weighted based ensemble methods. Trans Eng Comput Sci. 2022;3(2):127.