Original Paper

Zakhriya Alhassan, Department of Computer Science, Durham University, Durham, UK and Computer Science Department, College of Computer Science and Engineering, University of Jeddah, Jeddah, Saudi Arabia.
Matthew Watson, Department of Computer Science, Durham University, Durham, UK.
David Budgen, Department of Computer Science, Durham University, Durham, UK.
Riyad Alshammari, College of Public Health and Health Informatics, Health Informatics Department, King Saud bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia and King Abdullah International Medical Research Center, Ministry of the National Guard - Health Affairs, Riyadh, Saudi Arabia.
Ali Alessa, Department of Information Technology Programs, Institute of Public Administration, Riyadh, Kingdom of Saudi Arabia.
Noura Al Moubayed, Department of Computer Science, Durham University, UK.

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

Abstract

Background:
Predicting the risk of glycated hemoglobin (HbA1c) 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 (EHR) data for identifying such patients can ultimately help provide better health outcomes.

Objective:
Our study investigates the performance of predictive models to forecast HbA1c elevation levels by employing several machine learning models. We also investigate utilizing the patient's EHR longitudinal data in the performance of the predictive models. Explainable methods have been employed to interpret the decisions made by the blackbox models.

Methods:
This study employed Multiple Logistic Regression, Random Forest, Support Vector Machine and Logistic Regression models, as well as a deep learning model (Multi-layer perceptron) to classify patients with normal (<5.7%) and elevated (≥5.7%) levels of HbA1c. 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 dataset from Saudi Arabia with 18,844 unique patient records.

Results:
The machine learning models achieved promising results for predicting current HbA1c elevation risk. When employed with longitudinal data, the machine learning models outperformed the Multiple Logistic Regression model employed in the comparative study. The multi-layer perceptron model achieved an accuracy of 83.22% for the AUC-

ROC when used with historical data. All models showed 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 HbA1c levels (≥5.7% or less). Utilizing the patient's longitudinal data improved the performance and affected the relative importance for the predictors used. The models showed results that are consistent with comparable studies.

Keywords: Glycated Hemoglobin HbA1c; Prediction; Machine Learning; Deep Learning; Neural Network; Multi-Layer Perceptron; Electronic Health Records; Time-series Data; Longitudinal Data; Diabetes.

Introduction

The level of glycated hemoglobin (HbA1c) 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), HbA1c provides a long-term measure of a patient's blood glucose levels [3]. The HbA1c 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 Glycated Haemoglobin (HbA1c) in patient blood is considered as a cut-off 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%≤ HbA1c <6.5%) can indicate the future onset of diabetes. Therefore, HbA1c can act as an early predictor for the potential development of Type-2 Diabetes Mellitus (T2DM) [2]. Ackermann et al suggested using the HbA1c test as a measure for identifying those adults who are at a greater risk of developing T2DM in the future [3].

Research has shown that reducing HbA1c levels can significantly reduce the possibility of developing serious complications. Hence, close monitoring of HbA1c levels is recommended for all diabetic patients and also for those with the potential for developing diabetes [5]. It is also suggested that diabetic and non-diabetic patients with raised HbA1c 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, diagnosis, physician's notes, and treatments that are stored in Electronic Health Records (EHR). These are collected on an irregular basis, according to clinical needs, and stored with an associated timestamp.

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 is: “Can HbA1c prediction be improved by using machine learning and utilizing longitudinal data that are normally available in EHR systems?”.

This paper reports an investigation into the performance of machine learning models to predict current HbA1c levels as a binary classification problem using the EHR data. Non-diabetic patients with an HbA1c level of 5.7% or more are considered to have an elevated HbA1c, while those with lower levels than that 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 work is the first to employ machine learning models that use longitudinal data from EHR systems for the purpose of HbA1c elevation risk prediction. This work is also the first to utilize explainable machine learning techniques to explain the classification decisions made by the black box models (SVM and MLP) in predicting HbA1c elevation risk (≥5.7%), in order to better understand the behavior of the model.

Related Work

EHR data has 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 genes 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 in diagnosing diabetes, and discovering its related patterns, using EHR data [11-15]. However, we are not aware of any studies that have explored machine learning models for the prediction of current elevated HbA1c levels using EHR data from a non-diabetic population, as well as the impact of patient longitudinal data on the effectiveness of such predictive machine learning models.

Several studies have investigated the association between HbA1c levels and clinical variables using statistical models [16] [17]. A study by Rose et al [18] discussed the correlation between RBS and HbA1c levels. Stanley et al [19] used a linear regression model for imputation of missing HbA1c data. Their model calculates HbA1c levels for patient records with missing HbA1c values as continuous and categorical values and uses 4 predictors extracted from an EHR system: RBS, FBS, along with age and gender, as predictors to calculate the level of HbA1c for a diabetic population. Simone et al [20] used linear regression models to predict HbA1c levels after 6 years for non-diabetic patients using different populations.

A study by Wells et al [21] in 2018 was the first to focus on predicting current HbA1c elevation levels for non-diabetic patients using an EHR dataset. Multiple Logistic Regression (MLR) was employed to calculate the probability of a patient having an elevated HbA1c level (≥5.7%). The dataset was extracted from an EHR system used in the USA. The authors used 8 independent variables fitted to the model using Restricted Cubic Splines (RCS) 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 HbA1c levels for non-diabetic patients. In addition, the experimental dataset used by Wells et al to train and test their model was imbalanced with 74% of the samples having normal HbA1c levels ($<5.7\%$) and only 26% of the samples having elevated HbA1c levels ($\geq 5.7\%$).

We have performed a differentiated replication of the study by Wells et al [21] using the more balanced KAIMRC dataset [24]. While 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

To study the impact of employing advanced predictive models with EHR data to predict current HbA1c levels, we employed the Multiple Logistic Regression (MLR), Random Forest (RF), Support Vector Machine (SVM) and Logistic Regression (LR) models; as well as a deep learning model, Multi-layer Perceptron (MLP) [25]. The problem was formulated into binary classification problem whereby the target variable, HbA1c level, was encoded with 1 when the level of HbA1c is 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 dataset (detailed in the Dataset subsection). The performance of the models was investigated using current visit data only and also 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.

Using black box machine learning models in healthcare 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 healthcare 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 SHAP (SHapley Additive exPlanations) values [28] and LIME scores [29] techniques have therefore been employed 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 use SHAP values to interpret our black box models, as they can be efficiently calculated, and their use enables a global view of the model to be constructed through the computation of SHAP values from across the whole dataset.

SHAP values are computed using the feature's mean marginal contribution across different coalitions of all features. Shapley 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 calculate are not too greatly affected by the approximation method used, we also compute the LIME [29] scores for the models, across the entire dataset. LIME tries to estimate locally faithful linear explanations (i.e. 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 are able to apply both methods to both of our black box classification models (SVM and MLP).

Dataset

The data used in this study is taken from the King Abdullah International Research Center (KAIMRC) dataset. The data has been collected from King Abdulaziz Medical City located in the central and western regions of Saudi Arabia (KSA), which the World Health Organization (WHO) ranked as the second highest in the Middle East for prevalence of diabetes, and 17th in the world [30]. According to the International Diabetes Federation (IDF), 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 dataset 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 is to identify non-diabetic patients that are at a high risk of HbA1c elevation, all patients previously diagnosed with hyperglycemia were eliminated from the experimental dataset. The remaining cohort formed our experimental dataset, and was categorized by using the American Diabetes Association's (ADA) guidelines [31]. Patients with HbA1c readings of more than 5.7% are considered as being in the pre-diabetic range while those with less than 5.7% are considered to be in the normal range.

Most medical datasets are imbalanced [32] [33] [34]. Such imbalances occur when the proportion of one class of patients in the dataset is greater than its counterpart class [35] [36]. However, unusually, our experimental dataset is not imbalanced. Slightly over half of the patients in our experimental dataset (52.1%) were found to have elevated levels of HbA1c (≥5.7%) while 47.9% of patients had normal HbA1c levels (<5.7%). This can be ascribed to the high incidence of diabetes in the region from which the dataset was collected [37].

A detailed illustration of the patients' class distribution (HbA1c levels) by age groups and gender is shown in Figure 1. This shows that as the age of patients increases, so the proportion of patients who have elevated HbA1c levels is steadily increasing. The dataset also exhibits a balanced gender distribution, with 49.4% of the patients being male and 50.6% female. However, the proportion of male patients with elevated levels of HbA1c (≥5.7%) is greater than for the female patients. Also, female patients with normal levels of HbA1c (<5.7%) made more visits than males. Table 1 shows the profile for the distribution of HbA1c elevation levels organized by gender.

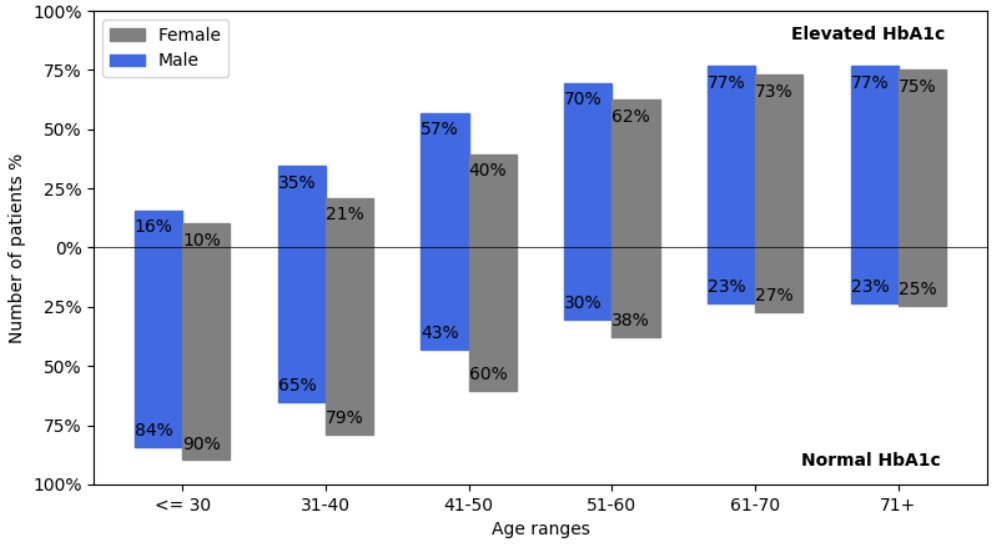

Figure 1. HbA1c Elevation levels distributed over age range and gender in the KAIMRC dataset (before sampling).

Table 1. Profile for the class distribution over gender.

Characteristics		HbA1c <5.7%	HbA1c ≥5.7%
Number of patients (Total:18,844)	Total (%)	9,018 (47.9%)	9,826 (52.1%)
	Male%	41 .73%	56.42%
	Female%	58.27%	43.58%
	Total	79,607	77,993

Number of visits (Total: 157,600)	Male%	39.72%	53.32%
	Female%	60.28%	46.68%

Feature Selection and Data Sampling

Six main variables (features) were extracted from the KAIMRC EHR dataset to be used in this study. These features were selected firstly for their theoretical association with hyperglycemia and secondly for their availability in the KAIMRC dataset, and are: Age, Body Mass Index (BMI), Estimated Glomerular Filtration Rate (eGFR), Random Blood Sugar (RBS), Total cholesterol (CHOL) and non-high-density lipoprotein (non-HDL). For the lab codes of the features used, refer to Table 1 in Multimedia Appendix 1. 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 KAIMRC dataset.

Feature	Unit	HbA1c <5.7%	HbA1c ≥5.7%	*P Value*
Age mean (SD)	Years	43.94 (16.38)	58.92 (15.12)	*<0.001*
BMI mean (SD)	Kg/m^2	29.11 (6.75)	30.90 (6.55)	*<0.001*
eGFR mean (SD)	$mL/min/1.73\ m^2$	100.03 (29.22)	85.81 (28.239)	*<0.001*
RBS mean (SD)	mmol/L	5.45 (1.26)	7.88 (4.19)	*<0.001*
Cholesterol mean (SD)	mmol/L	4.65 (1.07)	4.42 (1.20)	*<0.001*
non-HDL mean (SD)	mmol/L	3.45 (1.01)	3.37 (1.115)	*<0.001*

It is very common in clinical practice that physicians may require that some lab tests and vital signs be recorded frequently. 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 aim at predicting the HbA1c levels (≥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 includes independent patients, to ensure only unseen patients data are used for testing the models. Since we aim at identifying patients with elevated levels of HbA1c from non-diabetic population, patients previously diagnosed with diabetes were excluded. We also excluded non-adult 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 dataset from 114,057 patients with 750,709 visits to 18,844 unique patients with 157,600 visits.

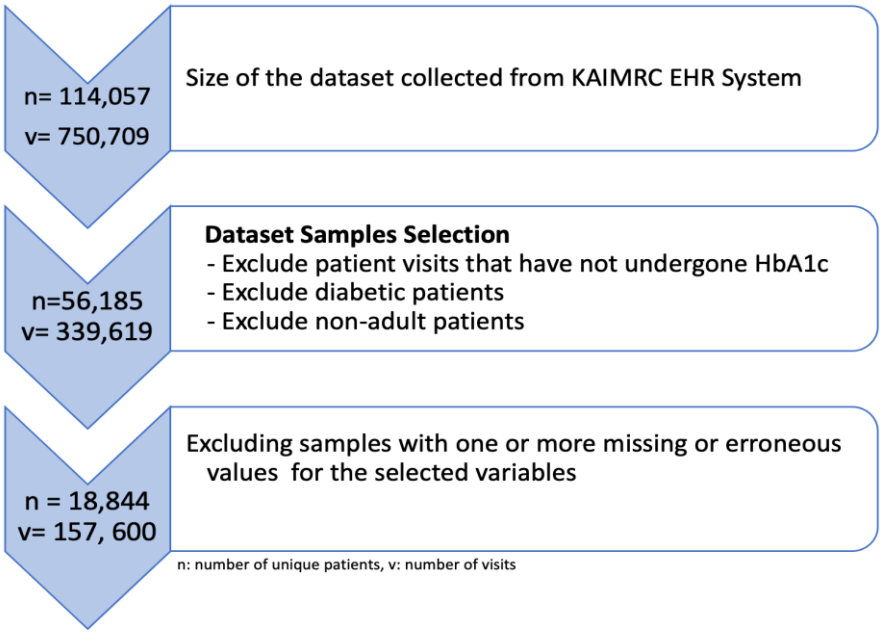

Figure 2. Details of the sampling approach performed on the KAIMRC dataset.

The inputs (input features space) for the models used in this study were continuous values. Values for age, eGFR, RBS and CHOL features were directly available in the KAIMRC dataset. The values for the BMI and non-HDL 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 is described by the selected features, represented as $x_1, x_2, \cdots, x_n$ Those features are formed as episodes based on the time-stamped values available in each visit (v_i).

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

Here x_{ij} is the feature value at a patient visit v_i $(0 < i \geq s, 0 < j \geq n)$; 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 is set to 6 as explained earlier.

If the number of visits (longitudinal time-series visits) for a patient is fewer than s, the input for this patient is 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 is more than s, the Piece-wise Aggregation Approximation (PAA) technique [38] is applied to the data for these visits to take account of 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 is:

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

where $\tilde{x}_i$ represents the approximated value for x and r is the total number of visits for a patient. 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 is then concatenated with the current visit data to form the final input for the deep learning model. Since the MLR, RF, SVM and LR models are not capable of handling the multi-dimensional data (formed as matrices), for these the output of the PAA was re-organized into a single-dimensional input by vectorizing the matrix used in equation (Eq1) as below:

$$Input = [x_{11} \quad x_{12} \quad x_{13} \quad \cdots \quad x_{sn}] \qquad \text{(Eq:3)}$$

The last data pre-processing task before training the predictive models was data scaling. The experimental dataset was scaled using the normalization technique that re-scales 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 Multiple Logistic Regression (MLR) model used by Wells et al, 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 the 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 (RCS) technique with 3-knots. When using the longitudinal input, the variables that caused collinearity were excluded.

Random Forest (RF) 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 dataset. 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 Gini, with a value of 100 for the number of trees parameters.

Logistic regression (LR) is commonly used to solve binary classification problems. It calculates the odds ratio of the variables, and is similar to multiple linear regression but uses a binomial distribution of the dependent variable (i.e. 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 (SVM) 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 dataset into regions, one 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 [45] [44]. The kernel function used in SVM model employed is Radial Base Function (RBF) with a value of 1 for the cost parameter (C).

Multi-layer perceptron (MLP), also known as a feed-forward neural network, is one of the most common deep learning approaches. MLP 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 non-linear functions. MLP uses a backpropagation algorithm to update the weights and biases within the hidden layers to minimize the output error rate [46] [25].

To optimize the MLP model, fine tuning of the structure and hyperparameters has been performed, involving 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 or $Y2$ for elevated HbA1c). A relu activation function was used in the 3 hidden layers and a sigmoid 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.

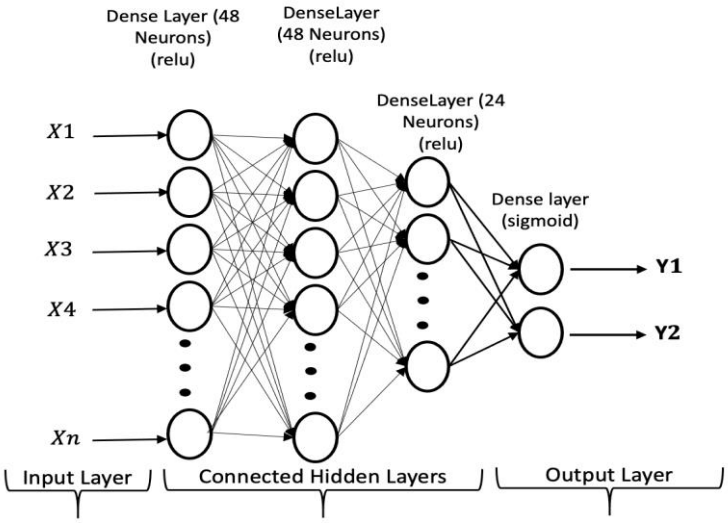

Figure 3. The structure used for multi-layer perceptron (MLP) trained with the longitudinal data.

Evaluation of Model Performance

The models all employed the same data pre-processing, training, and testing techniques. The models were validated using the 10-fold cross-validation technique. The K-fold CV is one of the most commonly approximation approaches used 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), 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, F1-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 compute the SHAP values and LIME scores for all samples in our dataset and then calculate 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 the MLR (statistical model employed by Wells et al), as well as 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 the MLR and RF using the balanced accuracy measure.

Table 3 also shows that all models, including the MLR, achieved better performance using all reported measures when they are 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 HbA1c levels prediction.

Model	With longitudinal data	AUC-ROC, % (SD[f])	Balanced Accuracy, % (SD)	Accuracy, % (SD)	F1, % (SD)	Precision, % (SD)	PR-AUC, % (SD)
MLR[a]	No	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	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[b]	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[c]	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[d]	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)
MLP[e]	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)

[a]MLR: Multiple Logistic Regression.
[b]RF: Random Forest.
[c]SVM: Support Vector Machine.
[d]LR: Logistic Regression.
[e]MLP: Multi-Layer Perceptron.
[f]SD: Standard deviation.

Figure 4 summarizes the 10-folds 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 as well as MLP with and without longitudinal data, as the measures for these models show a small variation between the folds. As shown in Figures 4 and 5, the SD values for MLR with and without longitudinal data are larger

than for the rest of the models. This indicates that the machine learning models used can not only enhance the performance, but also improve the classification confidence for HbA1c prediction.

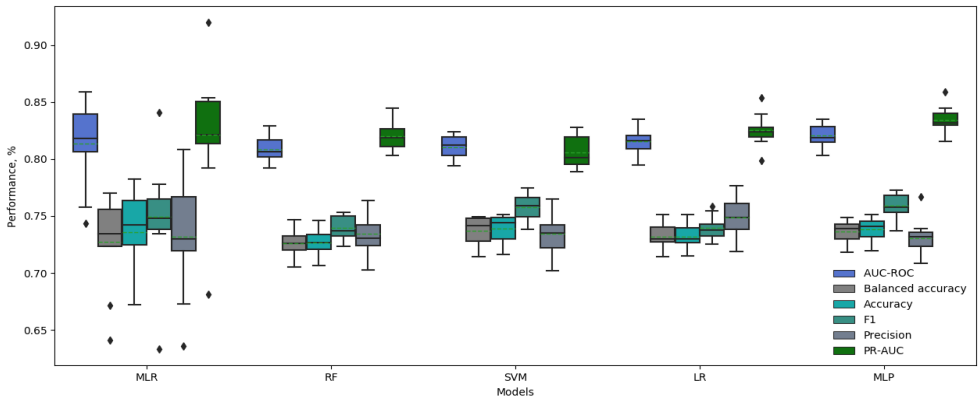

Figure 4. Boxplots showing the detailed 10-folds performance of all models trained without longitudinal data.

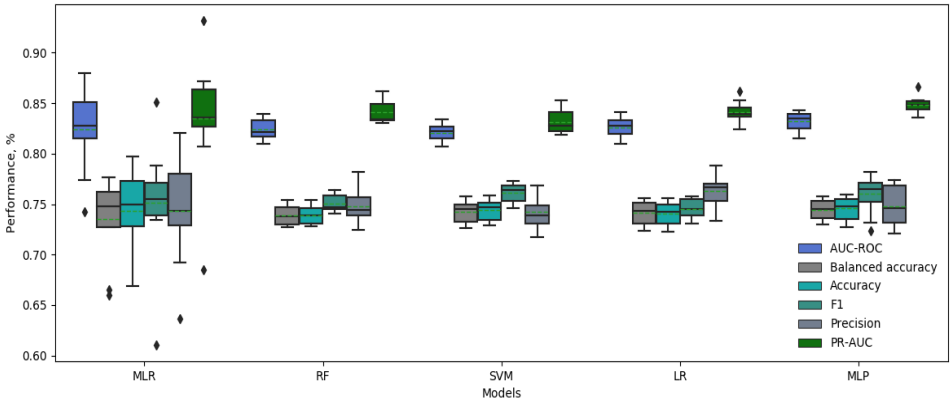

Figure 5. Boxplots showing the detailed 10-folds performance of all models trained with longitudinal data.

Table 4 shows the ranked order of importance of the set of predictors used for training the models. Further detail on the actual importance values for each model is 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 MLP 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 are identical to those obtained using LIME, providing greater confidence in the explainability methods used (see Multimedia Appendix 6).

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

Model	Longitudinal data	1st	2nd	3rd	4th	5th	6th
MLR	No	Age	RBS	BMI	CHOL	Non-HDL	eGFR
RF	No	Age	RBS	BMI	eGFR	CHOL	Non-HDL
	Yes	RBS	Age	CHOL	eGFR	Non-HDL	BMI
LR	No	RBS	Age	Non-HDL	CHOL	BMI	eGFR
	Yes	RBS	Age	Non-HDL	eGFR	CHOL	BMI
SVM (SHAP & LIME)	No	Age	RBS	BMI	Non-HDL	CHOL	eGFR
	Yes	RBS	Age	CHOL	Non-HDL	BMI	eGFR
MLP (SHAP & LIME)	No	RBS	Age	Non-HDL	CHOL	BMI	eGFR
	Yes	RBS	Age	eGFR	CHOL	Non-HDL	BMI

Table 4 and the figures in Multimedia Appendix 6 show that all of the models are heavily and interchangeably reliant on Age and RBS when making classification decisions. The RF and SVM models, when trained with longitudinal data, ranks RBS over Age. Figures 6 and 7 highlight the importance our best performing model, MLP, places upon the features in our dataset using SHAP and LIME, respectively. Both figures show that the RBS contributes the most to the MLP's final prediction, whilst the patient's BMI contributes the least.

For all models trained with longitudinal data, BMI is ranked lower than when the models are trained without longitudinal data. However, the importance value produced for the BMI variable from the models is still not insignificant (see 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 BMI, rendering it less important.

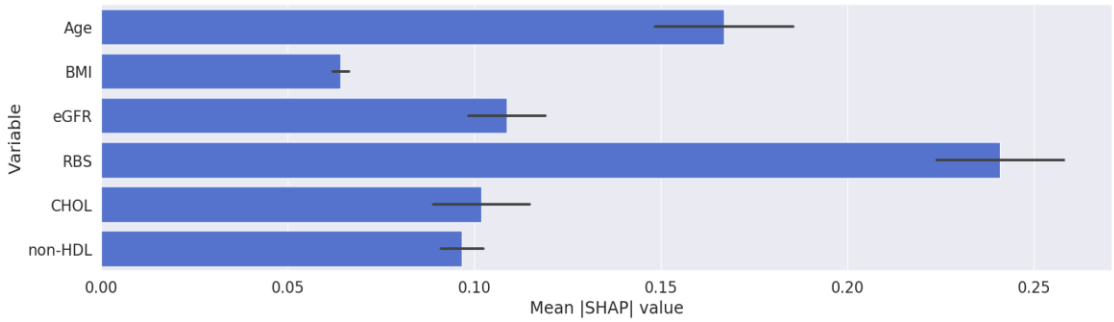

Figure 6. Relative importance of predictors obtained from MLP trained with longitudinal data using SHAP.

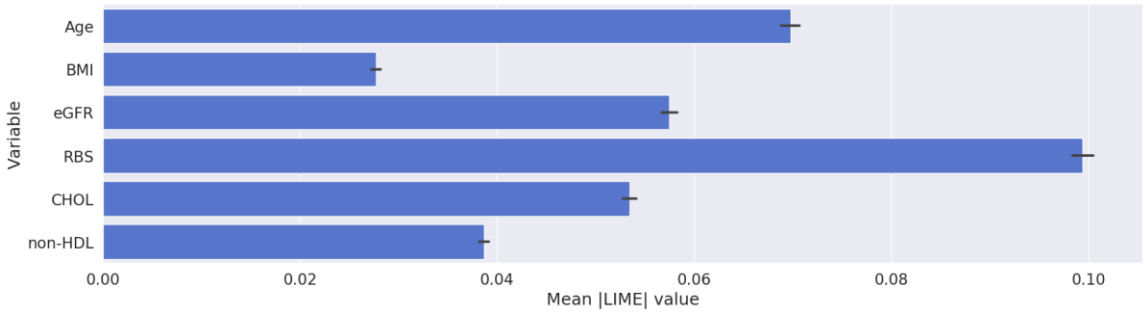

Figure 7. Relative importance of predictors obtained from MLP trained with longitudinal data using LIME.

When using the MLP and LR models trained on the longitudinal data the eGFR variable is ranked higher than CHOL and BMI, in contrast to when these are trained on the current visit only. None of the other models trained with the current visit only, except RF, consider 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 dataset. These figures highlight how different inputs have different SHAP values. The patient in Figure 8 (for whom our model correctly predicts elevated HbA1c levels (≥5.7%)) has a higher RBS value than the patient in Figure 9 (for whom our model correctly predicts normal HbA1c levels (<5.7%)). This explains why our MLP model places much more importance on the RBS value of the patient in Figure 6.

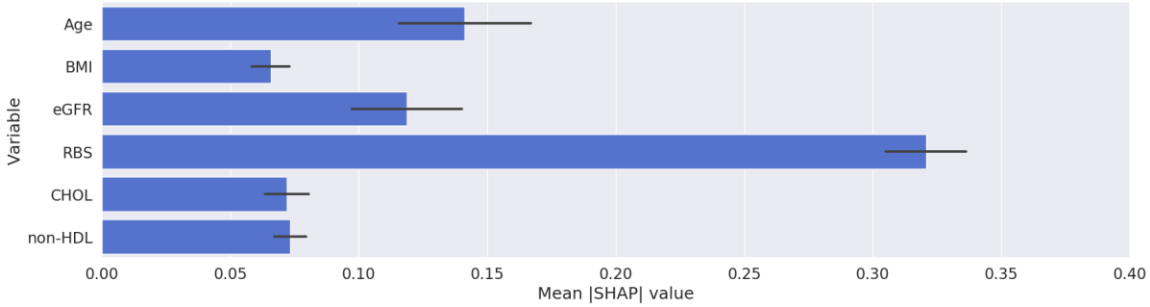

Figure 8. An example shows the SHAP values for a randomly selected sample patient with elevated HbA1c levels (≥5.7%).

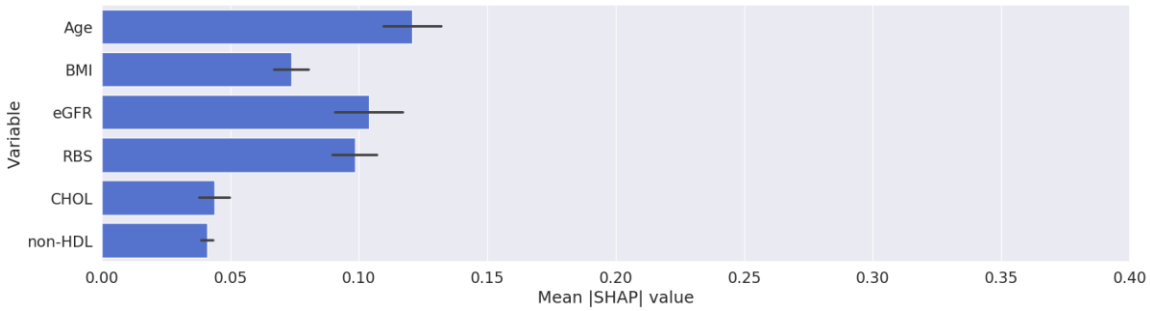

Figure 9. An example shows the SHAP values for a randomly selected sample patient with normal HbA1c levels (<5.7%).

The task of predicting HbA1c elevation risk can be challenging. Figure 10 provides a visualization of the datapoints for the 2 classes (pre-diabetic with ≥5.7%) and (normal with <5.7%) after mapping the datapoints (for the test data) into 2 dimensions using t-SNE [50]. The overlap in the datapoints visualized in the figure demonstrates the challenge of separating the patients with and without elevated levels of HbA1c (≥5.7%) in the KAIMRC dataset. We avoided intensive feature engineering techniques in the sampling approach used. However, the approaches adopted are able to achieve promising results with an accuracy of 83.22% for the AUC-ROC using MLP with historical data.

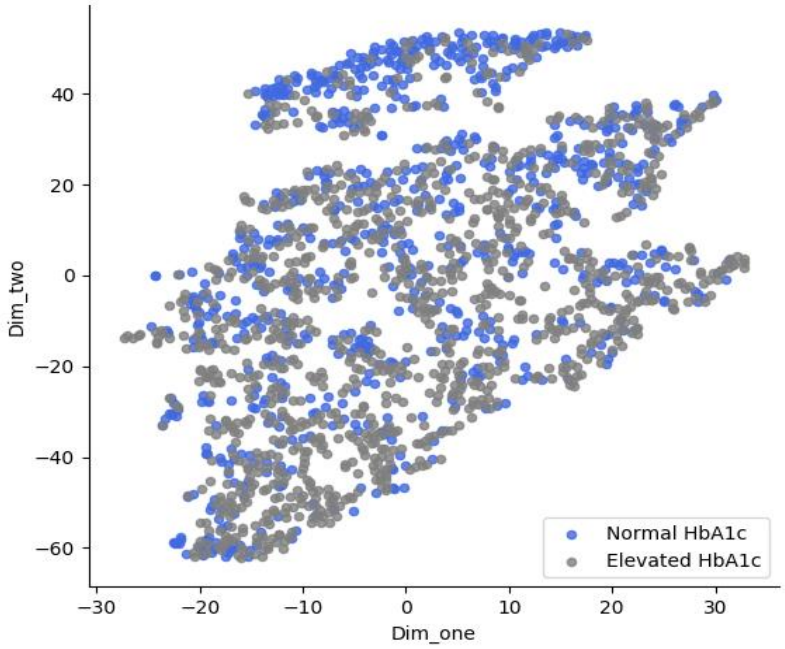

Figure 10. 2-Dimension visualization using t-SNE for a randomly selected subset of the data.

In summary, all models show promising results for predicting the current HbA1c elevation levels (≥5.7%) using EHR data. The results emphasize that the HbA1c predictive models can exhibit more learnability when they are trained with the patient longitudinal observations that are normally available from EHR systems.

Discussion and conclusion

EHR systems were adopted for the purpose of improving healthcare 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.

The sampling approach used did not affect the balanced nature of the dataset used. As shown in Figure 2, there were 56,185 unique patients before removing the records with 1 or more missing values. The number of unique patients with elevated HbA1c levels (≥5.7) before removing the incomplete records was 27,354 with 48.68% (27,354/56,185). The number of unique patients with normal HbA1c levels was 28,831 with 51.32% (28,831/56,185). We would argue that the absence or the presence of the HbA1c readings is not random. Being a sample collected from the population of Saudi Arabia, the likelihood of a patient taking an HbA1c test is large because of the prevalence of diabetes [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 HbA1c levels (≥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 dataset 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. Figure 12 details the number of visits made by patients (after removing the outliers) over HbA1c levels. Both figures show that the majority of the patients have made relatively few visits. 52% (8713/ 16818) of the patients have made 4 visits or fewer during 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 employed. However, we hypothesize that the longitudinal behavior of the features used can be enriched by employing 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 be likely to improve the prediction performance of our chosen models.

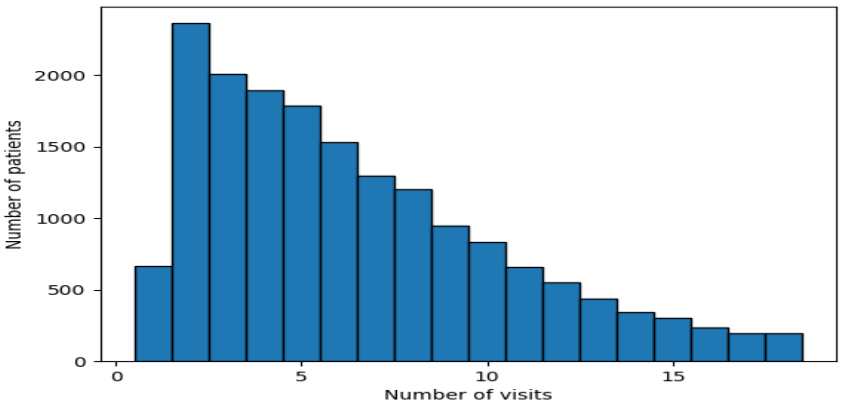

Figure 11. Histogram showing the trend in the number of visits made by patients.

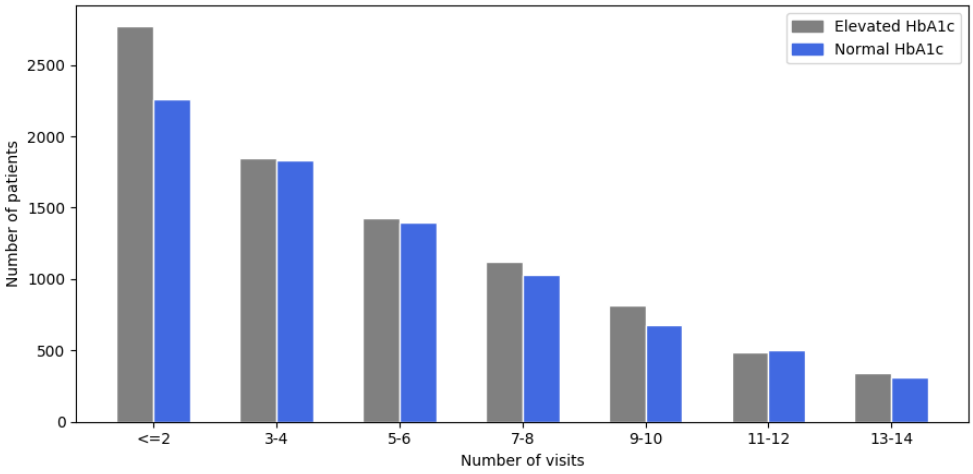

Figure 12. The details for the number of visits made over number of patients across HbA1c levels.

Variations in the data/model produce slightly different attribution values. However, due to the critical nature of many healthcare applications, it is always important to verify that our models make 'sensible' predictions. Without the use of SHAP/LIME, this would be hard to verify for any non-linear 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 has involved investigating the use of other deep learning models along with the MLP, such as Long-Short Term Memory (LSTM) and Bidirectional LSTM [25, 53] for HbA1c prediction.

Table 5 reports the results of using these models. The MLP model achieved similar performance to the LSTM and BiLSTM models using all reported measures. This suggests that directly modelling the temporal dynamics in the data is not very helpful. This could be due to the short lengths of the time series, or to weak temporal dependency.

Table 5. LSTM and BiLSTM Classifiers performance for current HbA1c levels prediction.

Model	With longitudinal data	AUC-ROC, % (SD[f])	Balanced Accuracy, % (SD)	Accuracy, % (SD)	F1, % (SD)	Precision, % (SD)	PR-AUC, % (SD)
LSTM[a]	Yes	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[b]	Yes	83.16% (0.87)	74.21% (1.24)	74.30% (1.15)	75.46% (1.39)	75.19% (2.36)	84.75% (0.75)

[a]LSTM: Long-Short Term Memory.
[b]BiLSTM: Bidirectional LSTM.

Generalizing our findings using other datasets is challenging because of the accessibility and privacy restrictions that apply to medical datasets. For this reason, and because of the lack of similar studies that have employed machine learning for HbA1c prediction using EHR data, comparing the performance achieved by the models outlined in this work with those developed by other researchers will require the availability of alternative anonymized datasets.

Conclusions

We believe that this study is the first to investigate the performance of machine learning models used with EHR data for predicting current HbA1c elevation risk (≥5.7%) for non-diabetic patients. It is also the first to investigate employing the longitudinal data that are normally stored on EHR systems to enhance the prediction of HbA1c 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 dataset with more variables and the recording of longitudinal behavior over longer periods.

Acknowledgements

We would like to acknowledge the contribution by King Abdullah International Research Center (KAIMRC) for providing the dataset under the approved projects: "Diabetes Early Warning System, Research Protocol SP14/042 ", "Finding the Common Related Diseases with Diabetes using Data Mining Association Techniques, Research Protocol SP15/064 " and extension project number RYD-17-417780-187503 to collect the newest dataset. 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, 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 dataset. All authors participated in reviewing the manuscript.

Conflicts of Interest

None declared.

Abbreviations

HbA1c: Glycated Hemoglobin.
KAIMRC: King Abdullah International Medical Research Center.
EHR: Electronic Health Records.
T2DM: Type-2 Diabetes Mellitus.
KSA: Kingdom of Saudi Arabia.
RCS: Restricted Cubic Splines.
SHAP: SHapley Additive exPlanations.
BMI: Body Mass Index.
eGFR: estimated Glomerular Filtration Rate.
CHOL: Total Cholesterol.
Non-HDL: Non-High Density Lipoprotien.
RBS: Random Blood Sugar.
FBS: Fasting Blood Sugar.
PAA: Piece-wise Aggregation Approximation.
AUR-ROC: Area Under the Receiver Operating Characteristic.
PR-AUC: Precision-Recall Area Under the Curve.
MLR: Multiple Logistic Regression.
RF: Random Forest.
SVM: Support Vector Machine.
LR: Logistic Regression.
MLP: Multi-Layer Perceptron.
SD: Standard Deviation.
LSTM: Long-Short Term Memory.
BiLSTM: Bidirectional LSTM.

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;323(15):1021-5.

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;120(8):720-7.
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;40(1):11-7.
4. Organization WH. Use of glycated haemoglobin (HbA1c) in diagnosis of diabetes mellitus: abbreviated report of a WHO consultation. World Health Organization, 2011.
5. Khaw K-T, 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. Annals of internal medicine. 2004;141(6):413-20.
6. Association AD. 2. Classification and diagnosis of diabetes: standards of medical care in diabetes—2018. Diabetes care. 2018;41(Supplement 1):S13-S27.
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;274(6):547-60.
8. McKinney BA, Reif DM, Ritchie MD, Moore JH. Machine learning for detecting gene-gene interactions. Applied bioinformatics. 2006;5(2):77-88.
9. Goldenberg SL, Nir G, Salcudean SE. A new era: artificial intelligence and machine learning in prostate cancer. Nature Reviews Urology. 2019;16(7):391-403.
10. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary use of EHR: data quality issues and informatics opportunities. Summit on Translational Bioinformatics. 2010;2010:1.
11. Perveen S, Shahbaz M, Keshavjee K, Guergachi A. Prognostic modeling and prevention of diabetes using machine learning technique. Scientific reports. 2019;9(1):1-9.
12. Esteban S, Tablado MR, 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;152:53-70.
13. Miotto 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;6(1):1-10.
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;338:b880.
15. Alhassan Z, McGough AS, Alshammari R, Daghstani T, Budgen D, Al Moubayed N, editors. Type-2 diabetes mellitus diagnosis from time series clinical data using deep learning models. International Conference on Artificial Neural Networks; 2018: Springer.
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;29(2):352-5.
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;31(8):1473-8.
18. Rose E, Ketchell DS. Does daily monitoring of blood glucose predict hemoglobin A1c levels? Clinical Inquiries, 2003 (MU). 2003.
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. Statistics, Optimization & Information Computing. 2014;2(2):93-104.
20. Rauh SP, Heymans MW, Koopman AD, Nijpels G, Stehouwer CD, Thorand B, et al. Predicting glycated hemoglobin levels in the non-diabetic general population: Development and validation of the DIRECT-DETECT prediction model-a DIRECT study. PLoS One. 2017;12(2):e0171816.
21. Wells BJ, Lenoir KM, Diaz-Garelli J-F, Futrell W, Lockerman E, Pantalone KM, et al. Predicting Current Glycated Hemoglobin Values in Adults: Development of an Algorithm From the Electronic Health Record. JMIR medical informatics. 2018;6(4):e10780.
22. Baan CA, Ruige JB, Stolk RP, Witteman J, Dekker JM, Heine RJ, et al. Performance of a predictive model to identify undiagnosed diabetes in a health care setting. Diabetes care. 1999;22(2):213-9.
23. Griffin S, Little P, Hales C, Kinmonth A, Wareham N. Diabetes risk score: towards earlier detection of type 2 diabetes in general practice. Diabetes/metabolism research and reviews. 2000;16(3):164-71.
24. Alhassan Z, Budgen D, Alshammari R, Al Moubayed N. Predicting Current Glycated Hemoglobin Levels in Adults From Electronic Health Records: Validation of Multiple Logistic Regression Algorithm. JMIR Medical Informatics. 2020;8(7):e18963.
25. LeCun Y, Bengio Y, Hinton G. Deep learning. nature. 2015;521(7553):436-44.
26. Ahmad MA, Eckert C, Teredesai A, editors. Interpretable machine learning in healthcare. Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics; 2018.
27. Lipton ZC. The mythos of model interpretability. Queue. 2018;16(3):31-57.
28. Lundberg SM, Lee S-I, editors. A unified approach to interpreting model predictions. Advances in neural information processing systems; 2017.
29. Ribeiro MT, Singh S, Guestrin C, editors. " Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; 2016.
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;12(4):359-68.
31. Association AD. [cited 2020 11/7/2020]; Available from: https://www.diabetes.org/a1c.

32. Batista GE, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD explorations newsletter. 2004;6(1):20-9.
33. Zhang L, Yang H, Jiang Z. Imbalanced biomedical data classification using self-adaptive multilayer ELM combined with dynamic GAN. Biomedical engineering online. 2018;17(1):181.
34. Rahman MM, Davis DN. Addressing the class imbalance problem in medical datasets. International Journal of Machine Learning and Computing. 2013;3(2):224.
35. Longadge R, Dongre S. Class imbalance problem in data mining review. arXiv preprint arXiv:13051707. 2013.
36. Alhassan Z, Budgen D, Alshammari R, Daghstani T, McGough AS, Al Moubayed N, editors. Stacked Denoising Autoencoders for Mortality Risk Prediction Using Imbalanced Clinical Data. 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA); 2018: IEEE.
37. Alqurashi KA, Aljabri KS, Bokhari SA. Prevalence of diabetes mellitus in a Saudi community. Annals of Saudi medicine. 2011;31(1):19-23.
38. Keogh E, Chakrabarti K, Pazzani M, Mehrotra S, editors. Locally adaptive dimensionality reduction for indexing large time series databases. Proceedings of the 2001 ACM SIGMOD international conference on Management of data; 2001.
39. Zhao J, Papapetrou P, Asker L, Boström H. Learning from heterogeneous temporal data in electronic health records. Journal of biomedical informatics. 2017;65:105-19.
40. McDonald JH. Handbook of biological statistics: sparky house publishing Baltimore, MD; 2009.
41. Breiman L. Random forests. Machine learning. 2001;45(1):5-32.
42. Rawlings JO, Pantula SG, Dickey DA. Applied regression analysis: a research tool: Springer Science & Business Media; 2001. ISBN: 0387984542.
43. Sperandei S. Understanding logistic regression analysis. Biochemia medica: Biochemia medica. 2014;24(1):12-8.
44. Vapnik V. The nature of statistical learning theory: Springer science & business media; 2013. ISBN: 1475732643.
45. Noble WS. What is a support vector machine? Nature biotechnology. 2006;24(12):1565-7.
46. Gardner MW, Dorling S. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmospheric environment. 1998;32(14-15):2627-36.
47. Goodfellow I, Bengio Y, Courville A, Bengio Y. Deep learning: MIT press Cambridge; 2016.
48. Bobadilla J, Ortega F, Hernando A, Gutiérrez A. Recommender systems survey. Knowledge-based systems. 2013;46:109-32.
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;12(1):82.
50. Maaten Lvd, Hinton G. Visualizing data using t-SNE. Journal of machine learning research. 2008;9(Nov):2579-605.

51. Al-Zahrani JM, Aldiab A, Aldossari KK, Al-Ghamdi S, Batais MA, Javad S, et al. Prevalence of Prediabetes, Diabetes and Its Predictors among Females in Alkharj, Saudi Arabia: A Cross-Sectional Study. Annals of Global Health. 2019;85(1).
52. Naqvi S, Naveed S, Ali Z, Ahmad SM, Khan RA, Raj H, et al. Correlation between glycated hemoglobin and triglyceride level in type 2 diabetes mellitus. Cureus. 2017;9(6).
53. Schuster M, Paliwal KK. Bidirectional recurrent neural networks. IEEE transactions on Signal Processing. 1997;45(11):2673-81.