

Original Paper

Improving Current Glycated Hemoglobin Prediction in Adults: Use of Machine Learning Algorithms With Electronic Health Records

Zakhriya Alhassan^{1,2}, PhD; Matthew Watson¹, MSc; David Budgen¹, PhD; Riyadh Alshammari³, PhD; Ali Alessa⁴, PhD; Noura Al Moubayed¹, PhD

¹Department of Computer Science, Durham University, Durham, United Kingdom

²College of Computer Science and Engineering, University of Jeddah, Jeddah, Saudi Arabia

³National Center for Artificial Intelligence, Saudi Data and Artificial Intelligence Authority, Riyadh, Saudi Arabia

⁴Department of Information Technology Programs, Institute of Public Administration, Riyadh, Saudi Arabia

Corresponding Author:

Noura Al Moubayed, PhD

Department of Computer Science

Durham University

Mountjoy Centre

Durham, DH1 3LE

United Kingdom

Phone: 44 1913 341724 ext 41749

Email: noura.al-moubayed@durham.ac.uk

Abstract

Background: Predicting the risk of glycated hemoglobin (HbA_{1c}) elevation can help identify patients with the potential for developing serious chronic health problems, such as diabetes. Early preventive interventions based upon advanced predictive models using electronic health records data for identifying such patients can ultimately help provide better health outcomes.

Objective: Our study investigated the performance of predictive models to forecast HbA_{1c} elevation levels by employing several machine learning models. We also examined the use of patient electronic health record longitudinal data in the performance of the predictive models. Explainable methods were employed to interpret the decisions made by the black box models.

Methods: This study employed multiple logistic regression, random forest, support vector machine, and logistic regression models, as well as a deep learning model (multilayer perceptron) to classify patients with normal (<5.7%) and elevated (≥5.7%) levels of HbA_{1c}. We also integrated current visit data with historical (longitudinal) data from previous visits. Explainable machine learning methods were used to interrogate the models and provide an understanding of the reasons behind the decisions made by the models. All models were trained and tested using a large data set from Saudi Arabia with 18,844 unique patient records.

Results: The machine learning models achieved promising results for predicting current HbA_{1c} elevation risk. When coupled with longitudinal data, the machine learning models outperformed the multiple logistic regression model used in the comparative study. The multilayer perceptron model achieved an accuracy of 83.22% for the area under receiver operating characteristic curve when used with historical data. All models showed a close level of agreement on the contribution of random blood sugar and age variables with and without longitudinal data.

Conclusions: This study shows that machine learning models can provide promising results for the task of predicting current HbA_{1c} levels (≥5.7% or less). Using patients' longitudinal data improved the performance and affected the relative importance for the predictors used. The models showed results that are consistent with comparable studies.

(*JMIR Med Inform* 2021;9(5):e25237) doi: [10.2196/25237](https://doi.org/10.2196/25237)

KEYWORDS

glycated hemoglobin HbA_{1c}; prediction; machine learning; deep learning; neural network; multilayer perceptron; electronic health records; time series data; longitudinal data; diabetes

Introduction

Background

The level of glycated hemoglobin (HbA_{1c}) is used to measure the average glucose concentration in red blood cells [1,2]. Unlike other glucose blood tests, such as random blood sugar (RBS) and fasting blood sugar (FBS), HbA_{1c} provides a long-term measure of a patient's blood glucose levels [3]. The HbA_{1c} test can therefore provide physicians with a reliable means of monitoring a patient's hyperglycemia without requiring the patient to undertake overnight fasting prior to being tested.

A concentration of 6.5% for the HbA_{1c} in patient blood is considered as the cutoff point for the diagnosis of diabetes [4]. However, patients with a concentration of less than 6.5% are not completely excluded from a diabetes diagnosis, as the range of elevation levels ($5.7\% \leq \text{HbA}_{1c} < 6.5\%$) can indicate the future onset of diabetes. Therefore, HbA_{1c} can act as an early predictor for the potential development of type-2 diabetes mellitus (T2DM) [2]. Ackermann et al [3] suggested using the HbA_{1c} test as a measure for identifying those adults who are at a greater risk of developing T2DM in the future.

Research has shown that reducing HbA_{1c} levels can significantly reduce the possibility of developing serious complications. Hence, close monitoring of HbA_{1c} levels is recommended for all diabetic patients and those with the potential for developing diabetes [5]. It is also suggested that diabetic and nondiabetic patients with raised HbA_{1c} levels should be clinically checked and monitored as a preventive intervention to avoid developing T2DM [6].

Currently, the clinical data collected from patient visits consists of a set of readings for vital signs and lab tests, diagnoses, physicians' notes, and treatments that are stored in electronic health records (EHRs). These are collected on an irregular basis, according to clinical needs, and stored with an associated time stamp.

In recent years, machine learning models have shown powerful capabilities for analyzing and understanding complex data across a wide variety of applications. Our research question for this study was as follows: "Can HbA_{1c} prediction be improved by using machine learning with longitudinal data that are normally available in EHR systems?"

This paper reports an investigation into the performance of machine learning models to predict current HbA_{1c} levels as a binary classification problem using EHR data. Nondiabetic patients with an HbA_{1c} level of 5.7% or more are considered to have an elevated HbA_{1c}, while those with levels lower than this are considered normal. The models combine current visit data with extra features (independent variables) extracted from previous visits by patients. We used explainable methods to rank the features in order of their importance to the decision made by each of the models. To the best of our knowledge, this study is the first to employ machine learning models that use longitudinal data from EHR systems for the purpose of HbA_{1c} elevation risk prediction. This study is also the first to use

explainable machine learning techniques to explain the classification decisions made by black box models, support vector machine (SVM), and multilayer perceptron (MLP), in predicting HbA_{1c} elevation risk ($\geq 5.7\%$), in order to better understand the behavior of the model.

Related Work

EHR data have been intensively investigated for a variety of medical decision support tasks [7]. These tasks include the analysis of complex patterns and prediction of major medical events (for example, diagnostic imaging and gene interactions) [8,9]. Several studies have demonstrated the successful employment of EHR data with prediction models [10]. For instance, machine learning has been intensively used with EHR data in diagnosing diabetes and discovering its related patterns [11-15]. However, we are not aware of any studies that have explored machine learning models for the prediction of current elevated HbA_{1c} levels using EHR data from a nondiabetic population or the impact of patient longitudinal data on the effectiveness of such predictive machine learning models.

Several studies have investigated the association between HbA_{1c} levels and clinical variables using statistical models [16,17]. A study by Rose et al [18] discussed the correlation between RBS and HbA_{1c} levels. Stanley et al [19] used a linear regression model for imputation of missing HbA_{1c} data. Their model calculates HbA_{1c} levels for patient records with missing HbA_{1c} values as continuous and categorical values and uses 4 predictors extracted from an EHR system—RBS, FBS, age, and gender—as predictors to calculate the level of HbA_{1c} for a diabetic population. Simone et al [20] used linear regression models to predict HbA_{1c} levels after 6 years for nondiabetic patients using different populations.

A study by Wells et al [21] in 2018 was the first to focus on predicting current HbA_{1c} elevation levels for nondiabetic patients through use of an EHR data set. Multiple logistic regression (MLR) was employed to calculate the probability of a patient having an elevated HbA_{1c} level ($\geq 5.7\%$). The data set was extracted from an EHR system used in the United States. The authors used 8 independent variables fitted to the model using restricted cubic splines with 3 knots to formulate the final equation. The performance of the MLR model was compared to that of the models used by Baan et al [22] and Griffin et al [23]. However, the models by Baan and Griffin aimed at predicting the onset of patients' diabetes rather than predicting HbA_{1c} levels for nondiabetic patients. In addition, the experimental data set used by Wells et al to train and test their model was imbalanced with 74% of the samples having normal HbA_{1c} levels (5.7%) and only 26% of the samples having elevated HbA_{1c} levels ($\geq 5.7\%$).

We performed a differentiated replication of the study by Wells et al [21] using the more balanced King Abdullah International Medical Research Center (KAIMRC) data set [24]. Although the significant variables identified in our replication were in general agreement with those of the original study, there were some differences in the ranking of importance for these,

suggesting that such models do need to be “tuned” to the characteristics of different populations.

Methods

Study Design

To study the impact of using advanced predictive models with EHR data to predict current HbA_{1c} levels, we employed the MLR, random forest (RF), SVM, and logistic regression (LR) models, as well as a deep learning model, MLP [25]. The problem was formulated into a binary classification problem whereby the target variable, HbA_{1c} level, was encoded as 1 when the level of HbA_{1c} was 5.7% or more and with 0 otherwise. The results obtained from using these models were compared to those obtained from employing the model used by Wells et al with the KAIMRC data set (detailed in the Data Set subsection).

The performance of the models was investigated using current visit data only and with additional longitudinal data from current and previous visits. The performance of each model was evaluated using measures commonly employed in clinical applications. For the SVM and MLP models, the relative importance of the features was also calculated using explainable machine learning techniques.

Explainable Methods for Black Box Models

Using black box machine learning models in health care can have adverse effects on the trust and confidence placed in their outcomes; the risk of misclassification is potentially too high for clinicians to confidently use black box models for high risk health care decisions, and not being able to interpret a model’s decision exacerbates this problem [26]. Explainable methods for machine learning models allow interpretable outcomes that can expose the reasons behind the decision made by the model [27]. This transparency provides both health professionals and patients with the confidence and trust in the outcome of the models. The widely used Shapley Additive Explanations (SHAP) values [28] and local interpretable model-agnostic explanations (LIME) score [29] techniques have therefore been used to provide a degree of transparency to our deep learning model.

SHAP values are derived from Shapley values used in game theory and provide a method of calculating the contribution of each feature (variable) to the final prediction via the GradientSHAP approximation. This is achieved for each feature by comparing the prediction the model makes when the feature is present with the prediction obtained when the feature takes some baseline value [28]. Consequently, the SHAP values for a given input “explain” how each feature affects the output of the model when compared to the baseline (or “default”) output of the model. We used SHAP values to interpret our black box models, so they could be efficiently calculated, and their use enabled a global view of the model to be constructed through the computation of SHAP values from across the whole data set.

SHAP values were computed using the feature’s mean marginal contribution across different coalitions of all features. SHAP

values themselves are computationally intensive to compute, and so approximation methods are commonly used when calculating the values.

To ensure that the SHAP values we calculated were not too greatly affected by the approximation method used, we also computed the LIME [29] scores for the models across the entire data set. LIME tries to estimate locally faithful linear explanations (ie, explanations that correspond to how the model behaves around the instance being explained) for any classifier. LIME achieves this by creating local linear classifiers that approximate the behavior of the original model in the vicinity of the data being explained. As linear models are inherently interpretable through their parameters, they can be used to generate explanations of the original model. Both SHAP and LIME have the advantage that they are model-agnostic techniques, and so we were able to apply both methods to both of our black box classification models (SVM and MLP).

Data Set

The data used in this study were taken from the KAIMRC data set. The data were collected from King Abdulaziz Medical City located in the central and western regions of Saudi Arabia, an area which has been ranked second in the Middle East and seventeenth in world in diabetes prevalence by the World Health Organization (WHO) [30]. According to the International Diabetes Federation, the diabetes prevalence rate in Saudi Arabia is 18.3%. Therefore, the availability of the data from this population provides considerable opportunities for research into the early prediction of diabetes.

The data set contains a full history of patient details, vital signs, and lab test readings for each patient visit for the period from 2016 to the end of 2018. As the aim of this study was to identify nondiabetic patients that are at a high risk of HbA_{1c} elevation, all patients previously diagnosed with hyperglycemia were excluded from the experimental data set. The remaining cohort formed our experimental data set and was categorized by using the American Diabetes Association’s guidelines [31], in which patients with HbA_{1c} readings of more than 5.7% are considered as being in the prediabetic range, while those with less than 5.7% are considered to be in the normal range.

Most medical data sets are imbalanced [32-34]. These imbalances occur when the proportion of one class of patients in the data set is greater than its counterpart class [35,36]. However, unusually, our experimental data set was not imbalanced. Slightly over half of the patients in our experimental data set (9826/18,844, 52.14%) were found to have elevated levels of HbA_{1c} ($\geq 5.7\%$) while 47.86% (9018/18,844) of patients had normal HbA_{1c} levels ($< 5.7\%$). This can be ascribed to the high incidence of diabetes in the region from which the data set was collected [37].

A detailed illustration of the patients’ class distribution (HbA_{1c} levels) by age groups and gender is shown in Figure 1. This shows that as the age of patients increased, so did the proportion of patients who had elevated HbA_{1c} levels. The data set also exhibited a balanced gender distribution, with 49.40% (9308/18,844) of the patients being male and 50.60%

(9536/18,844) being female. However, the proportion of male patients with elevated levels of HbA_{1c} ($\geq 5.7\%$) was greater than that of the female patients. Also, female patients with normal

levels of HbA_{1c} ($< 5.7\%$) made more visits than did males. Table 1 shows the profile for the distribution of HbA_{1c} elevation levels organized by gender.

Figure 1. HbA_{1c} elevation levels distributed over age range and gender in the King Abdullah International Medical Research Center (KAIMRC) data set (before sampling). HbA_{1c}: glycated hemoglobin.

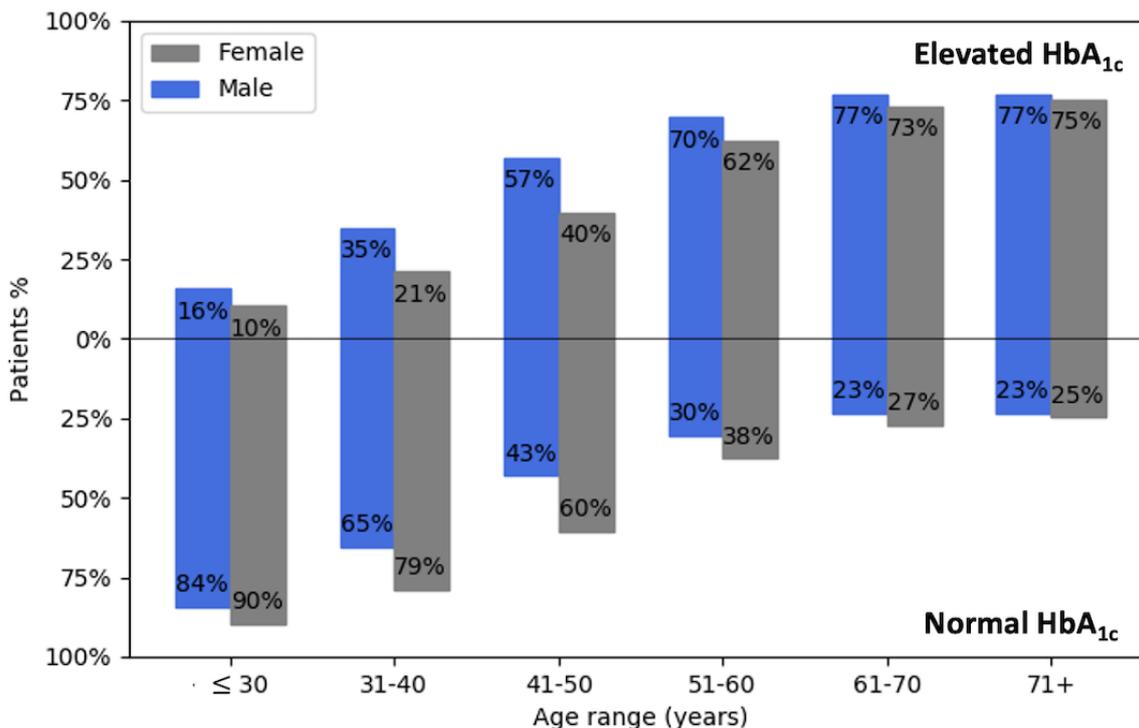


Table 1. Profile for the class distribution over gender.

Characteristics	HbA _{1c} ^a <5.7%, n/N (%)	HbA _{1c} $\geq 5.7\%$, n/N (%)
Number of patients (N=18,844)		
Total	9018/18,844 (47.86)	9826/18,844 (52.14)
Male	3764/9018 (41.74)	5544/9826 (56.42)
Female	5253/9018 (58.26)	4282/9826 (43.58)
Number of visits (N=157,600)		
Total	79,607/157,600 (50.51)	77,993/157,600 (49.49)
Male	31,620/79,607 (39.72)	41,591/77,993 (53.32)
Female	47,987/79,607 (60.28)	36,402/77,993 (46.68)

^aHbA_{1c}: glycated hemoglobin.

Feature Selection and Data Sampling

Six main variables (features) were extracted from the KAIMRC EHR data set to be used in this study. These features, which were selected first for their theoretical association with hyperglycemia and second for their availability in the KAIMRC data set, were the following: age, BMI, estimated glomerular

filtration rate (eGFR), RBS, total cholesterol, and non-high-density lipoprotein. The lab codes of the features used are available in Multimedia Appendix 1 Table S1. The descriptive statistics (using the data for the current visit only for unique patients), units, and P values for the selected features are presented in Table 2.

Table 2. Descriptive statistics of the selected features from the King Abdullah International Medical Research Center (KAIMRC) data set.

Feature	HbA _{1c} ^a 5.7%, mean (SD)	HbA _{1c} 5.7%, mean (SD)	P value
Age (years)	43.94 (16.38)	58.92 (15.12)	<0.001
BMI (Kg/m ²)	29.11 (6.75)	30.90 (6.55)	<0.001
eGFR ^b (ml/min/1.73 m ²)	100.03 (29.22)	85.81 (28.239)	<0.001
RBS ^c (mmol/L)	5.45 (1.26)	7.88 (4.19)	<0.001
CHOL ^d mean (mmol/L)	4.65 (1.07)	4.42 (1.20)	<0.001
non-HDL ^e mean (mmol/L)	3.45 (1.01)	3.37 (1.115)	<0.001

^aHbA_{1c}: glycated hemoglobin.

^beGFR: estimated glomerular filtration rate.

^cRBS: random blood sugar.

^dCHOL: total cholesterol.

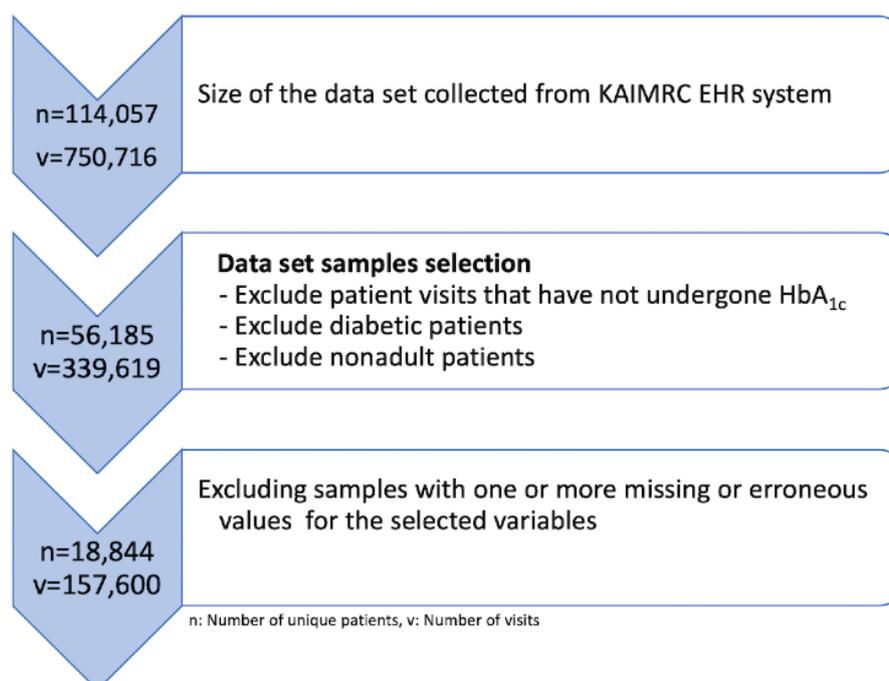
^enon-HDL: non-high-density lipoprotein.

It is very common in clinical practice that physicians may require some lab tests and vital signs to be frequently recorded. In these cases, the average value of all readings taken on a given day (the basic time interval used for this study) was used. For inpatient visits, only data for the first day were considered, and, where there were missing values, the first available values from the visit were used.

For the purpose of this study, we aimed at predicting the HbA_{1c} levels ($\geq 5.7\%$) for current (last) patient visits only. Unlike the sampling approach used by Wells et al, which was based on independent hospital visits for patients (including for the same

patients), the sampling approach used in this study included independent patients to ensure only unseen patients data were used for testing the models. Although we aimed to identify patients with elevated levels of HbA_{1c} from a nondiabetic population, patients previously diagnosed with diabetes were excluded. We also excluded nonadult patients and those with erroneous or missing values [24]. Figure 2 shows the details of the tasks performed to refine the sample selection. This resulted in a reduction in the size of the experimental data set from 114,057 patients with 750,709 visits to 18,844 unique patients with 157,600 visits.

Figure 2. Details of the sampling approach performed on the KAIMRC data set. EHR: electronic health record; HbA_{1c}: glycated haemoglobin; KAIMRC: King Abdullah International Medical Research Center.



The inputs (input features space) for the models used in this study were continuous values. Values for age, eGFR, RBS and total cholesterol features were directly available in the KAIMRC

data set. The values for the BMI and non-high-density lipoprotein variables were calculated from other available features using the formulae in [Multimedia Appendix 2](#).

Input Preparation for the Models

The input structure for the deep learning model was organized as a matrix, based on current and previous time-stamped patient visits. It contained the current visit data concatenated with approximated values for the selected features from all previous visits, which we refer to as the “Approximated Time Series Data”.

Each patient visit was described by the selected features, represented as x_1, x_2, \dots, x_n . These features were formed as episodes based on the time-stamped values available in each visit (v_i).

$$\text{Input} = \begin{bmatrix} v_1 & x_{11} & x_{12} & \cdots & x_{1n} \\ v_2 & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ v_s & x_{s1} & x_{s2} & \cdots & x_{sn} \end{bmatrix} \quad (1)$$

Here, x_{ij} is the feature value at a patient visit ($0 < i \leq s, 0 < j \leq n$); s is the number of time series steps (the length of the input sequence); and n is the number of features for each time step, which was set to 6 as explained earlier.

If the number of visits (longitudinal time series visits) for a patient was fewer than s , the input for this patient was padded out with the mean value of the available visits to compensate for the missing time series data (Multimedia Appendix 3 shows an example of the padding approach used). Where the number of longitudinal visits for a patient was more than s , the piecewise aggregation approximation (PAA) technique [38] was applied to the data for these visits to account for all data from patient visits.

PAA transforms the longitudinal time series data using s as a number of sliding windows (or segments) into a reduced number of time steps data (approximated) employing the mean value of the series falling within that window (segment) [39]. We tested the models with several values for the size of the sliding window (s), and 3 was shown to be the optimal value. The formula used to calculate the approximated time-series data was as follows:

$$\tilde{x}_i = \frac{s}{r} \sum_{j=(r-\frac{1}{s})(i-1)+1}^{(r-\frac{1}{s})i} x_j, s < r - 1 \quad (2)$$

Where \tilde{x}_i represents the approximated value for x , r is the total number of visits for a patient, and s is the reduced number of time series steps (Multimedia Appendix 4 shows an example of the PAA technique used).

The approximated time series data forming the output of the PAA was then concatenated with the current visit data to form the final input for the deep learning model. As the MLR, RF, SVM, and LR models are not capable of handling multidimensional data (formed as matrices), the output of the PAA was reorganized for these into a single-dimensional input by vectorizing the matrix used in equation 1 as below:

$$\text{Input} = [x_{11}x_{12}x_{13} \dots x_{sn}] \quad (3)$$

The last data preprocessing task before training the predictive models was data scaling. The experimental data set was scaled using the normalization technique that rescales the ranges of each of the features to be between 0 and 1 using minimum and maximum values of that feature.

Predictive Models and Experimental Setups

As a baseline comparison, we employed the MLR model used by Wells et al [21], and compared the results from this with those from 4 commonly used machine learning models.

The MLR model is used to create a mathematical equation that can best calculate the probability of a value by assigning weights (coefficients) to the independent variables (features) based on their importance [40]. In this study we employed the same approach used by Wells et al by which the continuous features were fitted into the MLR model using restricted cubic splines technique with 3 knots. When we used the longitudinal input, the variables that caused collinearity were excluded.

Random forest is an algorithm very commonly used for classification. It combines several decision trees that are generated during the training process. Each decision tree is trained using a random subset of the training data set. The final classification is then based on the majority voting results of all generated decision trees [41]. The quality function used in the employed RF model is the Gini importance, with a value of 100 for the number of tree parameters.

Logistic regression is commonly used to solve binary classification problems. It calculates the odds ratio of the variables and is similar to MLR but uses a binomial distribution of the dependent variable (ie, more than 1). Thus, it includes a logit function that handles different types of relationships between the dependent and independent variables [42,43].

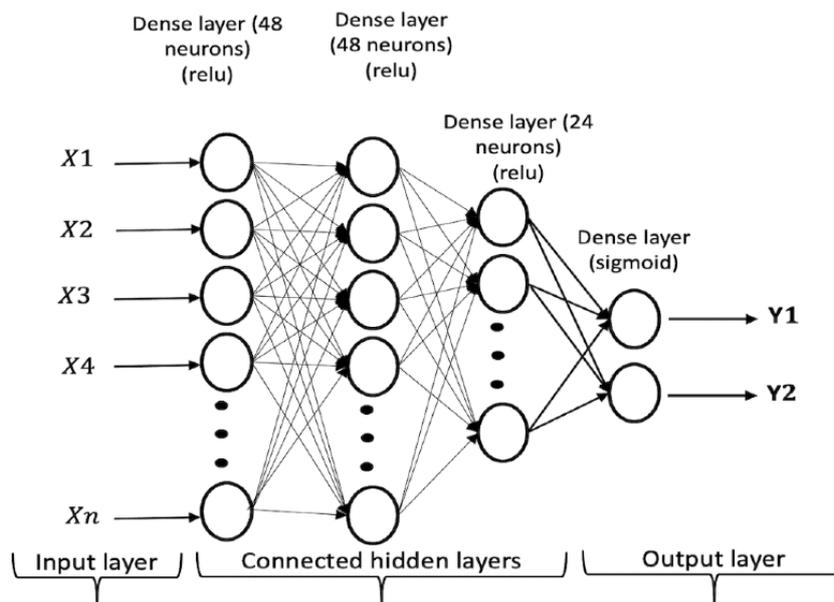
Support vector machine was introduced by Vapnik [44] in 1998. It can solve both classification and regression problems. It uses the training feature space to decide on the separation boundaries (hyperplane) that best divides the training data set into regions, 1 for each class. The very close points to the hyperplanes are the support vectors. SVMs also use kernels to help enhance class separation by mapping the training features into a higher dimensional space with an increased number of dimensions [44,45]. The kernel function used in the SVM model employed is a radial base function with a value of 1 for the cost parameter (C).

A multilayer perceptron, also known as a feed-forward neural network, is one of the most common deep learning approaches. It is mainly used to address supervised learning problems by learning the dependencies between the input layer (the features or variables) and output layer (the classification decision) using a fully connected hidden layer in between. The layers, including hidden ones, contain a number of neurons that are connected to the neurons of the next and previous layers via weights and nonlinear functions. MLP uses a backpropagation algorithm to update the weights and biases within the hidden layers to minimize the output error rate [25,46].

To optimize the MLP model, fine-tuning of the structure and hyperparameters was performed and involved the number of hidden layers and neurons, activation functions, optimizers, and loss functions. The optimized structure of the MLP model used in this study contained 3 hidden layers. The number of neurons in the hidden layers were 48, 48, and 24, respectively. The final layer (the output layer) contained 2 neurons for the final output

of the model ($Y1$ for normal HbA_{1c} or $Y2$ for elevated HbA_{1c}). A rectified linear unit activation function was used in the 3 hidden layers, while a sigmoid was used in the output layer. The detailed structure of the MLP model is shown in [Figure 3](#). The model was trained using an Adam optimizer with mean squared error as the loss function.

Figure 3. The structure used for multilayer perceptron trained with the longitudinal data. relu: rectified linear unit.



Evaluation of Model Performance

The models all employed the same data preprocessing, training, and testing techniques. The models were validated using the 10-fold cross-validation technique. The k-fold cross-validation is one of the most commonly used approximation approaches for validating the obtained results [47,48]. For the MLP model, 100 epochs were used to train each fold.

As our measure for evaluating and comparing the performance of the proposed models, we used the area under the receiver operating characteristic (AUC-ROC) curve, which is equal to the concordance statistic [49]. We also report values for a set of measures that are commonly used in clinical applications: balanced accuracy (that calculates the recall average for each class), overall accuracy, F score, precision, and precision-recall area under the curve (PR-AUC).

To determine the importance that the black box models (SVM and MLP) place upon each variable, we first computed the SHAP values and LIME scores for all samples in our data set and then calculated the average absolute SHAP value and LIME score for each predictor.

Results

[Table 3](#) shows the performance metrics obtained using the MLR, RF, SVM, LR, and MLP models with and without the longitudinal data. The results show that the models achieved competitive performance using the reported measures. The LR and MLP models trained with and without the longitudinal data achieved better performance with regards to the AUC-ROC measure than did the MLR (statistical model employed by Wells et al) or the RF and SVM models (more details about AUC-ROC and PR-AUC curve plots are presented in [Multimedia Appendix 5](#)). The results also show that the SVM, LR, and MLP models trained with and without the longitudinal data achieved better performance than did the MLR and RF models using the balanced accuracy measure.

[Table 3](#) also shows that all models, including the MLR, achieved better performance using all reported measures when they were trained with the features from patients' longitudinal data. The MLP with longitudinal data slightly outperformed all other models with respect to the reported measures.

Table 3. Classifiers performance for current glycated hemoglobin level prediction.

Model	AUC-ROC ^a , % (SD)	Balanced accuracy, % (SD)	Accuracy, % (SD)	F score, % (SD)	Precision, % (SD)	PR-AUC ^b , % (SD)
MLR^c						
No ^d	81.38 (3.82)	72.74 (4.15)	73.59 (3.79)	74.91 (5.12)	73.20 (5.05)	82.14 (6.04)
Yes ^e	82.45 (4.09)	73.49 (4.19)	74.30 (4.02)	75.11 (6.00)	74.36 (5.26)	83.45 (6.29)
RF^f						
No	80.82 (1.14)	72.57 (1.17)	72.64 (1.14)	73.97 (1.04)	73.42 (1.84)	82.03 (1.35)
Yes	82.38 (1.04)	73.86 (0.98)	73.91 (0.95)	75.07 (0.86)	74.81 (1.68)	84.06 (1.17)
SVM^g						
No	81.05 (1.04)	73.69 (1.35)	73.88 (1.33)	75.76 (1.18)	73.42 (1.90)	80.56 (1.48)
Yes	82.04 (0.89)	74.25 (1.11)	74.40 (1.08)	76.08 (0.92)	74.20 (1.65)	83.16 (1.19)
LR^h						
No	81.51 (1.26)	73.18 (1.10)	73.17 (1.08)	73.96 (1.03)	74.88 (1.69)	82.49 (1.46)
Yes	82.59 (1.04)	74.11 (1.15)	74.05 (1.13)	74.55 (0.98)	76.31 (1.72)	84.13 (1.04)
MLPⁱ						
No	82.07 (1.06)	73.61 (1.04)	73.83 (1.03)	75.87 (1.10)	73.07 (1.62)	83.42 (1.19)
Yes	83.22 (0.92)	74.45 (1.18)	74.55 (1.18)	75.99 (1.95)	74.78 (2.07)	84.85 (0.78)

^aAUC-ROC: area under the receiver operating characteristic.

^bPR-AUC: precision-recall area under the curve.

^cMLR: multiple logistic regression.

^dWithout longitudinal data.

^eWith longitudinal data.

^fRF: random forest.

^gSVM: support vector machine.

^hLR: logistic regression.

ⁱMLP: multilayer perceptron.

Figure 4 summarizes the 10-fold performance achieved for the set of measures where the models were trained without longitudinal data, and Figure 5 shows the performance where they were trained with the longitudinal data. Both figures show a more consistent prediction trend for RF, LR, SVM, and MLP with and without longitudinal data, as the measures for these

models show a small variation between the folds. As shown in Figure 4 and Figure 5, the SD values for MLR with and without longitudinal data are larger than those for the other models. This indicates that the machine learning models used can not only enhance the performance, but can also improve the classification confidence for HbA_{1c} prediction.

Figure 4. Box plot showing the detailed 10-fold performance of all models trained without longitudinal data. AUC-ROC: area under the receiver operating characteristic; LR: logistic regression; MLP: multilayer perceptron; MLR: multiple logistic regression; PR-AUC: precision-recall area under the curve; RF: random forest; SVM: support vector machine.

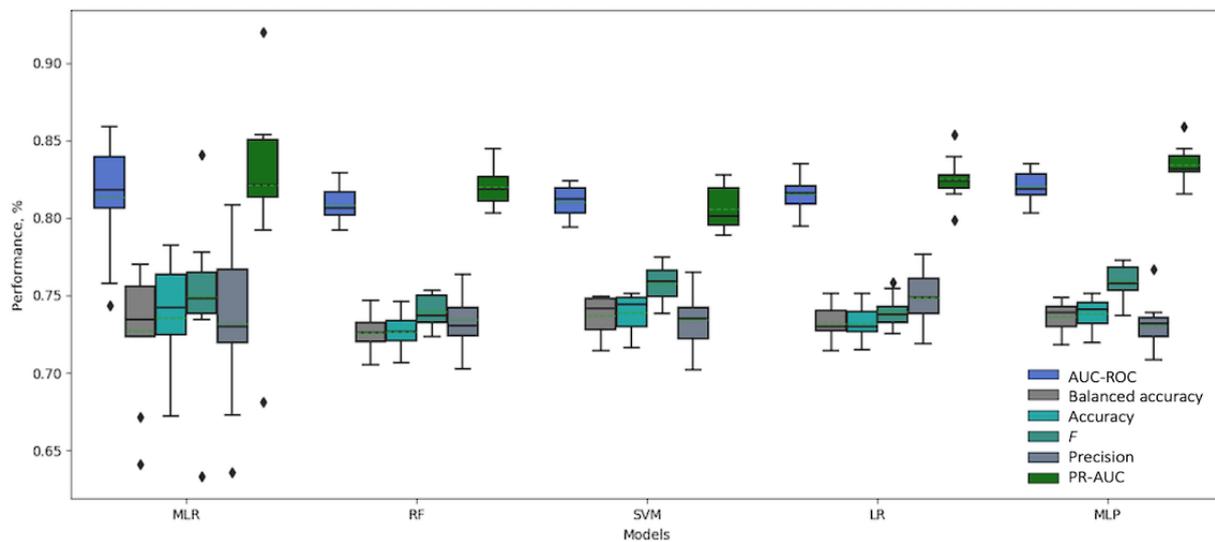


Figure 5. Boxplot showing the detailed 10-fold performance of all models trained with longitudinal data. AUC-ROC: area under the receiver operating characteristic; LR: logistic regression; MLP: multilayer perceptron; MLR: multiple logistic regression; PR-AUC: precision-recall area under the curve; RF: random forest; SVM: support vector machine.

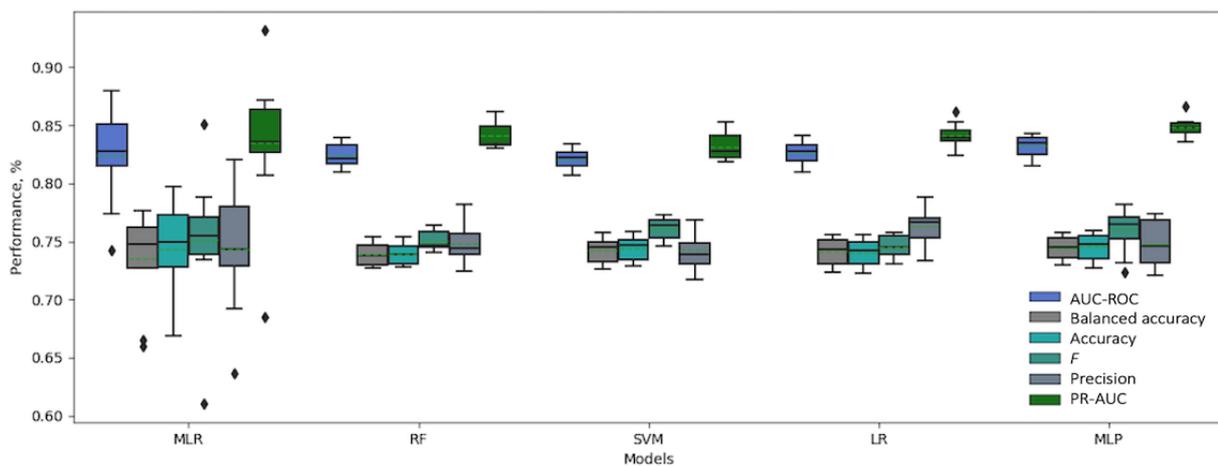


Table 4 shows the ranked order of importance of the set of predictors used for training the models. Further details on the actual importance values for each model are provided in Multimedia Appendix 6 (refer to Multimedia Appendix 7 for more details of the MLR and LR calculator). Calculating the importance of the predictors for the MLR models using

vectorized longitudinal data was not possible due to the collinearity caused by having multiple variables for BMI. The order of importance results obtained using the SHAP method for both the SVM and MLP were identical to those obtained using LIME and provided greater confidence in the explainable methods used (see Multimedia Appendix 6).

Table 4. Order of importance of predictors for the models.

Model	Importance rank					
	1st	2nd	3rd	4th	5th	6th
MLR^a						
No ^b	Age	RBS ^c	BMI	CHOL ^d	Non-HDL ^e	eGFR ^f
RF^g						
No	Age	RBS	BMI	eGFR	CHOL	Non-HDL ^h
Yes ^h	RBS	Age	CHOL	eGFR	Non-HDL	BMI
LRⁱ						
No	RBS	Age	Non-HDL	CHOL	BMI	eGFR
Yes	RBS	Age	Non-HDL	eGFR	CHOL	BMI
SVM^j (SHAP^k & LIME^l)						
No	Age	RBS	BMI	Non-HDL	CHOL	eGFR
Yes	RBS	Age	CHOL	Non-HDL	BMI	eGFR
MLP^m (SHAP & LIME)						
No	RBS	Age	Non-HDL	CHOL	BMI	eGFR
Yes	RBS	Age	eGFR	CHOL	Non-HDL	BMI

^aMLR: multiple logistic regression.

^bWithout longitudinal data.

^cRBS: random blood sugar.

^dCHOL: total cholesterol.

^enon-HDL: non-high-density lipoprotein.

^feGFR: estimated glomerular filtration rate.

^gRF: random forest.

^hWith longitudinal data.

ⁱLR: logistic regression.

^jSVM: support vector machine.

^kSHAP: Shapley Additive Explanations.

^lLIME: local interpretable model-agnostic explanations.

^mMLP: multilayer perceptron.

Table 4 and the figures in Multimedia Appendix 6 show that all of the models were heavily and interchangeably reliant on age and RBS when making classification decisions. The RF and SVM models, when trained with longitudinal data, ranked RBS over age. Figure 6 and Figure 7 highlight the importance that

our best performing model, MLP, placed upon the features in our data set using SHAP and LIME, respectively. Both figures show that the RBS contributed the most to the MLP's final prediction, while the patient's BMI contributed the least.

Figure 6. Relative importance of predictors obtained from the multilayer perceptron trained with longitudinal data using SHAP. CHOL: total cholesterol; eGFR: estimated glomerular filtration rate; non-HDL: non-high-density lipoprotein; RBS: random blood sugar; SHAP: Shapley Additive Explanations.

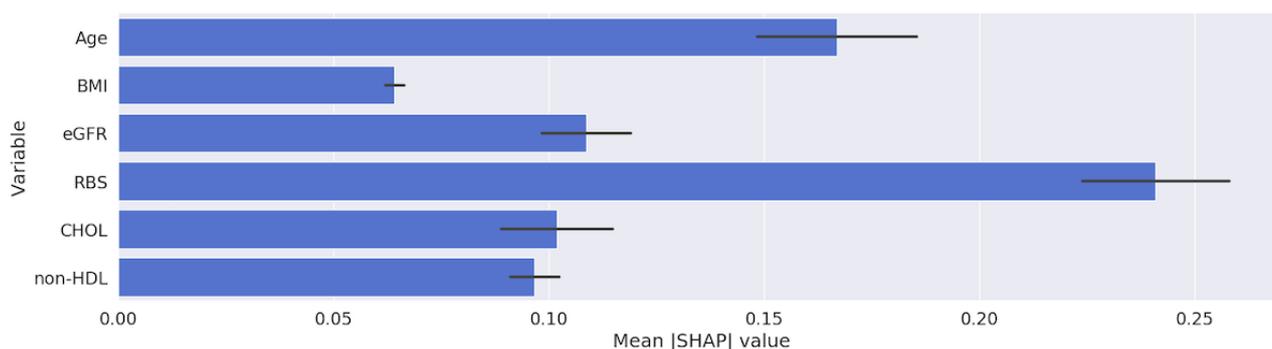
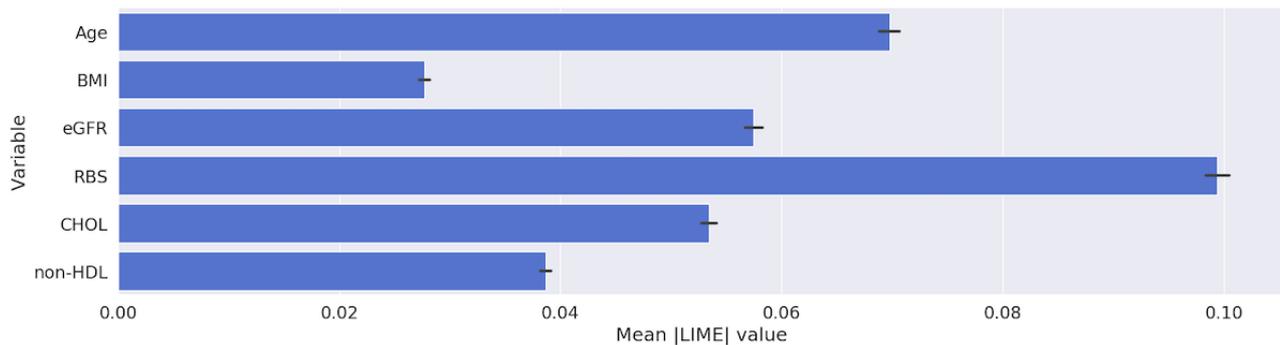


Figure 7. Relative importance of predictors obtained from multilayer perceptron trained with longitudinal data using LIME. CHOL: total cholesterol; eGFR: estimated glomerular filtration rate; LIME: local interpretable model-agnostic explanations; non-HDL: non-high-density lipoprotein; RBS: random blood sugar.



For all models trained with longitudinal data, BMI was ranked lower than when the models were trained without longitudinal data. However, the importance value produced for the BMI variable from the models was still not insignificant (see the figures in Multimedia Appendix 7). This indicates that models are able to find subtle relationships in the longitudinal data that are more relevant to the prediction than is BMI, rendering it less important.

When MLP and LR models trained on the longitudinal data were used, the eGFR variable was ranked higher than total cholesterol and BMI, in contrast to when these were trained on the current visit only. None of the other models trained with the current visit only, except for RF, considered it important. Again,

we ascribe this to the information that the model learns from the variations of eGFR values between a patient’s visits (longitudinal EHR data).

SHAP values are calculated on the sample level. Figures 8 and 9 illustrate the SHAP values for 2 randomly selected sample patients from our data set. These figures highlight how different inputs have different SHAP values. The patient in Figure 8 (for whom our model correctly predicted elevated HbA_{1c} levels of ≥5.7%) had a higher RBS value than did the patient in Figure 9 (for whom our model correctly predicted normal HbA_{1c} levels of <5.7%). This explains why our MLP model placed much more importance on the RBS value of the patient in Figure 6.

Figure 8. An example showing the SHAP values for a randomly selected sample with elevated glycated hemoglobin levels (≥5.7%). CHOL: total cholesterol; eGFR: estimated glomerular filtration rate; non-HDL: non-high-density lipoprotein; RBS: random blood sugar; SHAP: Shapley Additive Explanations.

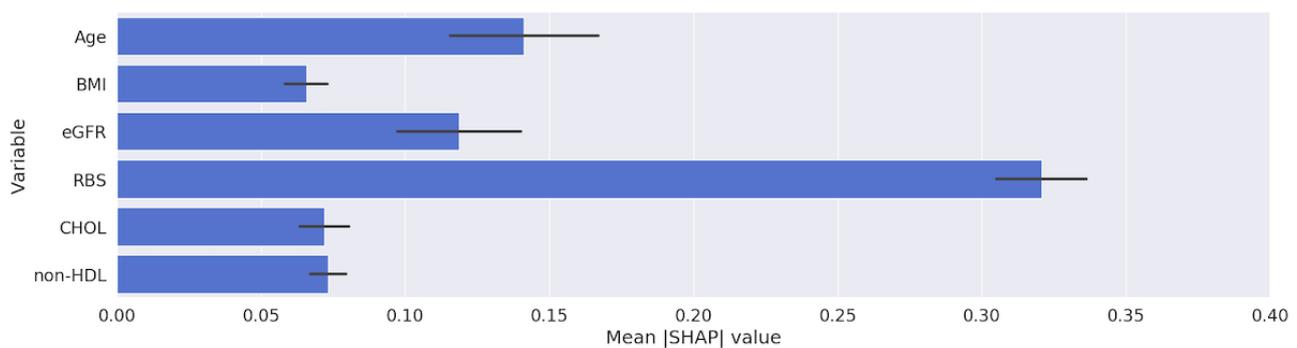
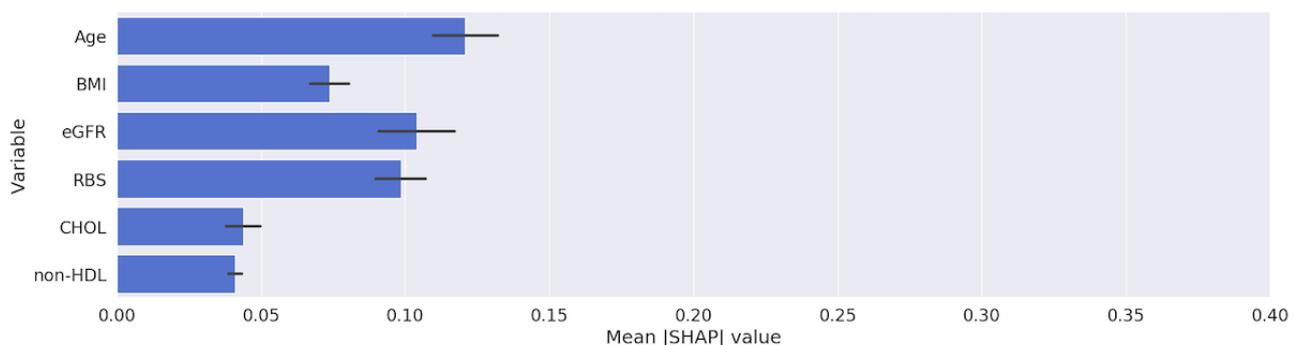


Figure 9. An example showing the SHAP values for randomly selected sample with normal glycated hemoglobin levels (<5.7%). CHOL: total cholesterol; eGFR: estimated glomerular filtration rate; non-HDL: non-high-density lipoprotein; RBS: random blood sugar; SHAP: Shapley Additive Explanations.



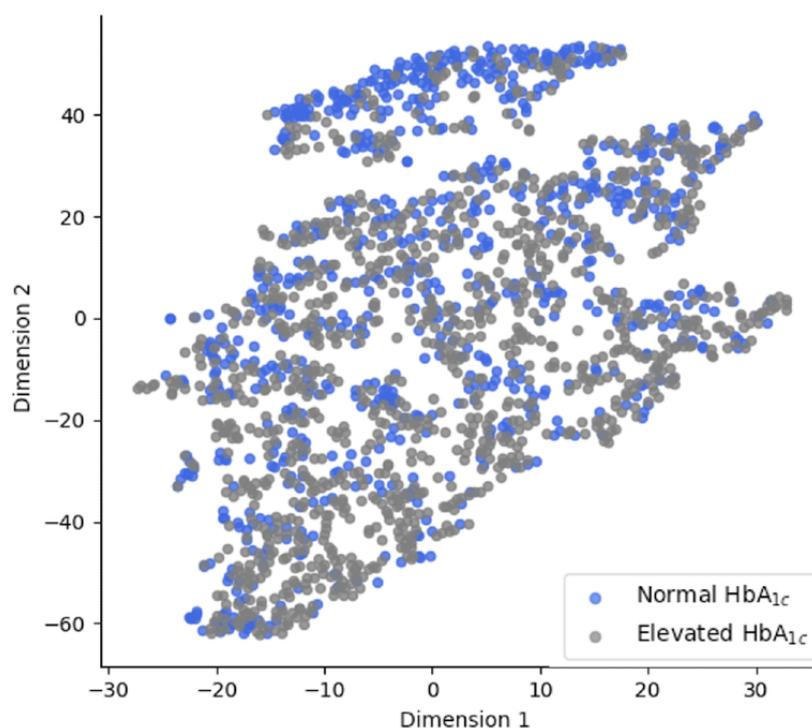
The task of predicting HbA_{1c} elevation risk can be challenging. Figure 10 provides a visualization of the data points for the 2

classes (prediabetic with ≥5.7%; normal with <5.7%) after mapping of the data points (for the test data) into 2 dimensions

with t-distributed stochastic neighbor embedding was performed [50]. The overlap in the data points visualized in the figure demonstrates the challenge of separating the patients with and without elevated levels of HbA_{1c} ($\geq 5.7\%$) in the KAIMRC data

set. We avoided intensive feature engineering techniques in the sampling approach used. However, the approaches adopted were able to achieve promising results with an accuracy of 83.22% for the AUC-ROC using MLP with historical data.

Figure 10. Two-dimensional visualization using t-distributed stochastic neighbor embedding for a randomly selected subset of the data. HbA_{1c}: glycated hemoglobin.



In summary, all models showed promising results for predicting the current HbA_{1c} elevation levels ($\geq 5.7\%$) with EHR data. The results emphasize that the HbA_{1c} predictive models can exhibit more learnability when they are trained with the longitudinal patient data observations typically available from EHR systems.

Discussion

Strengths and Limitations

EHR systems were adopted for the purpose of improving health care outcomes and were not originally intended for research purposes [19]. Patient data stored in EHR systems can be obtained at irregular intervals, as lab instructions are carried out with different frequencies based on the physician's decisions and a patient's visit patterns. It is very common that medical data extracted from EHR systems suffer from problems such as irregularity, incompleteness, and noisy and imbalanced data [13]. These can be challenging obstacles for any technology used for predictive analytics.

In our study, the sampling approach used did not affect the balanced nature of the data set used. As shown in Figure 2, there were 56,185 unique patients present before removal of the records with 1 or more missing values. The number of unique patients with elevated HbA_{1c} levels ($\geq 5.7\%$) before removal of the incomplete records was 27,354, resulting in a retention of

48.68% (27,354/56,185). The number of unique patients with normal HbA_{1c} levels was 28,831, resulting in a retention of 51.32% (28,831/56,185). We would argue that the absence or the presence of the HbA_{1c} readings is not random, as the sample was collected from the population of Saudi Arabia and thus the likelihood of a patient taking an HbA_{1c} test is large because of the prevalence of diabetes in this country [51]. This may affect the reproducibility of this work using different populations from different countries especially those with lower rates of diabetes.

It is hoped that these outcomes will encourage further investigation into the predictability of current HbA_{1c} levels ($\geq 5.7\%$) using more of the readings normally provided in EHR data. For example, other important readings such as FBS and triglycerides have shown clinical correlations with diabetes [52]. In addition, our data set contained only 3 years of patient data, which limits the number of patient visits recorded. Figure 11 shows the number of visits made by patients from 2016 to 2018, while Figure 12 details the number of visits made by patients (after removal of the outliers) over HbA_{1c} levels. Both figures show that the majority of the patients have made relatively few visits: 52% (8713/16818) of the patients made 4 visits or fewer over the 3 years (1.3 visit per year). This also justifies the size of the sliding window ($s=3$) as the optimal input size for the models used. However, we hypothesize that the longitudinal behavior of the features used can be enriched

by including more values obtained over longer periods. Therefore, incorporating more features and their longitudinal behavior over longer periods into the models used in this study would likely improve the prediction performance of our chosen models.

Figure 11. Histogram showing the trend in the number of visits made by patients.

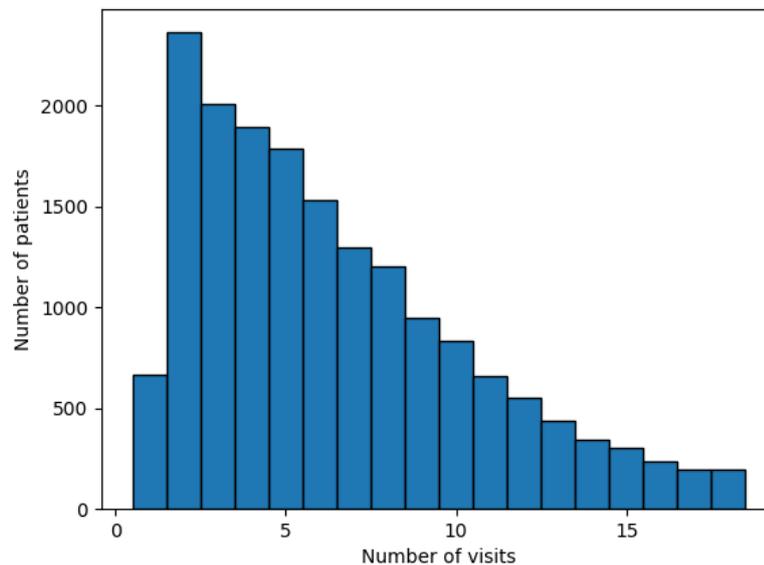
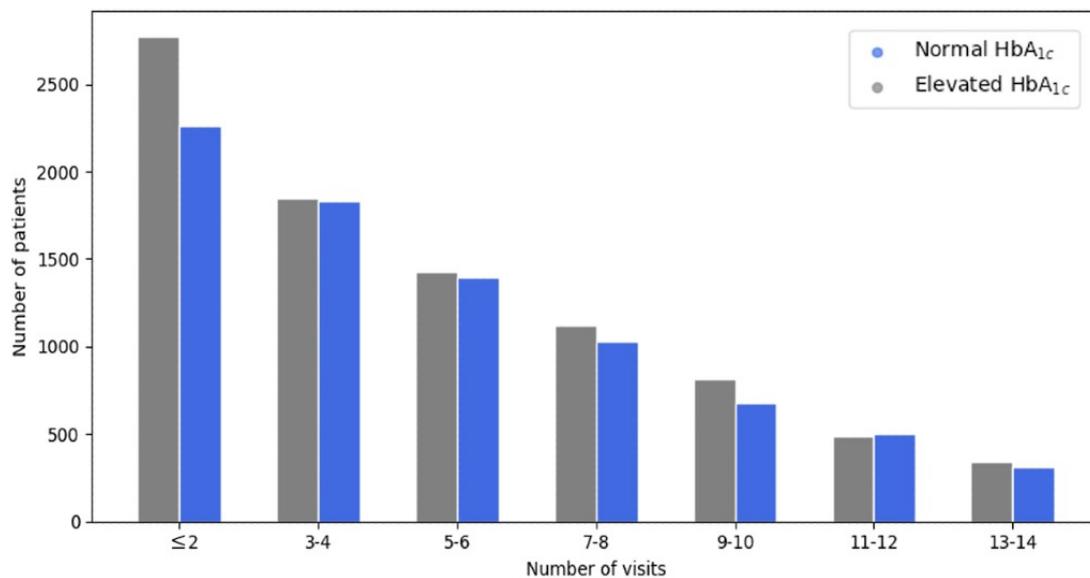


Figure 12. The details for the number of visits made over number of patients. HbA_{1c}: glycated hemoglobin.



Variations in the data or model produce slightly different attribution values. However, due to the critical nature of many health care applications, it is always important to verify that the models make “sensible” predictions. Without the use of SHAP/LIME, this would be hard to verify for any nonlinear model. Although it is possible to see that the models have high performance, we would be unable to verify that a model is not making spurious correlations. Furthermore, through the use of SHAP, we can verify that MLPs trained on the longitudinal data are learning to use the extra information contained in the longitudinal data (as indicated by the higher importance of

eGFR), allowing us to pinpoint the reason these models gain higher performance.

To investigate the effect of temporal dependencies in the data, this study investigated the use of other deep learning models along with the MLP, including long short-term memory (LSTM) and bidirectional LSTM [25,53] for HbA_{1c} prediction. Table 5 reports the results of using these models. The MLP model achieved similar performance to the LSTM and bidirectional LSTM models according to all reported measures. This suggests that directly modeling the temporal dynamics in the data is not very helpful. This could be due to the short lengths of the time series or a too-weak temporal dependency.

Table 5. LSTM and BiLSTM Classifiers performance trained with longitudinal data for current HbA_{1c} levels prediction.

Model	AUC-ROC ^a , % (SD)	Balanced Accuracy, % (SD)	Accuracy, % (SD)	F score, % (SD)	Precision, % (SD)	PR-AUC ^b , % (SD)
LSTM ^c	83.26% (0.91)	74.17% (1.05)	74.59% (1.23)	75.64% (1.50)	74.59% (3.26)	81.88% (0.95)
BiLSTM ^d	83.16% (0.87)	74.21% (1.24)	74.30% (1.15)	75.46% (1.39)	75.19% (2.36)	84.75% (0.75)

^aAUC-ROC: area under the receiver operating characteristic.

^bPR-AUC: precision-recall area under the curve.

^cLSTM: long short-term memory.

^dBiLSTM: bidirectional LSTM.

Generalizing our findings using other data sets is challenging because of the accessibility and privacy restrictions that apply to medical data sets. For this reason, and because of the lack of similar studies that have used machine learning for HbA_{1c} prediction with EHR data, comparing the performance achieved by the models outlined in this study with those developed by other researchers will require the availability of alternative anonymized data sets.

Conclusions

We believe that this study is the first to investigate the performance of machine learning models used with EHR data for predicting current HbA_{1c} elevation risk ($\geq 5.7\%$) for

nondiabetic patients. It is also the first to investigate employing the longitudinal data that are normally stored on EHR systems to enhance the prediction of HbA_{1c} elevation levels. Our findings show that the MLP model achieves better results when a patient's longitudinal data are combined with current visit data, and the use of longitudinal data also affects the relative importance for the predictors used.

As this work formed a continuation of previous work [24], we avoided changing the sampling approach used. However, studying the impact of applying different sampling approaches could be valuable to explore in future work as would the use of a larger data set with more variables and the recording of longitudinal behavior over longer periods.

Acknowledgments

We would like to acknowledge the contribution the KAIMRC for providing the data set under the approved projects: Diabetes Early Warning System (research protocol no. SP14/042), Finding the Common Related Diseases with Diabetes using Data Mining Association Techniques (research protocol no. SP15/064,) and extension project (no. RYD-17-417780-187503) to collect the newest data set. The authors would also like to thank Cievert Ltd and the European Regional Development Fund for sponsoring this work.

Authors' Contributions

ZA was responsible for implementing and building predictive models. ZA, MW, DB, and NAM were responsible for the design of the study and for writing the manuscript. ZA, MW, DB, and NAM were responsible for designing and validating the models. MW and ZA were responsible for analyzing the explainability of the machine learning model. ZA, AA, and RA were responsible for extracting and describing the data set. All authors participated in reviewing the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Lab test and diagnostic codes.

[\[PDF File \(Adobe PDF File\), 93 KB-Multimedia Appendix 1\]](#)

Multimedia Appendix 2

Formulae for the calculated variables.

[\[PDF File \(Adobe PDF File\), 77 KB-Multimedia Appendix 2\]](#)

Multimedia Appendix 3

An example of the padding approach used.

[\[PDF File \(Adobe PDF File\), 169 KB-Multimedia Appendix 3\]](#)

Multimedia Appendix 4

An example of the PAA technique.

[[PDF File \(Adobe PDF File\), 240 KB-Multimedia Appendix 4](#)]

Multimedia Appendix 5

AUC-ROC and PR-AUC curves for the models (with 10 folds) trained with longitudinal data.

[[PDF File \(Adobe PDF File\), 1011 KB-Multimedia Appendix 5](#)]

Multimedia Appendix 6

Variable relative importance charts for the models.

[[PDF File \(Adobe PDF File\), 578 KB-Multimedia Appendix 6](#)]

Multimedia Appendix 7

Multiple logistic regression (MLR) and logistic regression (LR) details.

[[PDF File \(Adobe PDF File\), 157 KB-Multimedia Appendix 7](#)]

References

1. Larsen ML, Hørder M, Mogensen EF. Effect of long-term monitoring of glycosylated hemoglobin levels in insulin-dependent diabetes mellitus. *New England Journal of Medicine* 1990 Oct 11;323(15):1021-1025. [doi: [10.1056/NEJM199010113231503](https://doi.org/10.1056/NEJM199010113231503)] [Medline: [2215560](#)]
2. Pradhan AD, Rifai N, Buring JE, Ridker PM. Hemoglobin A1c predicts diabetes but not cardiovascular disease in nondiabetic women. *The American Journal of Medicine* 2007 Aug;120(8):720-727. [doi: [10.1016/j.amjmed.2007.03.022](https://doi.org/10.1016/j.amjmed.2007.03.022)] [Medline: [PMC2585540](#)]
3. Ackermann RT, Cheng YJ, Williamson DF, Gregg EW. Identifying adults at high risk for diabetes and cardiovascular disease using hemoglobin A1c: National Health and Nutrition Examination Survey 2005-2006. *American Journal of Preventive Medicine* 2011 Jan;40(1):11-17. [doi: [10.1016/j.amepre.2010.09.022](https://doi.org/10.1016/j.amepre.2010.09.022)] [Medline: [21146762](#)]
4. World Health Organization. Use of glycated haemoglobin (HbA1c) in diagnosis of diabetes mellitus: abbreviated report of a WHO consultation. *World Health Organization* 2011:a. [Medline: [26158184](#)]
5. Khaw K, Wareham N, Bingham S, Luben R, Welch A, Day N. Association of hemoglobin A1c with cardiovascular disease and mortality in adults: the European prospective investigation into cancer in Norfolk. *Ann Intern Med* 2004 Sep 21;141(6):413. [doi: [10.7326/0003-4819-141-6-200409210-00006](https://doi.org/10.7326/0003-4819-141-6-200409210-00006)] [Medline: [15381514](#)]
6. American Diabetes Association. Classification and diagnosis of diabetes: standards of medical care in diabetes—2018. *Diabetes Care* 2017 Dec 08;41(Supplement 1):S13-S27. [doi: [10.2337/dc18-s002](https://doi.org/10.2337/dc18-s002)]
7. Coorevits P, Sundgren M, Klein GO, Bahr A, Claerhout B, Daniel C, et al. Electronic health records: new opportunities for clinical research. *Journal of internal medicine* 2013 Oct 18;274(6):547-560. [doi: [10.1111/joim.12119](https://doi.org/10.1111/joim.12119)] [Medline: [23952476](#)]
8. McKinney BA, Reif DM, Ritchie MD, Moore JH. Machine learning for detecting gene-gene interactions: a review. *Appl Bioinformatics* 2006;5(2):77-88 [FREE Full text] [doi: [10.2165/00822942-200605020-00002](https://doi.org/10.2165/00822942-200605020-00002)] [Medline: [16722772](#)]
9. Goldenberg SL, Nir G, Salcudean SE. A new era: artificial intelligence and machine learning in prostate cancer. *Nature Reviews Urology* 2019 May 15;16(7):391-403. [doi: [10.1038/s41585-019-0193-3](https://doi.org/10.1038/s41585-019-0193-3)] [Medline: [31092914](#)]
10. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary use of EHR: data quality issues and informatics opportunities. *Summit Transl Bioinform* 2010:1. [Medline: [21347133](#)]
11. Perveen S, Shahbaz M, Keshavjee K, Guergachi A. Prognostic modeling and prevention of diabetes using machine learning technique. *Scientific reports* 2019 Sep 24;9(1):1. [doi: [10.1038/s41598-019-49563-6](https://doi.org/10.1038/s41598-019-49563-6)] [Medline: [31551457](#)]
12. Esteban S, Rodríguez Tablado M, Peper FE, Mahumud YS, Ricci RI, Kopitowski KS, et al. Development and validation of various phenotyping algorithms for Diabetes Mellitus using data from electronic health records. *Computer Methods and Programs in Biomedicine* 2017 Dec;152:53-70. [doi: [10.1016/j.cmpb.2017.09.009](https://doi.org/10.1016/j.cmpb.2017.09.009)] [Medline: [29054261](#)]
13. Miozzo R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports* 2016 May 17;6(1):1-10. [doi: [10.1038/srep26094](https://doi.org/10.1038/srep26094)] [Medline: [27185194](#)]
14. Hippisley-Cox J, Coupland C, Robson J, Sheikh A, Brindle P. Predicting risk of type 2 diabetes in England and Wales: prospective derivation and validation of QDScore. *BMJ* 2009 Mar 17;338(mar17 2):b880-b880. [doi: [10.1136/bmj.b880](https://doi.org/10.1136/bmj.b880)] [Medline: [19297312](#)]
15. Alhassan Z, McGough A, Alshammari R, Daghestani T, Budgen D, Al MN. Type-2 diabetes mellitus diagnosis from time series clinical data using deep learning models. 2018 Presented at: International Conference on Artificial Neural Networks; 2018 Oct 4-7; Greece. [doi: [10.1007/978-3-030-01424-7_46](https://doi.org/10.1007/978-3-030-01424-7_46)]

16. McCarter RJ, Hempe JM, Chalew SA. Mean blood glucose and biological variation have greater influence on HbA1c levels than glucose instability: an analysis of data from the Diabetes Control and Complications Trial. *Diabetes Care* 2006 Jan 27;29(2):352-355. [doi: [10.2337/diacare.29.02.06.dc05-1594](https://doi.org/10.2337/diacare.29.02.06.dc05-1594)] [Medline: [16443886](https://pubmed.ncbi.nlm.nih.gov/16443886/)]
17. Nathan DM, Kuenen J, Borg R, Zheng H, Schoenfeld D, Heine RJ. Translating the A1C assay into estimated average glucose values. *Diabetes Care* 2008 Jun 07;31(8):1473-1478. [doi: [10.2337/dc08-0545](https://doi.org/10.2337/dc08-0545)] [Medline: [18540046](https://pubmed.ncbi.nlm.nih.gov/18540046/)]
18. Rose E, Ketchell D. Clinical inquiries. Does daily monitoring of blood glucose predict hemoglobin A1c levels? *J Fam Pract* 2003;1. [Medline: [12791231](https://pubmed.ncbi.nlm.nih.gov/12791231/)]
19. Xu S, Schroeder EB, Shetterly S, Goodrich GK, O'Connor PJ, Steiner JF, et al. Accuracy of hemoglobin A1c imputation using fasting plasma glucose in diabetes research using electronic health records data. *Stat., optim. inf. comput* 2014 Jun 01;2(2):93-104. [doi: [10.19139/68](https://doi.org/10.19139/68)]
20. Rauh SP, Heymans MW, Koopman ADM, Nijpels G, Stehouwer CD, Thorand B, et al. Predicting glycosylated hemoglobin levels in the non-diabetic general population: Development and validation of the DIRECT-DETECT prediction model - a DIRECT study. *PLoS ONE* 2017 Feb 10;12(2):e0171816. [doi: [10.1371/journal.pone.0171816](https://doi.org/10.1371/journal.pone.0171816)] [Medline: [28187151](https://pubmed.ncbi.nlm.nih.gov/28187151/)]
21. Wells BJ, Lenoir KM, Diaz-Garelli J, Futrell W, Lockerman E, Pantalone KM, et al. Predicting current glycosylated hemoglobin values in adults: development of an algorithm from the electronic health record. *JMIR Med Inform* 2018 Oct 22;6(4):e10780. [doi: [10.2196/10780](https://doi.org/10.2196/10780)] [Medline: [30348631](https://pubmed.ncbi.nlm.nih.gov/30348631/)]
22. Baan CA, Ruige JB, Stolk RP, Witteman JC, Dekker JM, Heine RJ, et al. Performance of a predictive model to identify undiagnosed diabetes in a health care setting. *Diabetes Care* 1999 Feb 01;22(2):213-219. [doi: [10.2337/diacare.22.2.213](https://doi.org/10.2337/diacare.22.2.213)] [Medline: [10333936](https://pubmed.ncbi.nlm.nih.gov/10333936/)]
23. Griffin SJ, Little PS, Hales CN, Kinmonth AL, Wareham NJ. Diabetes risk score: towards earlier detection of Type 2 diabetes in general practice. *Diabetes/metabolism research and reviews* 2000 May;16(3):164-171. [doi: [10.1002/1520-7560\(200005/06\)16:3<164::aid-dmrr103>3.0.co;2-r](https://doi.org/10.1002/1520-7560(200005/06)16:3<164::aid-dmrr103>3.0.co;2-r)] [Medline: [10867715](https://pubmed.ncbi.nlm.nih.gov/10867715/)]
24. Alhassan Z, Budgen D, Alshammari R, Al Moubayed N. Predicting current glycosylated hemoglobin levels in adults from electronic health records: validation of multiple logistic regression algorithm. *JMIR Med Inform* 2020 Jul 3;8(7):e18963. [doi: [10.2196/18963](https://doi.org/10.2196/18963)] [Medline: [32618575](https://pubmed.ncbi.nlm.nih.gov/32618575/)]
25. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015 May 27;521(7553):436-444. [doi: [10.1038/nature14539](https://doi.org/10.1038/nature14539)]
26. Ahmad M, Eckert C, Teredesai A. Interpretable machine learning in healthcare. 2018 Presented at: Proceedings of the ACM international conference on bioinformatics, computational biology, and health informatics; 2018 Aug 29-Sept 1; Washington DC. [doi: [10.1145/3233547.3233667](https://doi.org/10.1145/3233547.3233667)]
27. Lipton ZC. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *ACM* 2018 Jun;16(3):31-57. [doi: [10.1145/3236386.3241340](https://doi.org/10.1145/3236386.3241340)]
28. Lundberg S, Lee SI. A unified approach to interpreting model predictions. 2017 Presented at: Advances in neural information processing systems; 2017 Dec 4-9; Long Beach.
29. Ribeiro M, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. 2016 Presented at: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; 2016 Aug 13-16; San Francisco. [doi: [10.1145/2939672.2939778](https://doi.org/10.1145/2939672.2939778)]
30. Abdulaziz Al Dawish M, Alwin Robert A, Braham R, Abdallah Al Hayek A, Al Saeed A, Ahmed Ahmed R, et al. Diabetes mellitus in Saudi Arabia: a review of the recent literature. *Current diabetes reviews* 2016 Oct 26;12(4):359-368. [doi: [10.2174/1573399811666150724095130](https://doi.org/10.2174/1573399811666150724095130)] [Medline: [26206092](https://pubmed.ncbi.nlm.nih.gov/26206092/)]
31. Understanding A1C. American Diabetes Association. URL: <https://www.diabetes.org/a1c> [accessed 2020-11-07]
32. Batista GEAPA, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter* 2004 Jun;6(1):20-29. [doi: [10.1145/1007730.1007735](https://doi.org/10.1145/1007730.1007735)]
33. Zhang L, Yang H, Jiang Z. Imbalanced biomedical data classification using self-adaptive multilayer ELM combined with dynamic GAN. *BioMedical Engineering OnLine* volume 2018 Dec 4;17(1):1. [doi: [10.1186/s12938-018-0604-3](https://doi.org/10.1186/s12938-018-0604-3)] [Medline: [30514298](https://pubmed.ncbi.nlm.nih.gov/30514298/)]
34. Rahman MM, Davis DN. Addressing the class imbalance problem in medical datasets. *International Journal of Machine Learning and Computing* 2013;224-228. [doi: [10.7763/ijmlc.2013.v3.307](https://doi.org/10.7763/ijmlc.2013.v3.307)]
35. Longadge R, Dongre S, Malik L. Class imbalance problem in data mining review. *IJCSN* 2013;2(1):1-7.
36. Alhassan Z, Budgen D, Alshammari R, Daghestani T, McGough A, Al MN. Stacked denoising autoencoders for mortality risk prediction using imbalanced clinical data. 2018 Presented at: International Conference on Machine Learning and Applications (ICMLA); 2018 Dec 17; Orlando. [doi: [10.1109/icmla.2018.00087](https://doi.org/10.1109/icmla.2018.00087)]
37. Alqurashi KA, Aljabri KS, Bokhari SA. Prevalence of diabetes mellitus in a Saudi community. *Annals of Saudi Medicine* 2011 Jan;31(1):19-23. [doi: [10.4103/0256-4947.75773](https://doi.org/10.4103/0256-4947.75773)] [Medline: [21245594](https://pubmed.ncbi.nlm.nih.gov/21245594/)]
38. Keogh E, Chakrabarti K, Pazzani M, Mehrotra S. Locally adaptive dimensionality reduction for indexing large time series databases. 2001 Presented at: The 2001 ACM SIGMOD International Conference on Management of Data; 2001 May 21-25; Santa Barbara. [doi: [10.1145/375663.375680](https://doi.org/10.1145/375663.375680)]
39. Zhao J, Papapetrou P, Asker L, Boström H. Learning from heterogeneous temporal data in electronic health records. *Journal of Biomedical Informatics* 2017 Jan;65:105-119. [doi: [10.1016/j.jbi.2016.11.006](https://doi.org/10.1016/j.jbi.2016.11.006)] [Medline: [27919732](https://pubmed.ncbi.nlm.nih.gov/27919732/)]
40. McDonald J. *Handbook of Biological Statistics*. Baltimore, MD: Sparky House Publishing; 2009.

41. Breiman L. Random forests. *Machine learning* 2001;45(1):5-32.
42. Rawlings J, Pantula S, Dickey D. *Applied Regression Analysis*. New York: Springer; 2001:a.
43. Sperandei S. Understanding logistic regression analysis. *Biochemia Medica* 2014;12-18. [doi: [10.11613/bm.2014.003](https://doi.org/10.11613/bm.2014.003)] [Medline: [24627710](https://pubmed.ncbi.nlm.nih.gov/24627710/)]
44. Vapnik V. *The Nature of Statistical Learning Theory*. New York: Springer; 2013.
45. Noble WS. What is a support vector machine? *Nature Biotechnol* 2006 Dec;24(12):1565-1567. [doi: [10.1038/nbt1206-1565](https://doi.org/10.1038/nbt1206-1565)]
46. Gardner M, Dorling S. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Environment* 1998 Aug;32(14-15):2627-2636. [doi: [10.1016/s1352-2310\(97\)00447-0](https://doi.org/10.1016/s1352-2310(97)00447-0)]
47. Goodfellow I, Bengio Y, Courville A, Bengio Y. *Deep Learning*. Cambridge, MA: MIT Press; 2016.
48. Bobadilla J, Ortega F, Hernando A, Gutiérrez A. Recommender systems survey. *Knowledge-Based Systems* 2013 Jul;46:109-132. [doi: [10.1016/j.knosys.2013.03.012](https://doi.org/10.1016/j.knosys.2013.03.012)]
49. Austin PC, Steyerberg EW. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. *BMC medical research methodology* 2012 Jun 20;12(1):109-132. [doi: [10.1186/1471-2288-12-82](https://doi.org/10.1186/1471-2288-12-82)] [Medline: [22716998](https://pubmed.ncbi.nlm.nih.gov/22716998/)]
50. Maaten L, Hinton G. Visualizing data using t-SNE. *Journal of machine learning research*. (Nov) 2008;9:2579-2605.
51. Al-Zahrani J, Aldiab A, Aldossari K, Al-Ghamdi S, Batais M, Javad S. Prevalence of prediabetes, diabetes and its predictors among females in Alkharj, Saudi Arabia: a cross-sectional study. *Annals of Global Health* 2019;85(1):A. [doi: [10.5334/aogh.2467](https://doi.org/10.5334/aogh.2467)] [Medline: [31348623](https://pubmed.ncbi.nlm.nih.gov/31348623/)]
52. Naqvi S, Naveed S, Ali Z, Ahmad S, Khan R, Raj H. Correlation between glycated hemoglobin and triglyceride level in type 2 diabetes mellitus. *Cureus* 2017;9(6):1. [doi: [10.7759/cureus.1347](https://doi.org/10.7759/cureus.1347)] [Medline: [28713663](https://pubmed.ncbi.nlm.nih.gov/28713663/)]
53. Schuster M, Paliwal K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process* 1997;45(11):2673-2681. [doi: [10.1109/78.650093](https://doi.org/10.1109/78.650093)]

Abbreviations

- AUR-ROC:** area under the receiver operating characteristic
eGFR: estimated glomerular filtration rate
EHR: electronic health records
FBS: fasting blood sugar
HbA_{1c}: glycated hemoglobin
KAIMRC: King Abdullah International Medical Research Center
LIME: local interpretable model-agnostic explanations
LR: logistic regression.
LSTM: long short-term memory
MLP: multilayer perceptron
MLR: multiple logistic regression
PAA: piecewise aggregation approximation
PR-AUC: precision-recall area under the curve
RBS: random blood sugar
RF: random forest
SHAP: Shapley Additive Explanations
SVM: support vector machine
T2DM: type-2 diabetes mellitus
WHO: World Health Organization

Edited by C Lovis; submitted 23.10.20; peer-reviewed by S Veeranki, F Agakov, C Doogan; comments to author 13.11.20; revised version received 05.01.21; accepted 22.04.21; published 24.05.21

Please cite as:

Alhassan Z, Watson M, Budgen D, Alshammari R, Alessa A, Al Moubayed N

Improving Current Glycated Hemoglobin Prediction in Adults: Use of Machine Learning Algorithms With Electronic Health Records
JMIR Med Inform 2021;9(5):e25237

URL: <https://medinform.jmir.org/2021/5/e25237>

doi: [10.2196/25237](https://doi.org/10.2196/25237)

PMID:

©Zakhriya Alhassan, Matthew Watson, David Budgen, Riyadh Alshammari, Ali Alessa, Noura Al Moubayed. Originally published in JMIR Medical Informatics (<https://medinform.jmir.org>), 24.05.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on <https://medinform.jmir.org/>, as well as this copyright and license information must be included.