Comparison of Statistical Logistic Regression and RandomForest Machine Learning Techniques in Predicting Diabetes

Tahani Daghistani and Riyad Alshammari
Health Informatics Department, College of Public Health and Health Informatics, King Saud Bin Abdulaziz University for Health Sciences (KSAU-HS), King Abdullah International Medical Research Center (KAIMRC), Ministry of National Guard Health Affairs, Riyadh, KSA
Email: TahaniDaghistani@gmail.com, alshammarir@ksau-hs.edu.sa

***Abstract*—Diabetes is one of the global concerns in the healthcare domain and one of the leading challenges locally in Saudi Arabia. The prevalence of diabetes is anticipated to rise; early prediction of individuals at high risk of diabetes is a significant challenge. This study aims to compare RandomForest machine learning algorithm and Logistic Regression algorithm towards the prediction of diabetes. We analyzed 66,325 records that extracted from the Ministry of National Guard Hospital Affairs (MNGHA) databases in Saudi Arabia between 2013 and 2015. Both Machine Learning algorithms were applied to predict diabetes based on 18 risk factors. The evaluation criteria to compare the two algorithms were based on precision, Recall, True Positive rate, False Negative rate, F-measure and Area under the curve. The overall prevalence of diabetes in the data set is 64.47%. Male represents 55.50% of the data set while female represents 44.50%. For RandomForest (RF) model, the precision, Recall, True Positive Rate, False Positive Rate and F-measure value for predicting diabetes were 0.883, 0.88, 0.88, 0.188 and 0.876, respectively, while Logistic Regression model were only 0.692, 0.703, 0.703,0.454 and 0.675, respectively. Area under the ROC curve (AUC) value was 0.944 for the RF model and 0.708 for Logistic Regression model, which demonstrates higher predictive performance for RF than the Logistic Regression model. The RF algorithm showed superior prediction performance over Logistic Regression technique in predicting diabetes based on various matrices.**

***Index Terms*—diabetes, predictive model, machine learning, RandomForest, logistic regression**

I. Introduction

Diabetes is a severe disease that is increasing widely around the world. The seriousness of the disease lies in the complications occur as a result of either patient neglecting to check for diabetes or not receiving appropriate care promptly. The most common complications of Diabetes are heart disease, stroke, kidney disease and causes of death [1]. The global prevalence of diabetes for adult aged more than18 years old was 8.5% in 2014 accordance with World Health Organization (WHO) [2]. In parallel with increasing prevalence of diabetes, there is an increase in associated consequences of complications of diabetes. Hence, the death cases of diabetes complications are rising proportionally [3]. In 2015, there was an estimate of 1.6 million deaths as a direct caused by diabetes. In 2030, WHO anticipates that diabetes will be the seventh leading cause of death [2]. In Saudi Arabia, there is an excessive prevalence of diabetes that is expected to be more than 2.5 million patients by 2030 [4]. Diabetes type 2 was defined as a previous clinical diagnosis or an electronic medical record (EMR) [5].

Early prediction of diabetes type 2 is one of the prominent health research topics in Saudi Arabia. Diabetes Risk Score was the most convenient tool for prediction [6]. However, this method needs human intervention in decision-making. Nowadays, Computational models to predict the risk of diabetes can significantly support decision-making and assist self-disease management [7]. Therefore, machine learning is gaining attention in the health field as these techniques produce high performance in predicting diabetes. These models can be helpful in identifying those who are at high risk of having diabetes, for which prevention and control programs can be initiated to improve health outcomes [6], [8]. At the same time, these techniques reduce the human error in making the decision. Thus, decreasing health burden and utilizing health service resources [3]. Ideally, further development of models that incorporate prior knowledge would be auspicious for diabetes prediction [9]. The availability of a patient's health data could help to extract meaningful information and hidden knowledge.

The study aims to comparatively evaluate the performance of the machine learning based models in predicting diabetes mellitus. The two prediction approaches were applied on data sets collected from the Ministry of National Guard Health Affairs (MNGHA) hospital's databases from three regions of Saudi Arabia, mainly Central region (Riyadh city), Western region (Jeddah city) and Eastern region (Dammam and Al-Ahsa cities). Based on our best knowledge, this study is one of the most significant studies to date for early detection of diabetes concerning cohort size as well as the number of attributes considered.

Manuscript received September 25, 2019; revised April 6, 2020.

doi: 10.12720/jait.11.2.78-83

The remaining segments of this research article are arranged as follows: Section II presents the literature review on the challenges of using machine learning based models in predicting diabetes. Section III explains the methodology while Section IV presents the results. Discussion is discussed in Section V. Finally, Section VI includes the conclusion and future work.

II. Literature Review

Several studies have focused on the comparison between Logistic Regression models and various machine learning based models. Such prediction studies have become the central research area in health. Al-Mallah *et al.* [10] showed the superior of machine learning in prediction using different evaluation metrics. They aimed to predict All-Cause Mortality (ACM) for patients undergoing stress testing. They gathered ten years of follow up data for 34,212 patients. They achieved a sensitivity of 44.9% and specificity of 93.4% for Logistic regression and sensitivity of 87.4% and specificity of 97.2% for machine learning. Regarding area under the curve, Logistic Regression achieved 0.836 and machine learning achieved 0.923.

Dalakleidi *et al.* [11] implemented Evolving Artificial Neural Networks (EANNs), Bayesian-based algorithm, decision trees and logistic regression for predicting the development of diabetes and predicting one of the diabetes complications that is cardiovascular disease. The highest accuracy achieved by the EANNs model with an accuracy of 80.20% and Area under the Curve (AUC) of 0.849. The model could predict the complication with an accuracy of 92.86% and an AUC of 0.739 as well.

Several studies related to the diabetes research conducted for comparison purpose. Meng *et al.* [12] presented an experimental comparison of three different techniques for predicting diabetes or prediabetes on 735 patients using conventional risk factors. The four techniques used are logistic regression, decision tree and Artificial Neural Networks (ANNs). The decision tree model outperformed the other techniques, achieving an accuracy of 77.87% with a sensitivity of 80.68% and specificity of 75.13%. It followed by logistic regression, which achieved an accuracy of 76.13% with a sensitivity of 79.59% and a specificity of 72.74%. Wang *et al.* [13] developed a classification approach to identify people at high risk of type 2 diabetes. Total of (6,480) records were selected randomly as a training set to construct two models using an ANNs and Multivariate Logistic Regression (MLR). Total of (2,160) records were used as a validation set for performance comparison purpose. The predictive performance of the ANNs model was 86.93%, 79.14%, 31.86%, and 98.18% for sensitivity, specificity, positive and negative predictive value, respectively. While the predictive performance for MLR model was only 60.80%, 75.48%, 21.78%, and 94.52% in the same order. Also, performance analyzed by Area under the ROC curve (AUC) value; it showed more accurate predictive performance for ANN model (0.891) than the MLP model (0.744).

Since there are numerous studies, demonstrate the superiority of machine learning in predicting diabetes, paying attention to such studies and their results are required. Daghistani and Alshammari [14] had applied three classification techniques to construct a model to predict diabetes. Three machine-learning algorithms used, namely Self-Organizing Map (SOM), C4.5 and RandomForest. Recall and Precision were applied as evaluation criteria to compare the three algorithms. RandomForest achieved recall over 90% and precision over 65% using the test data set. The study by Selvakumar *et al.* [15] classified diabetes data using Binary Logistic Regression, Multilayer Perceptron (MLP) and k-Nearest Neighbor. They reached the best results using k-Nearest Neighbor with an accuracy of 80%.

Although a large number of prediction models being developed, poor performance is ultimately contributed to the usefulness of the models. Essential component affecting the performance of the model is what variables are used to develop that model. It is not a simple to identify risk factors; even we select them from the literature but may not reach statistical significance level [3]. Therefore, discovering novel risk factors are an addition aim of this study. Since the mentioned studies reflect data-driven research, the major gaps in diabetes research using machine learning are 1) the availability of data 2) the size of the dataset to provide proper training for an algorithm. The accessibility of full clinical and diagnostic data in EHR is the ideal way due to low cost. In contrast, data such as biological are expensive and more difficult and to collect. Therefore, it is less available. Moreover, other types of lack of data are lifestyle, behavior, and inheritance [16].

III. Methods

In this section, the methodology of this research article was explained. Description on how the data sets and features were obtained. This section also deals with the algorithms used in this research and their evaluation criteria.

A. Data Set and Features

The data sets included (66,325) records that were collected between 2013 and 2015 for all adult patients who had the Hemoglobin A1c (HgbA1c) test in their record while pediatric diabetic patients were excluded. The HgbA1c was used to identify/classify the patients as diabetic (HgbA1c >=7) or non-diabetic (HgbA1c <7). A few steps were taken for preparing data before analyzing it.

All records were merged, if the patient had multiple hospitalizations in the medical record. That applied by taking the last results for all attributes in each patient record.

All attributes with missing value more than 40% were excluded. The original data included lots of lab tests that do not have values for each patient. The remaining amount of missing data was treated as another attribute value and was processed as it without replacing it with the mean for continuous variable or mode for nominal attributes. However, Logistic Regression algorithm in Weka replaced missing values by using a ReplaceMissingValuesFilter,

and transformed nominal attributes into numeric attributes using a NominalToBinaryFilter.

Irrelevant attributes were excluded such as admission and discharge dates. In practice, such attributes do not increase the accuracy of the model in addition to increasing the complexity of the model.

Domain knowledge technique was employed for handling and removing implausible values. An example of the manual inspection was performed is rejecting errors such as out of range vital signs. Moreover, to clean data set from outliers, we removed values away from either the 25th or the 75th percentile that were detected by using a domain knowledge base. The goal of the research is to assist in identifying diabetic patients who are unaware of being diabetic by utilizing 18 attributes only [14]. A descriptive analysis of the attributes is shown in Table I. The attributes in data sets were categorized as follow:

1) Demographic attributes such as gender, age, and region;

2) Measurement attributes such as the Body Mass Index (BMI) and blood pressure;

3) Lab tests.

TABLE I. DESCRIPTIVE STATISTICS OF DIABETES RISK FACTORS

Risk Factors	Data
Region	
Central	54141 (81.63%)
Eastern	11085 (16.71%)
Western	1099 (1.66%)
Gender	
Male	36811 (55.50%)
Female	29514 (44.50%)
Age	
13-19	578 (0.87%)
20-34	4067 (6.13%)
35-44	4486 (6.76%)
45-64	23949 (36.11%)
65-84	29049 (43.80%)
>85	4196 (6.33%)
Body Mass Index (BMI)	30.77 ±8.92
Blood pressure	
High-BP	128.74 ±18.225
Low-BP	67.71 ±11.154
Lab Test	
eGFR	78.33 ±40.83
Mean corpuscular volume (MCV)	86.954 ±7.589
Mean corpuscular hemoglobin (MCH)	28.03 ±2.91036
Mean Corpuscular hemoglobin concentration (MCHC)	317.55 ±38.99
Red cell volume distribution width (RDW)	15.23 ±2.43
Platelet count (Plt)	273.70 ±125
Mean Platelet Volume (MPV)	8.55 ±1.38
White Blood Cell Count (WBC)	9.35 ±5.81
Red Blood Cell Count (RBC)	4.17 ±0.84
Hemoglobin (Hgb)	114.56 ±26.72
Hematocrit (Hct)	0.91 ±4.44

Values are mean ±SD and n (%).

B. Algorithms

In this study, several experiments were applied using Weka software [17] to select the algorithm that achieved the best performance for the prediction. In comparison to other supervised classification algorithms, RandonForest (decision tree) demonstrated better accuracy. Aside RF measures the importance of each attribute and identifying the most critical predictor among a large number of predictors. The way RF works is by generating several trees then choosing a feature in each node to be the split point randomly. Thus, accuracy improved due to decrease the error rate because of decreasing the correlation between the trees [18].

We also created another model using Logistic Regression (LR) algorithm to predict diabetes in Weka software. LR modeling data within short execution time with a ridge estimator, which provides embedded feature selection capability [19]. S. Le Cessie and J. C. Van Houwelingen LR was performed to classify risk factors for related to diabetes. The logistic function calculates the probability of diabetes y using the values of the predictive risk factors. The patient does not suffer from the disease if $y = 0$; otherwise, $y = 1$) [12].

C. Evaluation Criteria

10-fold cross-validation method was applied. It is a statistical technique working by partitioning the dataset into ten folds with equal size. Nine folds used for model training and the tenth used for model testing. After the tenth iterations were finished, the ten results averaged into single estimation [20].

Several matrices had been applied to select the best model in predicting diabetic patients that are True Positive rate, False Positive rate, Precision, Recall, Area under the Curve and F-measure. The metrics were calculated as:

- True Positive Rate (TPR): represent the number of patients who were classified as high risk of diabetes.
- False Negative Rate (FNR): represent the number of patients who were classified as low risk of diabetes or non-diabetic incorrectly.
- Precision: represents the percentage of diabetic patients that classified as positive and they were positive and it is calcuated based on formula 1

$$\text{Precision} = TP/(TP+FP) \quad (1)$$

- Recall: represent the percentage of diabetic patients that classified correctly and it is calcuated based on formula 2

$$\text{Recall} = TP/(TP+FN) \quad (2)$$

- *F-score*: represents the harmonic mean of precision and recall and it is calcuated based on formula 3

$$2 \times (\text{precision} \times \text{recall}) / (\text{precision} + \text{recall}) \quad (3)$$

- *ROC*: Receiver Operating Characteristic (ROC) Curve: It is a graphically way to display true positives versus false-positives across a series of cut-offs, and select the optimal cut-off for clinical use [21].

IV. RESULTS

As noted above, total of (66,325) records with 18 attributes were used with LR and RF based algorithms.

Male represent 55.50% of the data set, and female constitutes 44.50%. The majority of data either male or female belong to patients with age between 45 and 84 years old. The percentage of diabetic patients in the data set is 64.47%. The incidence of diabetes for both genders absorbed was more in age ranging 65-84 years (47.83%) for male and 48.6% for female than in age range 45-64 by 37.89% for male, 38.03% for female.

For Body Mass Index (BMI) and blood pressure measurements, we observed a female suffer from obesity more than male, the majority of females have 20-40 BMI value. Most males have 20-30 BMI value after that 30-40 value. Most patients in the study have normal blood pressure that represents 84.70% as shown in Table I.

TABLE II. CONFUSION MATRICES OF PREDICTING DIABETES USING ML AND LOGISTIC REGRESSION TECHNIQUES

	LR Model	RF Model
TPR	0.703	0.88
FNR	0.454	0.188
Precision	0.692	0.883
Recall	0.703	0.88
AUC	0.708	0.944
F-measure	0.675	0.876

The accuracy of prediction models is 70.3% and 88% for LR and RF, respectively. The metrics, namely TPR, FNR, Precision, Recall, Area Under the Curve and F-measure, were used to compare the overall performance of the two prediction models, the results of different evaluation matrices were presented in Table II. In comparing the prediction results for predicting diabetes, we found that the overall performance of Logistic Regression is less than the RandomForest. The average AUC values were 0.708 for LR and 0.944 for RF as shown in Fig. 1. As a result of all metrics of performance, RandomForest was considered the optimum prediction model in this study.

In comparing the prediction results for predicting diabetes, we found that the overall performance of LR is less than the RF. The average AUC values were 0.708 for LR and 0.944 for RF as shown in Fig. 1. As a result of all metrics of performance, RF was considered the optimum prediction model in this research study.

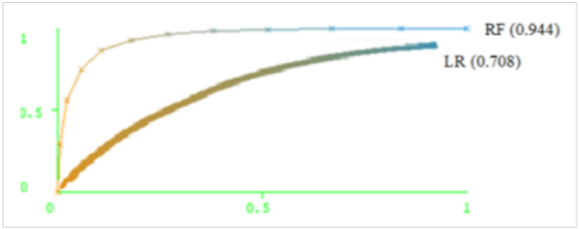

Figure 1. AUC for LR and RF models.

V. DISCUSSION

Early prediction of individuals at high risk of diabetes is an essential challenge in the health domain. In the present study, we compared RamdomForest model and logistic regression model in predicting diabetes based on risk factors. In comparison, machine-learning approaches proven the feasibility of using the massive data collected in electronic health records for diabetes risk forecasting [22]. Logistic Regression (LR) is often used to recognize significant risk factors that correlated with diabetes and has been used to develop a predictive model [13], [23].

As shown in Table I, there are attributes related to diabetes, such as gender, age, body mass index (BMI), blood pressure and 11 lab tests. To the best of our knowledge, this study, expand the list of the factors that used in several previous studies by adding laboratory attributes, so the models include both common risk factors for diabetes and less recognized risk factors. Therefore, discovering novel risk factor is an addition to this study; these factors were selected to build a predictive model for studying diabetes.

The input variables in the LR and RF models contained 11 lab test parameters that are eGFR, Mean Corpuscular Volume (MCV), Mean Corpuscular Hemoglobin (MCH), Mean Corpuscular Hemoglobin Concentration (MCHC), Red cell volume Distribution Width (RDW), Platelet count (Plt), Mean Platelet Volume (MPV), White Blood Cell Count (WBC), Red Blood Cell Count (RBC), Hemoglobin (Hgb) and Hematocrit (Hct). Hence, the generated model reduces the data dimension to few attributes, which allows our method to scale to more of beneficiaries factors. Nevertheless, our study could be generalized only to the Saudi population.

VI. CONCLUSION

In this paper we have analyzed health data collected from MNGHA databases, to perform a comparison of two machine learning algorithms. RandomForest and Logistic Regression used to construct models aiming to predict diabetes. After the generation of the prediction models, we observed that RandomForest performs with more accuracy and less error rate comparisons with Logistic Regression. We conclude that machine learning based algorithm has better prediction performance than the statistical-based algorithm. Future work is to use larger dataset to evaluate the robustness of the models.

APPENDIX LIST OF ABBREVIATIONS

MNGHA	Ministry of National Guard Hospital Affairs
ML	Machine learning
WHO	World Health Organization
Hct	Hematocrit
RF	RandonForest
LR	Logistic Regression
TP	True Positive
FN	False Negative
ROC	Receiver Operating Characteristic
BMI	Body Mass Index
MCV	Mean corpuscular volume
MCH	Mean corpuscular hemoglobin
MCHC	Mean Corpuscular hemoglobin concentration
RDW	Red cell volume distribution width

Plt	Platelet count
MPV	Mean Platelet Volume
WBC	White Blood Cell Count
RBC	Red Blood Cell Count
Hgb	Hemoglobin
Hct	Hematocrit
AROCs	Area under the Receiver Operating Curves
ACM	All-Cause Mortality
EANNs	Evolving Artificial Neural Networks
ANNs	Artificial Neural Networks
MLR	Multivariate Logistic Regression
SOM	Self-Organizing Map
MLP	Multilayer Perceptron

Availability of Data and Material

This project includes diabetes data from Ministry of National Guard Health Affairs (MNGHA), which was collected under IRB approval. According to IRB approval, the data will remain with the principal investigator (Dr. Riyad Alshammari - riyadalshammari@gmail.com) and the study investigators. Data sharing will be only on a collaborative basis under data sharing agreements.

Ethics Approval and Consent to Participate

This study does not contain human participants or animals performed by any of the authors. This study includes diabetes data from Ministry of National Guard Health Affairs (MNGHA), which was collected under IRB approval (IRB #: SP15/064). This study does not include informed consent because the study was a retrospective.

Conflict of Interest

None of the authors has any competing interests.

Author Contributions

TD collected and analyzed data, built data mining models, interpreted and visualized results. RS introduced the idea and designed the study, supervised and guided the study, reviewed and approved the final submission. Both authors read and approved the final manuscript.

Acknowledgment

This study was funded by the King Abdullah International Medical Research Center (KAIMRC), National Guard, Health Affairs, Riyadh, Saudi Arabia with research grant No. SP15/064.

References

[1] P. Kasemthaweesab and W. Kurutach, "Association analysis of Diabetes Mellitus (DM) with complication states based on association rules," in *Proc. 7th IEEE Conference on Industrial Electronics and Applications*, July 2012, pp. 1453-1457.

[2] WHO. Diabetes. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs312/en/

[3] G. S. Collins, S. Mallett, O. Omar, and L. M. Yu, "Developing risk prediction models for type 2 diabetes: A systematic review of methodology and reporting," *BMC Medicine*, vol. 9, no. 1, p. 103, 2011.

[4] WHO. Country and regional data on diabetes. [Online]. Available: http://www.who.int/diabetes/facts/world_figures/en/index2.html

[5] M. Alghamdi, M. Al-Mallah, S. Keteyian, C. Brawner, J. Ehrman, and S. Sakr, "Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project," *PloS ONE*, vol. 12, no. 7, p. e0179805, 2017.

[6] R. Alshammari and N. Almutairi, "Building diabetes early warning system using data mining techniques," *Journal of Medical Imaging and Health Informatics*, vol. 7, no. 3, pp. 655-659, 2017.

[7] K. Zarkogianni, E. Litsa, K. Mitsis, P. Y. Wu, C. D. Kaddi, C. W. Cheng, and K. S. Nikita, "A review of emerging technologies for the management of diabetes mellitus," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 12, pp. 2735-2749, 2015.

[8] N. Razavian, S. Blecker, A. M. Schmidt, A. Smith-McLallen, S. Nigam, and D. Sontag, "Population-level prediction of type 2 diabetes from claims data and analysis of risk factors," *Big Data*, vol. 3, no. 4, pp. 277-287, 2015.

[9] D. O. Shankaracharya, S. Samanta, and A. S. Vidyarthi, "Computational intelligence in early diabetes diagnosis: A review," *The Review of Diabetic Studies: RDS*, vol. 7, no. 4, p. 252, 2010.

[10] M. H. Al-Mallah, R. Elshawi, A. M. Ahmed, W. T. Qureshi, C. A. Brawner, M. J. Blaha, and S. Sakr, "Using machine learning to define the association between cardiorespiratory fitness and all-cause mortality (from the Henry Ford Exercise Testing Project)," *The American Journal of Cardiology*, 2017.

[11] K. Dalakleidi, K. Zarkogianni, A. Thanopoulou, and K. Nikita, "Comparative assessment of statistical and machine learning techniques towards estimating the risk of developing type 2 diabetes and cardiovascular complications," *Expert Systems*, 2017.

[12] X. H. Meng, Y. X. Huang, D. P. Rao, Q. Zhang, and Q. Liu, "Comparison of three data mining models for predicting diabetes or prediabetes by risk factors," *The Kaohsiung Journal of Medical Sciences*, vol. 29, no. 2, pp. 93-99, 2013.

[13] C. Wang, L. Li, L. Wang, Z. Ping, M. T. Flory, G. Wang, and W. Li, "Evaluating the risk of type 2 diabetes mellitus using artificial neural network: An effective classification approach," *Diabetes Research and Clinical Practice*, vol. 100, no. 1, pp. 111-118, 2013.

[14] T. Daghistani and R. Alshammari, "Diagnosis of diabetes by applying data mining classification techniques," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 7, pp. 329-332, 2016.

[15] S. Selvakumar, K. S. Kannan, and S. GothaiNachiyar, "Prediction of diabetes diagnosis using classification based data mining techniques," *International Journal of Statistics and Systems*, vol. 12, no. 2, pp. 183-188, 2017.

[16] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," *Computational and Structural Biotechnology Journal*, vol. 15, pp. 104-116, 2017.

[17] (2017). WEKA software: Machine learning group at the University of Waikato. [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/

[18] K. J. Archer and R. V. Kimes, "Empirical characterization of random forest variable importance measures," *Computational Statistics & Data Analysis*, vol. 52, no. 4, pp. 2249-2260, 2008.

[19] C. Chang, P. A. Verhaegen, and J. R. Duflou, "A comparison of classifiers for intelligent machine usage prediction," in *Proc. International Conference on Intelligent Environments*, June 2014, pp. 198-201.

[20] C. Shao, K. Paynabar, T. H. Kim, J. J. Jin, S. J. Hu, J. P. Spicer, and J. A. Abell, "Feature selection for manufacturing process monitoring using cross-validation," *Journal of Manufacturing Systems*, vol. 32, no. 4, pp. 550-555, 2013.

[21] C. M. Florkowski, "Sensitivity, specificity, Receiver-Operating Characteristic (ROC) curves and likelihood ratios: Communicating the performance of diagnostic tests," *The Clinical Biochemist Reviews*, vol. 29, suppl. 1, p. S83, 2008.

[22] S. Mani, Y. Chen, T. Elasy, W. Clayton, and J. Denny, "Type 2 diabetes risk forecasting from EMR data using machine learning," in *Proc. AMIA Annual Symposium Proceedings*, 2012, vol. 2012, p. 606.
[23] S. C. Lemon, J. Roy, M. A. Clark, P. D. Friedmann, and W. Rakowski, "Classification and regression tree analysis in public health: Methodological review and comparison with logistic regression," *Annals of Behavioral Medicine*, vol. 26, no. 3, pp. 172-181, 2003.

Tahani Daghistani received the BSc in computer and information science in the field of computer application from the king Saud University, and MSc with honor in health informatics from King Saud bi Abdulaziz University for Health Science, Saudi Arabia. Her current research work is focused on Data Mining, Machine Learning and Big data. She has published various research articles in these areas.

Dr. Riyad Alshammari is an Associate Professor in Computer Science, senior Data Scientist, and Joint-Associate Professor in Health Informatics at the department of Health Informatics, King Saud bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia. Dr. Alshammari studied at Dalhousie University, Halifax, N.S., Canada and graduated with a Bachelor of Computer Science in 2005, Master of Computer Science in 2008 and Doctor of Philosophy in Computer Science in 2012. Dr. Alshammari is specialized in big data, data modeling, artificial network and machine learning. Dr. Alshammari research interests include but not limited to the areas of artificial network, Machine Learning, Clinical Informatics, e-Health, and data analytics. He has published his research in leading peer-reviewed journals and international conferences, and has attended many conferences and workshops in the area of data science and artificial intelligent. Dr. Alshammari is a member of many IEEE societies and review committees of several IEEE international conferences and journals.