


Integrated statistical and machine learning analysis provides insight into key influencing symptoms for distinguishing early-onset type 2 diabetes

David A. Wood 

DWA Energy Limited, Lincoln, UK

Correspondence

David A. Wood, DWA Energy Limited,
Lincoln LN5 9JP, UK.
Email: dw@dwasolutions.com

Edited by Yi Cui

Funding information

None

Abstract

Background: Being able to predict with confidence the early onset of type 2 diabetes from a suite of signs and symptoms (features) displayed by potential sufferers is desirable to commence treatment promptly. Late or inconclusive diagnosis can result in more serious health consequences for sufferers and higher costs for health care services in the long run.

Methods: A novel integrated methodology is proposed involving correlation, statistical analysis, machine learning, multi-*K*-fold cross-validation, and confusion matrices to provide a reliable classification of diabetes-positive and -negative individuals from a substantial suite of features. The method also identifies the relative influence of each feature on the diabetes diagnosis and highlights the most important ones. Ten statistical and machine learning methods are utilized to conduct the analysis.

Results: A published data set involving 520 individuals (Sylthet Diabetes Hospital, Bangladesh) is modeled revealing that a support vector classifier generates the most accurate early-onset type 2 diabetes status predictions with just 11 misclassifications (2.1% error). Polydipsia and polyuria are among the most influential features, whereas obesity and age are assigned low weights by the prediction models.

Conclusion: The proposed methodology can rapidly predict early-onset type 2 diabetes with high confidence while providing valuable insight into the key influential features involved in such predictions.

KEYWORDS

error analysis, key feature influences, multi-*K*-fold cross-validation, symptom importance, type 2 diabetes screening

Highlights

- New integrated method combines statistical analysis and machine learning.
- Multi-*K*-fold validation reveals high-performing machine learning model setups.
- Statistical analysis of a suite of signs and symptoms identifies prediction challenges.
- Relative feature influences on prediction models contrast with correlations.
- Annotated confusion matrices provide detailed insight into misclassifications.
- Support vector classifier predicts early-onset type 2 diabetes with 2.1% errors.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Authors. *Chronic Diseases and Translational Medicine* published by John Wiley & Sons, Ltd on behalf of Chinese Medical Association.

1 | INTRODUCTION

Type 2 diabetes is a metabolic disease condition that renders individuals unable to produce sufficient insulin via pancreatic processes, and/or causes their bodies to become in some way resistant to insulin, rendering it less effective in regulating the body's energy supply.¹ Various blood-test criteria are used to diagnose diabetes. Glycated hemoglobin (HbA1c) $\geq 6.5\%$ (48 mmol/mol) and fasting (8 h without calorie intake) plasma glucose (FPG) levels of ≥ 126 mg/dl (7.0 mmol/L) are commonly used diagnostics. In some cases, an oral glucose tolerance test (OGTT) may also be conducted but this can be quite time-consuming. It involves taking initial blood samples, then having the patient drink a specified volume of glucose-containing liquid (~75 g). Additional blood samples are then taken from the patient at intervals up to about 3 h. The blood samples are tested to determine how quickly the glucose levels return to their initial levels. The HbA1c blood test is considered to be more specific for diagnosing early-onset diabetes and a slightly lower threshold has been recently recommended for that purpose of 6.03%.²

Type 2 diabetes is a global problem that is on the rise. About 537 million adults (aged 20–79 years) were living with diabetes in 2021, with that number expected to increase to about 643 million by 2030.³ Alarming, about one-in-two people living with diabetes remain undiagnosed and therefore are not being treated or taking prudent steps to mitigate its long-term health impacts. This is because early-onset diabetes conditions can extend over many years and cause life-threatening complications before detection in some cases.⁴ Early detection, treatments, and lifestyle changes have been shown to reduce risks of cardiovascular morbidity and mortality.⁵ Many blame a deteriorating diet and/or decrease in physical activity associated with modern, sedentary, urban lifestyles for the rapid increase in cases of type 2 diabetes, particularly in low- and middle-income countries.⁶

Blood testing (FPG, OGTT, HbA1c) offers the most definitive and reliable way to establish whether an individual is suffering from type 2 diabetes, or not. However, screening all individuals showing certain symptoms of type 2 diabetes using blood testing methods is both time-consuming and costly. It also requires substantial investment in testing infrastructure and analytical staff distributed at a local level. In most high population countries, especially developing countries with limited healthcare resources and budgets distributed unevenly across the nation, extensive routine blood test screening of large sectors of the population for type 2 diabetes is neither logistically nor financially feasible. Therefore, accurate techniques that exploit machine learning (ML) to consider assessments of groups of symptoms potentially related to type 2 diabetes displayed by individuals offer a quick, meaningful, and relatively cheap method to identify likely early-onset type 2 diabetes sufferers. A ML

assessment that an individual is likely suffering from type 2 diabetes enables clinicians to promptly recommend dietary and/or other lifestyle changes. Such machine-learning assessments can also be used to determine/filter which individuals require urgent blood-test assessment for early-onset type 2 diabetes, and which do not.

Attempts to diagnose the early onset of diabetes tend to use groups of signs and symptoms commonly associated with the disease. In particular, excessive thirst and/or dry mouth conditions, frequent urination, lack of energy/tiredness, slow healing wounds, recurrent skin infections, blurred vision, and tingling/partial numbness in hands and/or feet are symptoms that tend to be associated with early-onset type 2 diabetes.³ There is, therefore, a substantial ongoing research effort to develop reliable early-onset type 2 diabetes prediction models using suits of signs and symptoms exhibited by potential sufferers as predictors.⁷ This effort involves applying many distinct statistical,⁸ multivariate logistic regression (LGR),⁹ and machine and deep learning methods to datasets recording multiple signs and symptoms displayed by many hundreds of type 2 diabetes-positive and -negative individuals.^{10,11} Those datasets are used to train and validate models for deployment in the prediction of early-onset type 2 diabetes, in individuals previously unseen by the models, based on the signs and symptoms they display, rather than clinical tests.

In this study, a published data set involving 520 individuals is evaluated.^{12,13} Each individual has 16 criteria recorded relating to the signs and symptoms they display, and negative or positive status with respect to type 2 diabetes. Some of the data set individuals have recently tested positive for type 2 diabetes, others display some signs or symptoms commonly associated with diabetes but remain diabetes-negative. This study applies a novel integrated methodology combining correlation, statistical analysis, and ML to characterize the data set in detail. The primary objectives of the study are to: (1) Demonstrate the early-onset type 2 diabetes classification performances of three statistical and seven distinct ML models, and identify the model providing the most reliable binary (diabetes-negative/-positive) predictions with minimum misclassifications; (2) apply multi- K -fold cross-validation analysis and annotated confusion matrices to establish optimal configurations of the prediction models; and (3) identify the key type 2 diabetes signs and symptoms and their relative influences on the solutions derived by the high-performing classification models for this data set.

2 | MATERIALS AND METHODS

2.1 | Sylhet Diabetic Hospital data set

The data set evaluated was collected at the Sylhet Diabetic Hospital (Sylhet, Bangladesh) and previously

assessed using four ML models.¹² It has subsequently been placed in the public domain as a University of California, School of Information and Computer Science data set.¹³ The data set is referred to here as the Sylhet-520 data set.

The Sylhet-520 data set involves 16 criteria (independent variables) recorded for 520 individuals (328 males and 192 females), combined with a binary classification (dependent variable) of those patients based on blood tests into diabetes-positive and -negative categories (Table 1). The individuals included are those who were recently diagnosed with type 2 diabetes plus others experiencing some of the signs or symptoms but, at the time of the assessment, they tested negative for type 2 diabetes.

The individuals included in the data set can be further categorized according to gender, age, and type 2 diabetes blood test results (Table 2).

Fifteen of the independent variables (variables #2 to #16) are assessed in binary terms: variables #3 to #16 in “yes” or “no” answers; variable #2 distinguishes gender as male or female. Independent variable #1 is distinct in that it involves five age categories (Table 1). Variables #3 to #16 represent a spectrum of signs and symptoms (criteria) potentially associated with early-onset type 2 diabetes.

The 16 criteria applied to the Sylhet-520 data set individuals were selected by the clinical staff at the Sylhet Diabetic Hospital, based on multiple years of experience in dealing with many cases of type 2 diabetes and early-onset type 2 diabetes. These criteria clearly

TABLE 1 Early-onset diabetes signs and symptoms displayed by individuals (criteria normalized to scale -1 to +1)

| Criteria (I) # | Criteria description | ML variable type | “Yes” answer normalized value assigned | “No” answer normalized value assigned |
|--------------------|-----------------------------------|------------------|--|---------------------------------------|
| 1 | Age 20–35 years | I | Yes (-1.0) | N/A |
| 1 | Age 36–45 years | I | Yes (-0.5) | N/A |
| 1 | Age 46–55 years | I | Yes (0.0) | N/A |
| 1 | Age 56–65 years | I | Yes (+0.5) | N/A |
| 1 | Age >65 years | I | Yes (+1.0) | N/A |
| 2 | Gender | I | Male (-1.0) | N/A |
| 2 | Gender | I | Female (+1.0) | N/A |
| 3 | Polyuria | I | Yes (-1.0) | No (+1.0) |
| 4 | Polydipsia | I | Yes (-1.0) | No (+1.0) |
| 5 | Sudden weight loss | I | Yes (-1.0) | No (+1.0) |
| 6 | Weakness | I | Yes (-1.0) | No (+1.0) |
| 7 | Polyphagia | I | Yes (-1.0) | No (+1.0) |
| 8 | Genital thrush | I | Yes (-1.0) | No (+1.0) |
| 9 | Visual blurring | I | Yes (-1.0) | No (+1.0) |
| 10 | Itching | I | Yes (-1.0) | No (+1.0) |
| 11 | Irritability | I | Yes (-1.0) | No (+1.0) |
| 12 | Delayed healing | I | Yes (-1.0) | No (+1.0) |
| 13 | Partial paresis | I | Yes (-1.0) | No (+1.0) |
| 14 | Muscle stiffness | I | Yes (-1.0) | No (+1.0) |
| 15 | Alopecia | I | Yes (-1.0) | No (+1.0) |
| 16 | Obesity | I | Yes (-1.0) | No (+1.0) |
| Dependent variable | Class (type 2 diabetes- positive) | D | Yes (-1.0) | N/A |
| | Class (type 2 diabetes- negative) | D | Yes (+1.0) | N/A |

Note: “I” refers to independent variables; “D” refers to a binary dependent variable. Signs and symptoms check for early-onset type-2 diabetes conducted as a questionnaire supervised by clinicians for 520 individuals displaying newly diabetic or could-be diabetic characteristics at the Sylhet Diabetic Hospital in Bangladesh (320 positive and 200 negative for type 2 diabetes). The data set is in a public archive,¹³ and its details are provided by Islam et al.¹²

have specific relevance to the local population served by the Sylhet Diabetic Hospital. Therefore the exact set of criteria pertinent to other populations at different locations might vary slightly based on what are considered to be locally relevant factors. However, overall the 16 criteria selected for this data set of individuals are considered to be generally relevant based on the criteria discussed in other studies.^{3,6-11}

The criteria and the dependent variable assessments for each individual are converted, for this study, into normalized numerical ranges from -1.0 to $+1.0$. For the “yes/no” criteria assigned to an individual; a “yes” classification is assigned -1.0 and a “no” classification $+1.0$. For variable #2, “male” designation is assigned -1.0 and a “female” designation is assigned $+1.0$. For variable #1, the five age categories are respectively assigned -1.0 , -0.5 , 0 , $+0.5$, and $+1.0$. This normalized range of -1.0 to $+1.0$ for each independent variable avoids scaling biases that impact some classification models.

TABLE 2 Individuals categorized according to gender, age, and type 2 diabetes blood test result

| Age | Female | | Male | |
|-------------|--------|----------|-------|----------|
| | Total | Positive | Total | Positive |
| 20-35 years | 42 | 36 | 51 | 15 |
| 36-45 years | 46 | 45 | 92 | 32 |
| 46-55 years | 66 | 58 | 83 | 43 |
| 55-65 years | 23 | 19 | 66 | 37 |
| >65 years | 15 | 15 | 36 | 20 |

TABLE 3 Regression and machine learning models applied to predict diabetes onset risks based on influencing variables

| Model | Code | Type | Originator(s) | Examples of ML model applied in diabetes studies |
|---------------------------------|------|--|---|--|
| Logistical Regression | LGR | Probabilistic Classifier | Berkson (1944) ¹⁴ | Tabaei and Herman (2002) ¹⁵ Rajendra and Latifi (2021) ¹⁶ |
| Naïve Bayes Classifier | NBC | Probabilistic Classifier | Thomas Bayes' Theorem was proposed in 1763, Hand and Yu (2001) ¹⁷ | Theresa and Evangeline (2021) ¹⁸ |
| Quadratic Discriminant Analysis | QDA | Statistical Classifier with Quadratic Decision Surface | Fischer (1936; linear version) ¹⁹ Tharwat, 2016 (QDA vs. LDA) ²⁰ | Maniruzzaman et al. (2018) ²¹ |
| Adaptive Boosting | ADA | Boosted Tree ensemble | Freund and Schapire (1997) ²² | Vijayan and Anjali (2015) ²³ |
| Decision Tree | DT | Single tree | Quinlan (1986) ²⁴ | Ramezankhani et al. (2016) ²⁵ |
| K-Nearest Neighbor | KNN | Data Matching | Fix and Hodges (1951) ²⁶ | Sarkar et al. (2019) ²⁷ |
| Multi-Layer Perceptron | MLP | Artificial Neural Network | Rosenblatt (1958) ²⁸ | Bani-Salameh et al. (2021) ²⁹ |
| Random Forest | RF | Tree ensemble | Ho (1998) ³⁰ | Wang et al. (2021) ³¹ |
| Support Vector Classifier | SVC | Hyperplane Fit | Cortes and Vapnik (1995) ³² | Abbas et al. (2019) ³³ |
| Extreme Gradient Boosting | XGB | Boosted Tree ensemble | Chen and Guestrin (2016) ³⁴ | Wang et al. (2020) ³⁵ |

Note: Prediction models evaluated for type-2 diabetes status prediction using 520 data records from the Sylhet Diabetic Hospital published data set.¹³

For the dependent variable, diabetes-positive individuals are assigned a value of $+1.0$, whereas diabetes-negative individuals are assigned a value of -1.0 . This means that measured in numerical terms each misclassification error generated by a classification model will score a value of “2” in absolute terms (e.g., $+1$ less -1). This makes it possible to compute a metric determining the number of errors generated by each model solution (Section 2.3 and Supporting Information: Figure S1)

2.2 | Statistical and ML methods

Ten statistical and ML models are configured and applied to predict the binary classification of the dependent variable (“Class” in Table 1; diabetes-positive or -negative) from 16 potentially influential criteria. This suite of well-established ML methods (Table 3) is executed in Python code with the aid of published algorithms.³⁶ The mathematics underpinning these models are described extensively in the literature building on the work of their original developers. All of these models have been applied in recent years as part of diabetes prediction studies using a selection of potential influential variables with some examples cited in Table 3.

LGR is a statistical model widely applied to binary classification tasks based on values relating to independent variables. It involves linear mathematical relationships.¹⁴ The *Naïve Bayes classifier* (NBC) applies the Bayes Theorem to derive probabilistic classifications by making the simplistic assumption that each independent variable influences the dependent variable in an

entirely independent way.¹⁷ *Quadratic discriminant analysis* (QDA) is a statistical method involving a more general form of Bayesian discrimination with the assumption that the data records of each class follow Gaussian distributions.³⁷ QDA introduces more flexibility into linear discriminant analysis by allowing non-linear separation of the data records.^{19,20}

Four of the remaining seven methods are based on decision trees (DTs). DT employs a single tree structure to assign data records to its various branches,²⁴ connected by nodes based on splitting criteria (e.g., Gini Coefficient),³⁸ applied to the independent variables, and a depth constraint limiting the number of subdivision layers. *Random Forest* (RF) is an ensemble method employing a substantial set of DTs, each trained on different segments of the data set and independent variables, with predictions averaging the results of each DT.³⁰ *Adaptive boosting* (ADA), also known as Ada-boost,²² is an ensemble method employing a set of DTs as its base learners. It tweaks the DT values over a series of iterations giving more weight to those data records misclassified in the previous iteration. *Extreme gradient boosting* (XGB) is also an ensemble method boosting the performance of underlying DTs by progressively improving upon the residuals of the previous iteration. It does this by adjusting a regularization function combining L1 and L2 components.³⁴

K-nearest Neighbor (KNN) employs data matching as its classification technique,²⁶ involving no correlation or statistical assumptions and no regression calculations. It measures the collective differences between all the independent variables of data records in a data set and selects the K number of most closely matching data

records. K can vary typically between 2 and 25. Weights are applied to those closest matches depending on the magnitude of their distances from the data record to be classified, to predict that record's class.

Multi-layer Perceptron (MLP) models are shallow forms of a neural network with input and output layers fully connected by one or multiple hidden layers with adjustable numbers of nodes.²³ Weights and biases applied to the nodes and layers, contribute to calculations involving activation functions, and are adjusted over a series of training iterations. MLPs apply back-propagation or other optimization algorithms to progressively reduce a loss function.

Support vector classification (SVC) is a nonprobabilistic method involving linear and nonlinear components.³² It expresses the independent variables in multidimensional hyperspace space (one dimension for each variable) attempting to maximize the distance separating the dependent variable classes. That separation facilitates the optimal positioning of a linear boundary or hyperplane to classify the data records in the defined hyperspace. Kernel functions, which can be linear or nonlinear (e.g., radial basis function [RBF]), map lower dimensional data values into higher dimensions. Independent variable values lying close to the hyperplane are referred to as "support vectors," as they are instrumental in defining the optimum hyperplane's position.

These 10 models require some tuning adjustments to their hyperparameters to optimize their performance with respect to the Sylhet-520 data set evaluated. The hyperparameter values used to define the models applied are provided in Table 4. Several approaches

TABLE 4 Regression and machine learning model structures and hyperparameters applied

| Prediction models applied | Hyperparameter values applied |
|---|--|
| Regression models | |
| Logistical Regression Classifier (LGR) | L1 ratio = 0.5; solver = saga |
| Gaussian Naïve Bayes Classification (NBC) | Priors = none; variance smoothing = $1e^{-9}$ |
| Quadratic Discriminant Analysis (QDA) | Priors = none; regularization parameter = 0 |
| Adaptive Boosting (ADA) | Number of estimators = 1000; learning rate = 0.01; splitter = best; base estimator criterion = Gini; base estimator is DT with depth = 10 |
| Decision Tree (DT) | Maximum depth = 10,000; splitter = best; splitting criterion = entropy |
| K-Nearest Neighbor (KNN) | Number of nearest neighbors assessed $K = 10$; distance metric = Minkowski with $p = 2$ (Euclidian); neighbor selection algorithm = auto |
| Multi-layer Perceptron (MLP) | 3 Hidden layers with 100, 50, and 25 neurons; learning rate = adaptive; initial learn rate = 0.001; solver = adam; $\alpha = 0.01$; activation fn. = relu |
| Random Forest (RF) | Number of estimators = 1000; maximum depth = 50; splitting criterion = Gini |
| Support Vector Classifier (SVC) | Kernel = rbf; $C = 1$; $\gamma = 0.25$ |
| Extreme Gradient Boosting (XGB) | Number of estimators = 5000; maximum depth = 5; $\eta = 0.01$; subsample = 0.6; columns sampled per tree = 0.8; booster = gbtrees |

were adopted to derive these values, including trial-and-error tests of specific values, grid search,³⁹ and Bayesian optimization.⁴⁰

Another important consideration in classification models is to determine the appropriate number of available data records to assign between model training (the training subset) and model validation (the validation subset). This data record division is typically expressed as a fractional or percentage split ($X\%$ training subset: $Y\%$ validation subset) and referred to as “the split.” $X + Y = 1$ or 100% , and the Y data records are excluded from the training process, while X data records are excluded from trained model validation analysis. The selected split needs to be carefully selected to provide a sufficiently diverse number of data records to facilitate training. However, the split also needs to allocate sufficient data records for validation such that the validation results are statistically reproducible based on the application of multiple random splits.

Determining the split by repeated trials can be time-consuming and can lack statistical rigor unless a large number of trials are run. K -fold cross-validation is a technique that can rigorously determine the appropriate split to use with a specific data set.⁴¹ In this technique, a data set is divided randomly, multiple times into K data subsets, each of equal size. K values between 4 and 15 are typically found to be the most informative. Each of the K subsets is used once as the validation subset with the remaining data records allocated to the training subset. For fourfold cross-validation this means that four cases are required, whereas a 15-fold cross-validation involves 15 separate cases. As the K -fold splits are initially made randomly, it is statistically advisable to repeat the K -fold analysis in several runs. Hence, repeating a fourfold analysis three times results in 12 cases being evaluated, and repeating a 15-fold analysis three times results in 45 cases being evaluated.

Presenting and comparing the results of multiple K -folds in multiple runs is referred to as “multi- K -fold cross-validation analysis.” It is common practice in ML studies that include K -fold cross-validation analysis to consider or present the results of just one K -fold to justify using a specific training/testing split. However, analysis of multiple K -folds is substantially more informative than evaluating a single K -fold. Each K -fold analysis provides an expected mean and associated uncertainty value which, when considered collectively, reveal the optimum training/testing split to use for detailed ML analysis.

It is important to assess the results of multi- K -fold cross-validation analysis statistically. Based on the multiple runs made, this can be meaningfully performed by calculating the means and standard deviations of a loss function and comparing those statistics for the different K -folds considered. In this study, 4-, 5-, 10-, and

15-fold cross-validation analysis is conducted and mean and standard deviations of the mean absolute error (MAE) are compared. The results of that analysis are presented and interpreted in Section 3.2.

2.3 | Considered measures of classification error

Three statistical metrics are computed to assess the classification performances of the 10 models applied to the Sylhet-520 data set. These are MAE, root mean squared error, and coefficient of determination (R^2). For classification problems, it is also appropriate to assess other metrics focused specifically on the classification accuracy and types of misclassifications (e.g., false positives [FPs] and false negatives [FNs]). Such metrics, commonly applied to assess classification problems are: the absolute number of prediction errors (Error #); the percentage of errors relating to models applied to the complete data set (Error %); accuracy (A), precision (P), and recall (R); and balanced F1 score. These specific classification metrics are usefully displayed as part of an annotated confusion matrix, which for a binary classification problem displays four distinct compartments distinguishing true positives, true negatives, FP, and FN. The statistical and misclassification metrics computed in this study are defined in Figure S1.

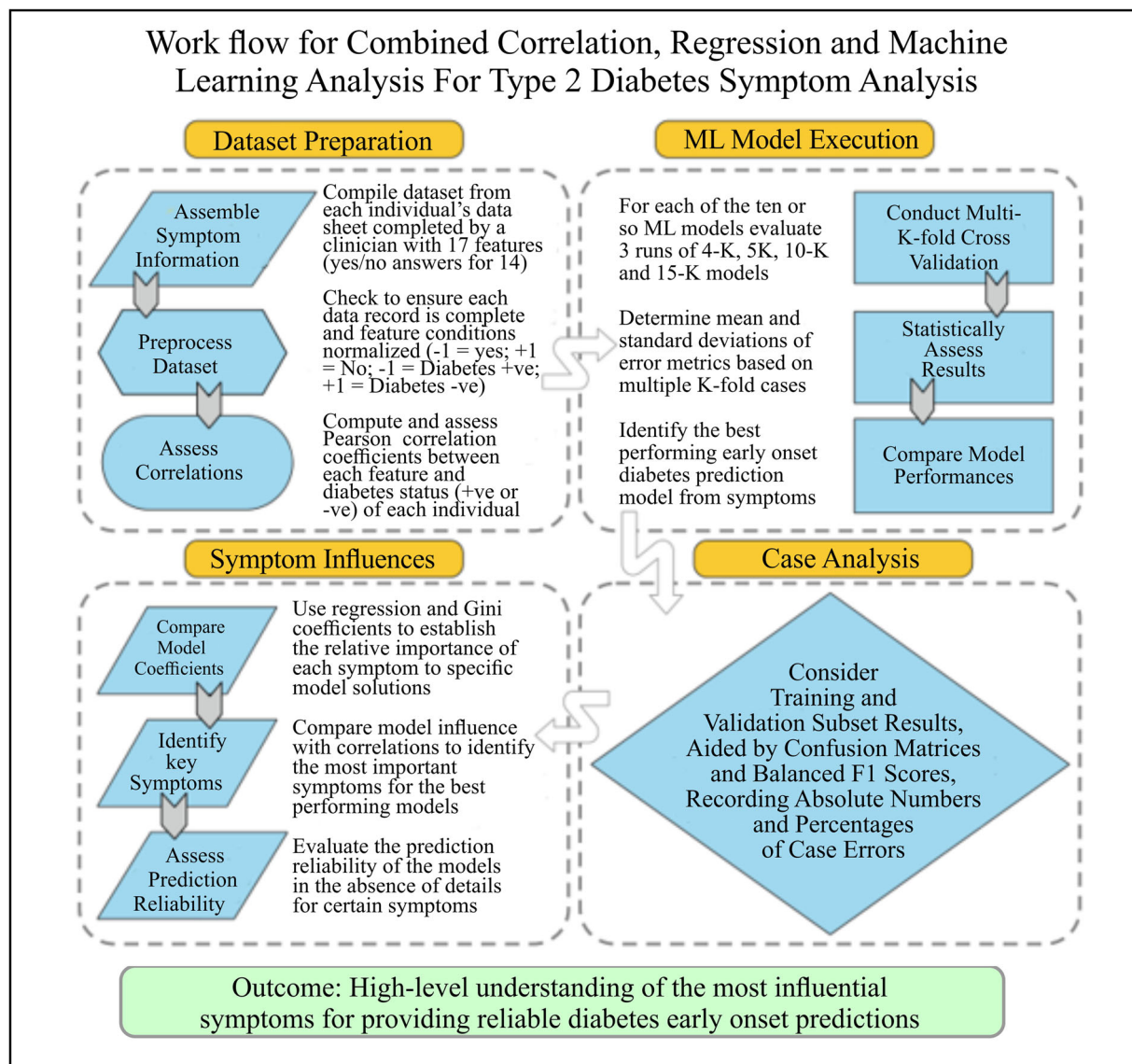
2.4 | Workflow of applied classification methodology

Figure 1 describes the components involved in the integrated workflow methodology applied in this study to characterize and classify the Sylhet-520 data set. It involves correlation, statistical analysis, ML models with multi- K -fold-cross-validation, annotated confusion matrices, and relative importance assessments of potentially influential criteria. This integrated approach provides more valuable insight into the early-onset type 2 diabetes data set that can be derived by merely conducting a misclassification error analysis of the classification models applied.

3 | RESULTS

3.1 | Characterizing the Sylhet type-2 diabetes symptom data set

Nine of the signs and symptoms covered in the assessments of data set individuals are shown to have substantially positive Pearson correlation coefficients with type 2 diabetes diagnosis (Figure 2).^{42,43}



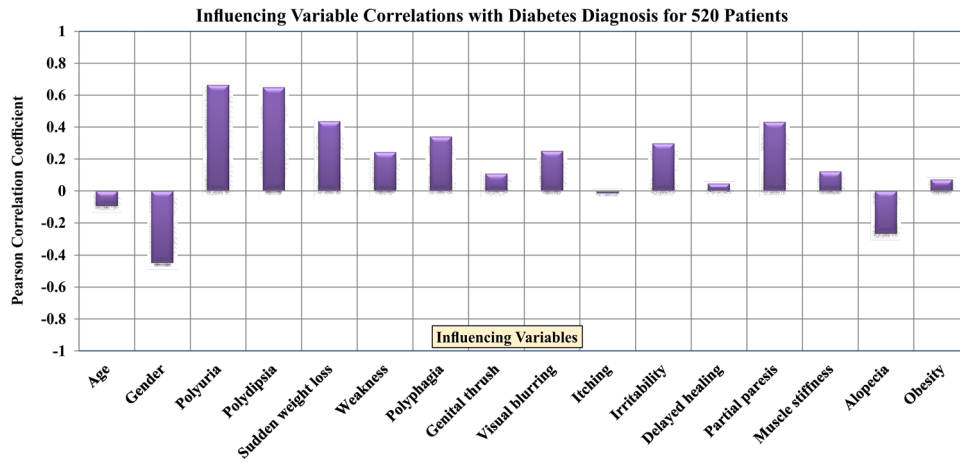


FIGURE 2 Pearson correlation coefficients for 16 criteria versus diabetes type 2 diagnosis for 520 individuals. Positive answers for the symptom questions with moderate to high positive correlation coefficients indicate that individuals suffering those symptoms are more likely than not to have tested positive for type 2 diabetes. The standard correlation coefficient scale varies from -1.0 to $+1.0$, with a value of -1.0 representing perfect negative correlation, a value of $+1.0$ representing perfect positive correlation, and a value of zero representing no correlation. Moderate positive correlations are considered to be those from $+0.25$ to $+0.5$, whereas high positive correlations are considered to be those $>+0.5$.

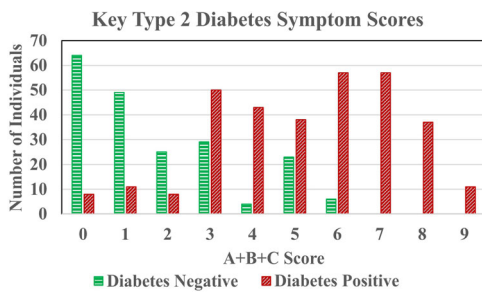


FIGURE 3 Collective scores for the nine features constituting the three categories A, B, and C. An A + B + C score of zero means that the individual was assigned “No” for all nine features. An A + B + C score of 9 means that the individual was assigned “Yes” for all nine features.

status of individuals with scores of between 3 and 6 (Figure 3).

The capability of distinguishing the diabetes status for individuals with A + B + C scores of 3–5 is improved if the relative contributions of category A, B, and C symptoms to the A + B + C score are considered. Figure 4 distinguishes the relative contributions of each category (A or B or C) to the A + B + C score. It is apparent that all the individuals for which the category A or category C symptoms make up more than 50% of A + B + C scores between 3 and 5 are type 2 diabetes-positive (Figure 4A,C). However, Figure 4B shows that this is not the case for the category B symptoms. Some diabetes-negative individuals display category B symptoms making up between 50% and 75% of their A + B + C scores between 3 and 6. This implies that category A and C symptoms are more useful in distinguishing between type 2 diabetes-positive and type 2 diabetes-negative individuals than category B symptoms.

3.2 | Multiple K-fold ML predictions

Statistical and ML classification methods taking into account all 16 independent variables (Table 1) for the 520 individuals in the Sylhet data set are able to classify the type 2 diabetes-negative and type 2 diabetes-positive patients more reliably than scoring (e.g., A + B + C score) or graphical techniques focused on just some of the signs and symptoms (Figures 2 and 3). Moreover, by employing rigorous multi-K-fold cross-validation analysis, the relative statistical robustness of these classification methods can be demonstrated.

Table 5 displays the mean MAE and standard deviation MAE for each classification model employed based on 4-, 5-, 10-, and 15-fold cross-validation analysis of the data set. Table 5 translates these results into the mean number of absolute errors (Error#) and mean percentage errors (Error%) for the number of cases run for each fold of cross-validation. These results are useful for determining the best-performing models and the data-record splits between training and validation subsets that provide the most reliable results.

The results reveal that the SVC model provides better predictions of the onset of type 2 diabetes than the other models. It is followed by RF, MLP, and XGB models, in that order (Table 5). On the other hand, the NBC, LGR, and QDA models, in that order, provide poorer predictions than the other models. For most models, the 10-fold cross-validation evaluations yield the lowest mean MAE values compared with the other K-folds applied to each model. This is not the case for the LGR and NBC models for which the fivefold evaluations outperform the other K-folds in terms of mean MAE, or for the ADA and KNN models for which the 15-fold evaluations outperform the other K-folds (Table 5). For

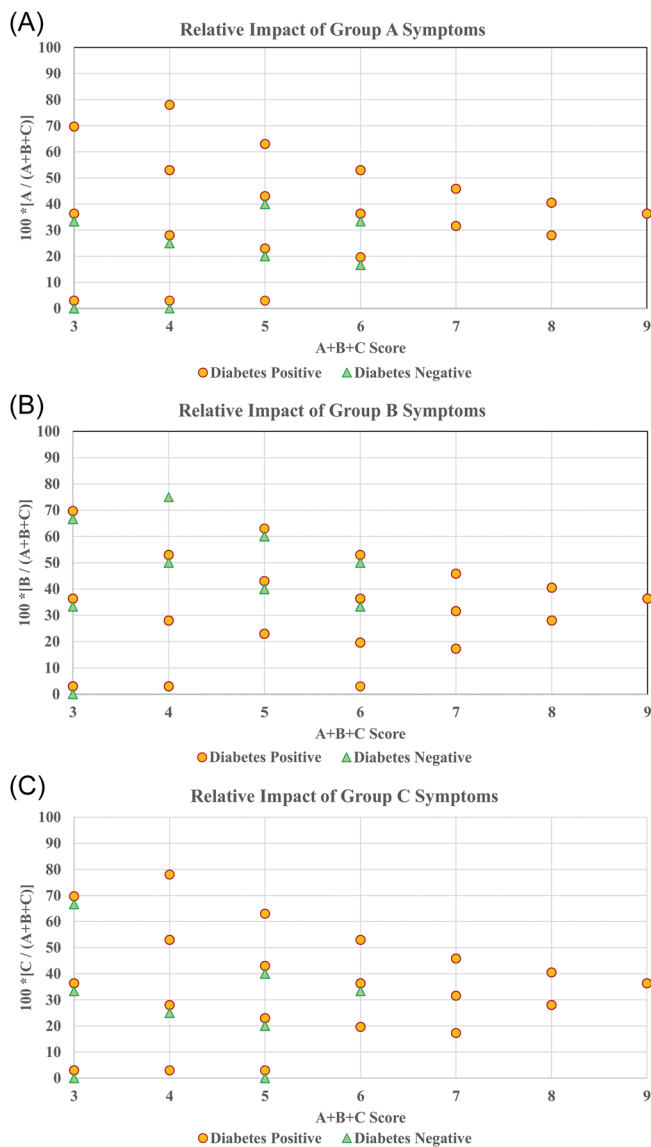


FIGURE 4 Relative contributions of symptom categories A, B, and C to the A + B + C scores: (A) considers category A (polyuria, polydipsia, and polyphagia) contributions; (B) considers category B (visual blurring, partial paresis, and muscle stiffness) contributions; and (C) considers category C (sudden weight loss, weakness, and irritability) contributions.

the RF and SVC models, the mean MAE values are only slightly lower for the 10-fold than for the 15-fold evaluation, although the MAE standard deviations are lower for the 10-fold case. These results suggest that it is more appropriate to use 90%:10% (training subset: validation subset) splits for this data set to obtain the most reliable prediction results from the best-performing models. The results for the 10- and 15-fold cross-validation analysis are displayed in Figure 5.

The mean number of prediction errors is only 11 for the SVC model applying the 10- and 15-fold evaluations (Table 6). Eleven errors constitute just over 2% of the 520 data records. Indeed, excluding the LGR, NBC, and

QDA models, the other seven models all achieve prediction errors of less than 5%. The superiority of the SVC model, followed by the RF, MLP, XGB, and ADA models, in that order, is confirmed in Tables 5, 6 and Figure 5.

The training and validation subset results for an example of one of the randomly selected solutions (i.e., Case Y; 1 of 30 model executions making up the 10-fold analysis) are presented for each model in Table 7. The trained model for Case Y is also applied to the full data set with the number of errors and percentage errors displayed for that case in Table 7.

It is apparent for Case Y that SVC is slightly outperformed by the ADA, MLP, RF, and XGB models (achieving one less incorrect prediction; four errors vs. five errors for SVC). However, the 10-fold cross-validation analysis has established that by averaging 30 such 90%:10% splits SVC outperforms those models. The standard deviations from Table 5 provide an indication of the kind of fluctuations to expect in MAE and error numbers for each case run. The Case Y results applied to the full data set (Table 7) are displayed in Figure 6A,B. The results are clearly consistent with the 10-fold analysis results, highlighting the inferior prediction performance of the NBC, LGR, and QDA models. However, these results also emphasize that the results of one random case should not be taken as an indication of a model's performance over multiple cases, confirming that the multi-K-fold analysis is most reliable for that purpose.

The computational execution times for each model are provided in Table 7 (including the time to conduct the 10-fold cross-validation analysis). The SVC model executes very quickly in comparison to the other high-performing models, making it potentially more attractive for automated systems involving larger datasets.

For individual case runs, such as Case Y, it is also worthwhile displaying the results in terms of an annotated confusion matrix, including calculations of accuracy, precision, recall, and balanced F1-score (see Supporting Information: Appendix SA for definitions of those metrics). Figure 7 illustrates confusion matrices for Case Y prediction results for the SVC and NBC models.

For the high-performing SVC model, Case Y generates zero false-negative prediction errors and just five FPs (Figure 7A). It achieves balanced F1 scores of about 0.99 for both type 2 diabetes-positive and -negative individuals. On the other hand, the poor-performing NBC model generates 30 false-negative prediction errors and 29 FPs for Case Y (Figure 7B). It achieves balanced F1 scores of about 0.90 for type 2 diabetes-positive and about 0.85 for type 2 diabetes-negative individuals.

The methodology applied (Figure 1) meaningfully integrates the analysis of multi-K-fold cross-validation (Figure 5), random case evaluation (Figure 6), and annotated confusion matrices (Figure 7) to make

TABLE 5 K-fold cross-validation analysis of 10 type 2 diabetes status prediction models expressed in terms of MAE

| MAE | 4-Fold (12 cases) | | 5-Fold (15 cases) | | 10-Fold (30 cases) | | 15-Fold (45 cases) | |
|-----|-------------------|--------|-------------------|--------|--------------------|---------------|--------------------|---------------|
| | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| LGR | 0.1551 | 0.0446 | 0.1487 | 0.0594 | 0.1500 | 0.0725 | 0.1542 | 0.1027 |
| NBC | 0.2269 | 0.0619 | 0.2244 | 0.0680 | 0.2282 | 0.0968 | 0.2208 | 0.1210 |
| QDA | 0.1167 | 0.0484 | 0.1103 | 0.0468 | 0.1026 | 0.0547 | 0.1038 | 0.0777 |
| ADA | 0.0872 | 0.0538 | 0.0769 | 0.0525 | 0.0692 | 0.0565 | 0.0644 | 0.0718 |
| DT | 0.1064 | 0.0476 | 0.0974 | 0.0509 | 0.0821 | 0.0513 | 0.0900 | 0.0732 |
| KNN | 0.1077 | 0.0562 | 0.1026 | 0.0495 | 0.0885 | 0.0660 | 0.0809 | 0.0684 |
| MLP | 0.0718 | 0.0484 | 0.0782 | 0.0602 | 0.0551 | 0.0532 | 0.0616 | 0.0716 |
| RF | 0.0667 | 0.0500 | 0.0628 | 0.0499 | 0.0526 | 0.0548 | 0.0527 | 0.0670 |
| SVC | 0.0577 | 0.0417 | 0.0487 | 0.0313 | 0.0423 | 0.0500 | 0.0437 | 0.0522 |
| XGB | 0.0808 | 0.0526 | 0.0718 | 0.0468 | 0.0590 | 0.0542 | 0.0681 | 0.0766 |

Note: The SVC model generates the lowest MAE values for the 10- and 15-fold solutions (in bold). MAE is expressed in terms of the value difference between negative type 2 diabetes (-1) and positive (+1) type 2 diabetes. Each incorrect prediction therefore contributes an absolute error value of 2 to the MAE computation.

Abbreviations: ADA, adaptive boosting; DT, decision tree; KNN, K-nearest neighbor; LGR, logistic regression; MAE, mean absolute error; MLP, multilayer perceptron; NBC, Naïve Bayes classifier; QDA, quadratic discriminant analysis; RF, random forest; SD, standard deviation; SVC, support vector classification; XGB, extreme gradient boosting.

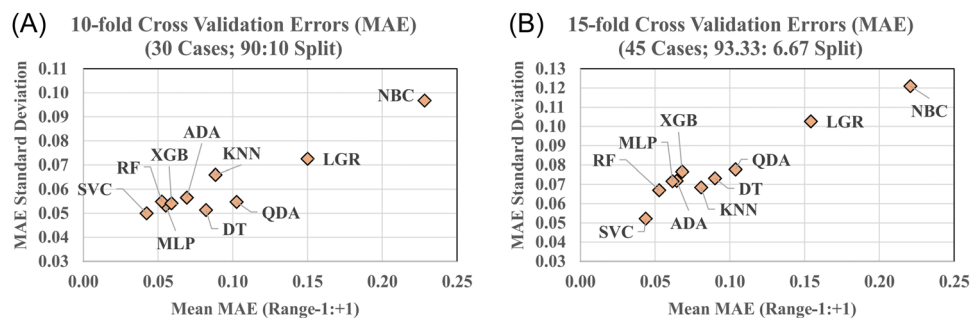


FIGURE 5 Standard deviation MAE versus mean MAE for 10 prediction models applied to the Sylhet Diabetic Hospital published data set for: (A) 10-fold cross-validation analysis; and (B) 15-fold cross-validation analysis. ADA, adaptive boosting; DT, decision tree; KNN, K-nearest neighbor; LGR, logistic regression; MAE, mean absolute error; MLP, multilayer perceptron; NBC, Naïve Bayes classifier; QDA, quadratic discriminant analysis; RF, random forest; SVC, support vector classification; XGB, extreme gradient boosting.

detailed comparisons of the prediction performances of the models considered. Combined with the statistical characterization of the data set, this provides maximum insight into the model capabilities as applied to this specific data set.

4 | DISCUSSION

Further insight into some of the prediction models can be gained by considering information that indicates the relative importance of each of the signs and symptoms to the model solutions. Such information can be gleaned from the LGR model by considering the absolute magnitudes of its regression coefficients, from the SVC model by considering the absolute magnitudes of its support vector coefficients, and for the tree/ensemble-

tree models (ADA, DT, RF, XGB) by considering their Gini (or entropy) coefficients. Unfortunately, it is not possible to extract such information from the NBC, QDA, KNN, and MLP models.

Figure 8 plots the relative importance assigned to each sign and symptom by the LGR, SVC, DT, RF, and XGB model solutions applied to the Sylhet-520 data set.

The relative importance of each variable in Figure 8 should be compared to their correlations with the diabetes diagnosis (Figure 2). It is apparent that polydipsia is given more weight by all of the models considered in Figure 8 than other signs or symptoms, even though polyuria has a higher correlation with the type 2 diabetes status of the 520 individuals collectively. Polyuria and gender are the next most important criteria for most of the models compared in Figure 8. On the

TABLE 6 K-fold cross-validation analysis of 10 type 2 diabetes status prediction models expressed in terms of actual numbers of errors and percentage errors

| Mean absolute and percentage errors | 4-Fold (12 cases) | | 5-Fold (15 cases) | | 10-Fold (30 cases) | | 15-Fold (45 cases) | |
|-------------------------------------|-------------------|---------|-------------------|---------|--------------------|--------------|--------------------|--------------|
| | Error # | Error % | Error # | Error % | Error # | Error % | Error # | Error % |
| LGR | 40.3 | 7.76% | 38.7 | 7.44% | 39.0 | 7.50% | 40.1 | 7.71% |
| NBC | 59.0 | 11.35% | 58.3 | 11.22% | 59.3 | 11.41% | 57.4 | 11.04% |
| QDA | 30.3 | 5.83% | 28.7 | 5.51% | 26.7 | 5.13% | 27.0 | 5.19% |
| ADA | 22.7 | 4.36% | 20.0 | 3.85% | 18.0 | 3.46% | 16.7 | 3.22% |
| DT | 27.7 | 5.32% | 25.3 | 4.87% | 21.3 | 4.10% | 23.4 | 4.50% |
| KNN | 28.0 | 5.38% | 26.7 | 5.13% | 23.0 | 4.42% | 21.0 | 4.04% |
| MLP | 18.7 | 3.59% | 20.3 | 3.91% | 14.3 | 2.76% | 16.0 | 3.08% |
| RF | 17.3 | 3.33% | 16.3 | 3.14% | 13.7 | 2.63% | 13.7 | 2.63% |
| SVC | 15.0 | 2.88% | 12.7 | 2.44% | 11.0 | 2.12% | 11.4 | 2.18% |
| XGB | 21.0 | 4.04% | 18.7 | 3.59% | 15.3 | 2.95% | 17.7 | 3.41% |

Note: The SVC model generates the lowest numbers of errors and percentage errors for the 10- and 15-fold solutions (in bold).

Abbreviations: ADA, adaptive boosting; DT, decision tree; KNN, K-nearest neighbor; LGR, logistic regression; MLP, multilayer perceptron; NBC, Naïve Bayes classifier; QDA, quadratic discriminant analysis; RF, random forest; SVC, support vector classification; XGB, extreme gradient boosting.

TABLE 7 Case Y (1 of 30, 10-fold cross-validation cases) results comparing training subset, validation subset, and full-data set prediction errors for models applied to the Sylhet Diabetic Hospital published data set¹³

| 90:10 Split Record # | Case Y training subset | | | | Case Y validation subset | | | | Case Y full data set | | | | | Ex time |
|----------------------|------------------------|--------|--------|---------|--------------------------|--------|--------|---------|----------------------|--------|--------|---------|---------|---------|
| | 468 | | | | 52 | | | | 520 | | | | | |
| Model | R ² | RMSE | MAE | Error # | R ² | RMSE | MAE | Error # | R ² | RMSE | MAE | Error # | Error % | |
| LGR | 0.7572 | 0.4804 | 0.1154 | 27 | 0.5752 | 0.6202 | 0.1923 | 5 | 0.7400 | 0.4961 | 0.1231 | 32 | 6.15% | 4.7 |
| NBC | 0.5504 | 0.6537 | 0.2137 | 50 | 0.2353 | 0.8321 | 0.3462 | 9 | 0.5206 | 0.6737 | 0.2269 | 59 | 11.35% | 4.5 |
| QDA | 0.8651 | 0.3581 | 0.0641 | 15 | 0.5752 | 0.6202 | 0.1923 | 5 | 0.8375 | 0.3922 | 0.0769 | 20 | 3.85% | 4.6 |
| ADA | 0.9730 | 0.1601 | 0.0128 | 3 | 0.9150 | 0.2774 | 0.0385 | 1 | 0.9675 | 0.1754 | 0.0154 | 4 | 0.77% | 95.0 |
| DT | 0.9730 | 0.1601 | 0.0128 | 3 | 0.8301 | 0.3922 | 0.0769 | 2 | 0.9594 | 0.1961 | 0.0192 | 5 | 0.96% | 4.6 |
| KNN | 0.9730 | 0.1601 | 0.0128 | 3 | 0.7451 | 0.4804 | 0.1154 | 3 | 0.9513 | 0.2148 | 0.0231 | 6 | 1.15% | 4.5 |
| MLP | 0.9730 | 0.1601 | 0.0128 | 3 | 0.9150 | 0.2774 | 0.0385 | 1 | 0.9675 | 0.1754 | 0.0154 | 4 | 0.77% | 17.7 |
| RF | 0.9730 | 0.1601 | 0.0128 | 3 | 0.9150 | 0.2774 | 0.0385 | 1 | 0.9675 | 0.1754 | 0.0154 | 4 | 0.77% | 32.1 |
| SVC | 0.9730 | 0.1601 | 0.0128 | 3 | 0.8301 | 0.3922 | 0.0769 | 2 | 0.9594 | 0.1961 | 0.0192 | 5 | 0.96% | 4.8 |
| XGB | 0.9730 | 0.1601 | 0.0128 | 3 | 0.9150 | 0.2774 | 0.0385 | 1 | 0.9675 | 0.1754 | 0.0154 | 4 | 0.77% | 40.2 |

Note: RMSE and MAE are expressed in terms of the value difference between negative type 2 diabetes (-1) and positive type 2 diabetes (+1), a value of 2. Model execution times (ex time) are expressed in seconds and include the time taken to conduct 10-fold cross-validation.

Abbreviations: ADA, adaptive boosting; DT, decision tree; KNN, K-nearest neighbor; LGR, logistic regression; MAE, mean absolute error; MLP, multilayer perceptron; NBC, Naïve Bayes classifier; QDA, quadratic discriminant analysis; RF, random forest; RMSE, root mean squared error; SVC, support vector classification; XGB, extreme gradient boosting.

other hand, sudden weight loss, weakness, polyphagia, visual blurring, and partial paresis are not assigned substantial weights despite having moderate correlation coefficients (about +0.2 to +0.4; Figure 2) with type 2 diabetes status.

The DT model stands out in that it assigns a more substantial weight (about 40% of the total weight it assigns) to polydipsia than the other models displayed in Figure 8. The ADA and XGB models also assign

relatively high weights to polydipsia. The relative weights assigned by the regression models LGR and SVC are quite similar in magnitude and quite distinct from those assigned by the tree/ensemble models. For instance, those two models assign comparable weights to gender, polyuria, polydipsia, itching, and irritability, at levels that are distinct from the other models. The RF model assigns weights that tend, for the most part, to fall between those of the other ensemble methods and

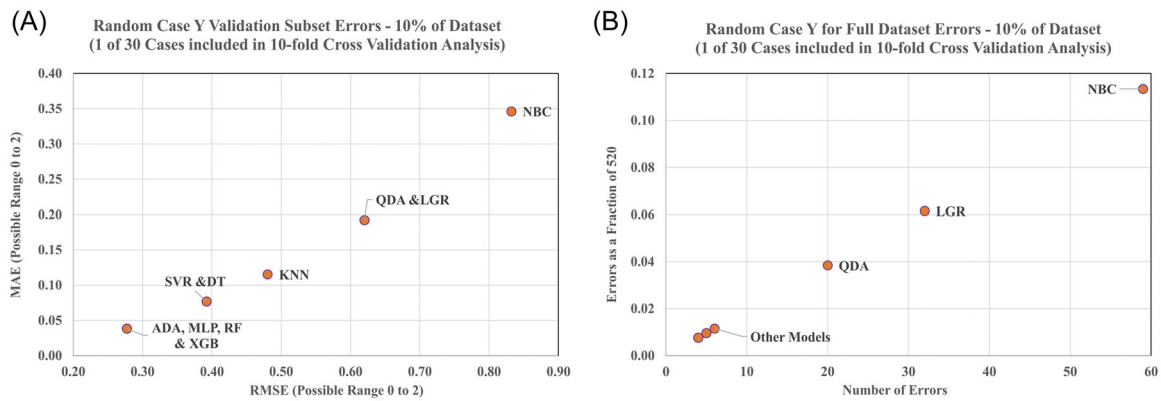


FIGURE 6 Case Y results for 10 prediction models applied to the Sylhet Diabetic Hospital published data set displaying: (A) MAE versus RMSE for the validation subset; and (B) error fraction versus error number for the case solution applied to the full data set (520 data records). ADA, adaptive boosting; KNN, *K*-nearest neighbor; LGR, logistic regression; MAE, mean absolute error; MLP, multilayer perceptron; NBC, Naïve Bayes classifier; QDA, quadratic discriminant analysis; RF, random forest; RMSE, root mean squared error; XGB, extreme gradient boosting.

| (A) SVC Case Y Solution Applied to Full Dataset | | | | (B) NBC Case Y Solution Applied to Full Dataset | | | |
|---|-------------------|-------------------|-------------------|---|-------------------|-------------------|-------------------|
| 520 | Record# | 325 | 195 | 520 | Record# | 319 | 201 |
| Totals | Class | Diabetes Positive | Diabetes Negative | Totals | Class | Diabetes Positive | Diabetes Negative |
| 320 | Diabetes Positive | 320 | 0 | 320 | Diabetes Positive | 290 | 30 |
| 200 | Diabetes Negative | 5 | 195 | 200 | Diabetes Negative | 29 | 171 |
| Precision: | | 98.46% | 100.00% | Precision: | | 90.91% | 85.07% |
| 5 | Errors | 5 | 0 | 59 | Errors | 29 | 30 |
| 99.04% % Successful Predictions | | | | 88.65% % Successful Predictions | | | |
| Class Prediction Performance: | | | | Class Prediction Performance: | | | |
| Accuracy | 0.9904 | | 0.9904 | Accuracy | 0.8865 | | 0.8865 |
| Precision | 0.9846 | | 1.0000 | Precision | 0.9091 | | 0.8507 |
| Recall | 1.0000 | | 0.9750 | Recall | 0.9063 | | 0.8550 |
| Balanced F1 Score | 0.9923 | | 0.9873 | Balanced F1 Score | 0.9077 | | 0.8529 |

FIGURE 7 Case Y annotated confusion matrices applied to full data set displaying accuracy, precision, recall, and balanced F1 score (see Appendix S1 for definitions of those metrics): (A) high-performing SVC model; and (B) poor-performing NBC model. NBC, Naïve Bayes classifier; SVC, support vector classification.

the SVC/LGR models. It is noticeable that obesity and age are assigned relatively low weights by all models.

Although the SVM and LGR models assign similar weights to most of the signs and symptoms they perform quite differently in their abilities to predict diabetes status. The SVM model performs much better than the LGR model (Tables 5 and 6, Figure 5). This is most likely explained in terms of the different mathematical concepts underpinning those two models. Whereas LGR is a binary, categorical classification model applying linear relationships between the independent variables and the dependent variable, SVC combines linear computations, associated with the definition of its optimum hyperplane, and nonlinear relationships associated with its RBF kernel used to locate that hyperplane. It appears that the nonlinear aspect of the RBF component of the SVM model provides it with a

substantial advantage over the linear LGR model in type 2 diabetes status predictions when applied to the studied data set.

Establishing distinct influences and relative importance of the signs and symptoms of the high-performing prediction models is important information to ascertain for two reasons. Firstly, it helps focus a clinician's attention on the key signs and symptoms to look for in type 2 diabetes screening. Secondly, it suggests that using one or other model, or correlation coefficients, to conduct feature selection to limit the number of signs and symptoms to input to other prediction models is probably not a good idea. Each feature exerts a different influence on each of the prediction models, therefore it is risky to disregard any of them. For instance, the "itching" criterion is assigned a low weight by the tree/ensemble methods and has a low correlation with

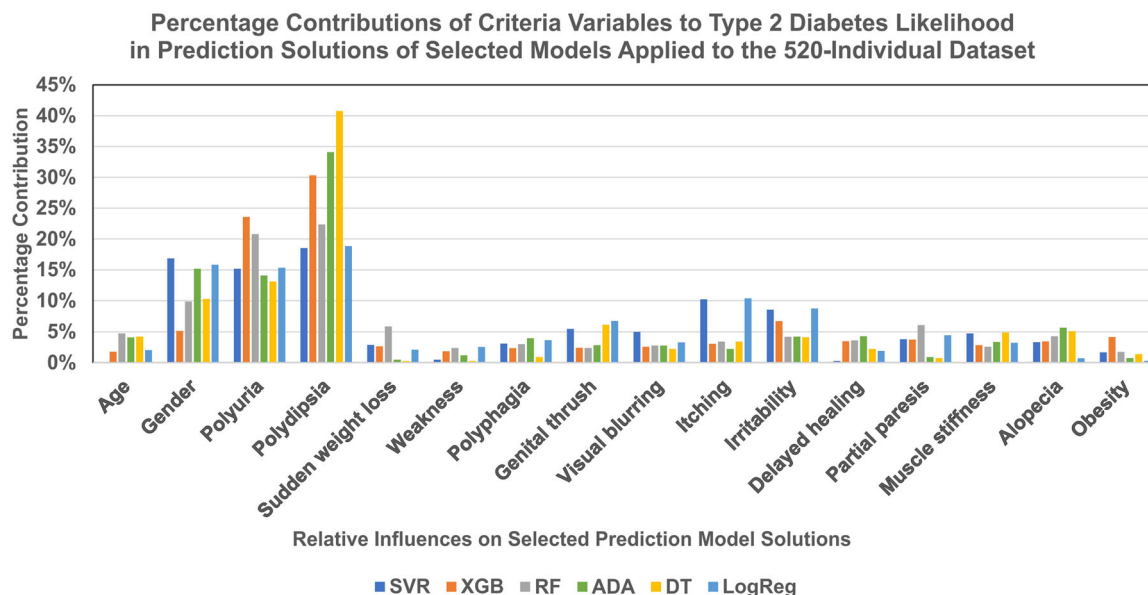


FIGURE 8 The influences of specific signs and symptoms on selected type 2 diabetes likelihood model solutions. ADA, adaptive boosting; DT, decision tree; RF, random forest; XGB, extreme gradient boosting.

diabetes status, yet it is assigned a relatively high 10% weight by the SVC and LGR models.

Although the SVC, RF, MLP, and XGB models perform well with the Sylhet-520 data set evaluated, further study is required to be able to generalize these models for application to other datasets involving larger numbers of individuals and broader geographic areas. One issue is the significance of weights assigned by the models to individual age and gender as influential variables. If these variables are to be used for larger datasets, it is important to ensure that the age and gender distributions of the individuals tested, and percentages of type 2 diabetes-positive and type 2 diabetes-negative individuals represented, are comparable with those observed in the region/nation being evaluated. The Sylhet-520 is somewhat skewed in this regard. For instance, of the 192 females represented in the data set 173 are diabetes-positive (90%), and of the 328 males represented in the data set 147 are diabetes-positive (45%). Hence, correlations between gender and diabetes status established for the Sylhet-520 data set are not likely to be meaningful when applied to other data sets.

5 | CONCLUSIONS

An integrated methodology combining correlations, statistical analysis, multiple ML models, multi-*K*-fold cross-validation, and confusion matrices, not only generates reliable and reproducible classifications of type 2 diabetes-positive and type 2 diabetes-negative individuals from a suite of 16 signs and symptoms (features) but it also provides insight to the relative influences of those

features in generating those predictions. Of the 10 ML and statistical prediction methods, applied to a published data set of 520 individuals, the Support Vector Classifier demonstrates the best performance. It generates only 11 misclassifications (2.1% error) based on an average of thirty 10-fold cross-validation runs. The RF, MLP, and XGB models also perform well, generating predictions with, on average, less than 3% misclassifications. On the other hand, the statistical methods, such as naïve Bayes, logistical regression, and QDA perform comparatively less accurately, generating between 5.2% and 11.4% average misclassification.

Multiple *K*-fold cross-validation is highly effective at identifying the best-performing prediction models and establishing the most reliable data record splits to apply to training and validation subsets. Comparison of 4-, 5-, 10-, and 15-fold analysis using multiple runs, and considering mean and standard deviations of error metrics, reveal the 10-fold configuration (involving 30 separate cases split 90% training: 10% validation) to generate the least errors for the SVC (and most other) models. This was closely followed by the 15-fold configuration. In contrast, the 4-fold configuration generated the poorest prediction performance.

Nine of the potentially influential features assessed were found to display the highest positive correlations with diabetes status in the data set studied. These features can be divided into metabolic impacts (polyuria, polydipsia, and polyphagia), mobility/muscular impacts (visual blurring, partial paresis, and muscle stiffness), and demeanor changes (sudden weight loss, weakness, and irritability). Statistical analysis reveals that the metabolic impacts and demeanor changes are

collectively more effective indicators of diabetes status than the mobility/muscular impacts. However, for individuals afflicted by less than three of the nine features that are more highly correlated with type 2 diabetes status, there is a high incidence of both type 2 diabetes-positive and -negative results. This makes it unreliable to use these few features for type 2 diabetes screening predictions.

Assessments of the relative influences of features on the solutions of high-performing ML prediction models reveal that polydipsia is assigned the highest weight. For the SVC model, this is followed by gender, polyuria, itching, and irritability, whereas obesity and age are assigned very low weights. These relative influences are quite distinct from the magnitude of the correlation coefficients of these features with diabetes status. The relative influences also vary in detail from one ML model to another. These findings suggest that caution is required when making decisions to disregard certain features because of their low correlations or low weights assigned by screening models. Features displaying low correlations with diabetes status or assigned low importance by one quick-to-execute ML or statistical model may have a substantial influence on other ML models.

Some of the feature influences are of specific relevance to the data set studied (e.g., gender and age), due to the makeup of the gender/age mix sampled, and the percentage of type 2 diabetes-positive and -negative individuals represented in each group. To generalize early-onset type 2 diabetes screening models for society as a whole, involving age and gender as potentially influencing features, makes it necessary to ensure sample distribution balances that are representative of the society being tested.

AUTHOR CONTRIBUTION

David A. Wood is the sole author of this study and has conducted all of its preparation, modeling, interpretation, analysis, and writing.

ACKNOWLEDGMENT

No external funding was received for this study.

CONFLICT OF INTEREST

The author declares no conflict of interest.

DATA AVAILABILITY STATEMENT

The data used in this study is publicly available and can be downloaded at UCS Data set¹³: <https://archive.ics.uci.edu/ml/datasets/Early%2Bstage%2Bdiabetes%2Brisk%2Bprediction%2Bdataset.#>

ETHICS STATEMENT

The data used in this study comes from a public database and the individuals involved in that database

are anonymous. No ethical or personal data disclosure issues have therefore arisen during this study.

ORCID

David A. Wood  <http://orcid.org/0000-0003-3202-4069>

REFERENCES

1. American Diabetes Association. Classification and diagnosis of diabetes: standards of medical care in diabetes. *Diabetes Care*. 2021;44(suppl 1):S15-S33. doi:10.2337/dc21-S002
2. Kaur G, Lakshmi P, Rastogi A, et al. Diagnostic accuracy of tests for type 2 diabetes and prediabetes: a systematic review and meta-analysis. *PLoS One*. 2020;15(11):e0242415. doi:10.1371/journal.pone.0242415
3. International Diabetes Federation. IDF Diabetes Atlas 10th Edition; 2021. Accessed April 15, 2022. <https://diabetesatlas.org/atlas/tenth-edition/>
4. Diabetes Australia. Failure to detect type 2 diabetes early costing \$700 million per year. 2018. Accessed April 15, 2022. <https://www.diabetesaustralia.com.au/mediarelease/failure-to-detect-type-2-diabetes-early-costing-700-million-per-year/>
5. Herman WH, Ye W, Griffin SJ, et al. Early detection and treatment of type 2 diabetes reduce cardiovascular morbidity and mortality: a simulation of the results of the Anglo-Danish-Dutch study of intensive treatment in people with screen-detected diabetes in primary care (ADDITION-Europe). *Diabetes Care*. 2015;38(8):1449-1455. doi:10.2337/dc14-2459
6. Sami W, Ansari T, Butt NS, Hamid M. Effect of diet on type 2 diabetes mellitus: a review. *Int J Health Sci*. 2017;11(2):65-71.
7. Kopitar L, Kocbek P, Cilar L, Sheikh A, Stiglic G. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Sci Rep*. 2020;10:11981. doi:10.1038/s41598-020-68771-z
8. Moura AM, Antunes M, Martins SO, Raposo JF. A statistical model to identify determinants of glycemic control in patients with type 2 diabetes with different pharmacotherapeutic profiles. *PLoS One*. 2020;15(7):e0235376. doi:10.1371/journal.pone.0235376
9. Zou D, Ye Y, Zou N, Yu J. Analysis of risk factors and their interactions in type 2 diabetes mellitus: a cross-sectional survey in Guilin, China. *J Diabetes Investig*. 2017;8(2):188-194. doi:10.1111/jdi.12549
10. Joshi RD, Dhakal CK. Predicting type 2 diabetes using logistic regression and machine learning approaches. *Int J Environ Res Public Health*. 2021;18(14):7346. doi:10.3390/ijerph18147346
11. Fregoso-Aparicio L, Noguez J, Montesinos L, Garcia-Garcia JA. Machine learning and deep learning predictive models for type 2 diabetes: a systematic review. *Diabetol Metab Syndr*. 2021;13:148. doi:10.1186/s13098-021-00767-9
12. Islam MMF, Ferdousi R, Rahman S, Bushra HY. Likelihood prediction of diabetes at early stage using data mining techniques. In: Konar D, Gupta M, Biswas S, Bhattacharyya S, eds. *Computer Vision and Machine Intelligence in Medical Image Analysis*. Springer; 2020:113-125. doi:10.1007/978-981-13-8798-2_12
13. UCI Dataset. Early stage diabetes risk prediction dataset. UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science. 2020. Accessed April 15, 2022. <https://archive.ics.uci.edu/ml/datasets/Early%2Bstage%2Bdiabetes%2Brisk%2Bprediction%2Bdataset.#>
14. Berkson J. Application of the logistic function to bio-assay. *J Am Stat Assoc*. 1944;39(227):357-365. doi:10.2307/2280041
15. Tabaei BP, Herman WH. A multivariate logistic regression equation to screen for diabetes: development and validation.

- Diabetes Care.* 2002;25(11):1999-2003. doi:10.2337/diacare.25.11.1999
16. Rajendra P, Latifi S. Prediction of diabetes using logistic regression and ensemble techniques. *Comput Methods Programs Biomed.* 2021;1:100032. doi:10.1016/j.cmpbup.2021.100032
 17. Hand DJ, Yu K. Idiot's Bayes—not so stupid after all? *Int Stat Rev.* 2001;69(3):385-399. doi:10.2307/1403452
 18. Theresa JS, Evangeline DJ. Classification of diabetes mellitus using Naive Bayes algorithm. In: Peter J, Fernandes S, Alavi A, eds. *Intelligence in Big Data Technologies—Beyond the Hype. Advances in Intelligent Systems and Computing.* Springer; 2021: 1167. doi:10.1007/978-981-15-5285-4_40
 19. Fischer RA. The use of multiple measurements in taxonomic problems. *Ann Eugen.* 1936;7:179-188. doi:10.1111/j.1469-1809.1936.tb02137.x
 20. Tharwat A. Linear vs. quadratic discriminant analysis classifier: a tutorial. *Int J Appl Pattern Recognit.* 2016;3(2):145. doi:10.1504/IJAPR.2016.079050
 21. Maniruzzaman M, Rahman MJ, Al-MehediHasan M, et al. Accurate diabetes risk stratification using machine learning: role of missing value and outliers. *J Med Syst.* 2018;42(5):92. doi:10.1007/s10916-018-0940-7
 22. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. *J Comput Syst Sci.* 1997;55:119-139. doi:10.1006/jcss.1997.1504
 23. Vijayan VV, Anjali C. Prediction and diagnosis of diabetes mellitus—a machine learning approach. 2015 *IEEE Recent Advances in Intelligent Computational Systems.* 2015. pp. 122-127. doi:10.1109/RAICS.2015.7488400
 24. Quinlan JR. Induction of decision trees. *Mach Learn.* 1986;1: 81-106. doi:10.1007/BF00116251
 25. Ramezankhani A, Hadavandi E, Pournik O, Shahrabi J, Azizi F, Hadaegh F. Decision tree-based modelling for identification of potential interactions between type 2 diabetes risk factors: a decade follow-up in a Middle East prospective cohort study. *BMJ Open.* 2016;6:e013336. doi:10.1136/bmjopen-2016-013336
 26. Fix E, Hodges JL Jr. *Discriminatory analysis, nonparametric discrimination: consistency properties.* USAF School of Aviation Medicine. Technical Report 4; 1951. p. 21.
 27. Sarkar IH, Faruque F, Alqahtani H, Kalim A. K-Nearest Neighbor learning based diabetes mellitus prediction and analysis for eHealth services. *EAI End Trans Scal Inf Sys.* 2019;7(26):e4. doi:10.4108/eai.13-7-2018.162737
 28. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol Rev.* 1958;65(6):386-408. doi:10.1037/h0042519
 29. Bani-Salameh H, Alkhatib SM, Abdalla M, et al. Prediction of diabetes and hypertension using Multi-Layer Perceptron neural networks. *Int J Model Simul Sci Comput.* 2021;12(2):2150012. doi:10.1142/S1793962321500124
 30. Ho TK. The random subspace method for constructing decision forests. *IEEE Trans Pattern Anal Mach Intell.* 1998;20(8):832-844. doi:10.1109/34.709601
 31. Wang X, Zhai M, Ren Z, et al. Exploratory study on classification of diabetes mellitus through a combined Random Forest Classifier. *BMC Med Inform Decis Mak.* 2021;21:105. doi:10.1186/s12911-021-01471-4
 32. Cortes C, Vapnik V. Support-vector networks. *Mach Learn.* 1995;20(3):273-297. doi:10.1007/BF00994018
 33. Abbas HT, Alic L, Erraguntla M, et al. Predicting long-term type 2 diabetes with support vector machine using oral glucose tolerance test. *PLoS One.* 2019;14(12):e0219636. doi:10.1371/journal.pone.0219636
 34. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Krishnapuram B, Shah M, Smola AJ, et al., eds. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* San Francisco, CA, USA. August 13-17, 2016. pp. 785-794. doi:10.1145/2939672.2939785
 35. Wang L, Wang X, Chen A, Jin X, Che H. Prediction of type 2 diabetes risk and its effect evaluation based on the XGBoost model. *Healthcare.* 2020;8(3):247. doi:10.3390/healthcare8030247
 36. SciKit Learn. Supervised and unsupervised machine learning models in Python. 2022. April 15, 2022. <https://scikit-learn.org/stable/>
 37. Huberty CJ. Discriminant analysis. *Rev Educ Res.* 1975;45(4): 543-598. doi:10.2307/1170065
 38. Gini C. Concentration and dependency ratios (published 1909 in Italian). *Riv di Polit Econ.* 1997;87:769-778.
 39. SciKit Learn. GridSearchCV: Exhaustive search over specified parameter values for an estimator in Python. 2022. April 15, 2022. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
 40. SciKit Learn. Bayesian optimization of hyperparameters in Python. 2022. April 15, 2022. <https://scikit-optimize.github.io/stable/modules/generated/skopt.BayesSearchCV.html>
 41. SciKit Learn. Cross-validation: evaluating estimator performance. 2022. April 15, 2022. https://scikit-learn.org/stable/modules/cross_validation.html
 42. Pearson K. On the dissection of asymmetrical frequency curves. *Phil Trans Roy Soc A.* 1894;185:71-110.
 43. Lawrence I, Lin K. A concordance correlation coefficient to evaluate reproducibility. *Biometrics.* 1989;45:255-268. doi:10.2307/2532051

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Wood DA. Integrated statistical and machine learning analysis provides insight into key influencing symptoms for distinguishing early-onset type 2 diabetes. *Chronic Dis Transl Med.* 2022;8:281-295. doi:10.1002/cdt3.39