Journal of Scientific & Industrial Research
Vol. 85, May 2026, pp. 433-444
DOI: 10.56042/jsir.v85i5.27426

Assessment of Early Diabetes Risk through a Random Forest Classifier

Pragati Choudhari[1], Shalini Goel[2], Sinu Nambiar[3]*, Sushama Shirke[4], Rupali Mangrule[5], Jyoti Deone[6] & Anant Sidhappa Kurhade[7,8]*

[1]School of Computer Science Engineering & Applications, D Y Patil International University, Akurdi, Pune 412 101, Maharashtra, India

[2]Department of Artificial Intelligence & Data Science, IIMT College of Engineering, Greater Noida 201 310, Uttar Pradesh, India

[3]Department of Computer Engineering & Technology, Dr Vishwanath Karad MIT World Peace University, Pune 411 038, Maharashtra, India

[4]Department of Computer Engineering, Army Institute of Technology, Dighi Hills, Pune 411 015, Maharashtra, India

[5]Department of Computer Science and Engineering, Maharashtra Institute of Technology, Chhatrapati Sambhajinagar 431 010, Maharashtra, India

[6]Department of Information Technology, D.Y. Patil Deemed to be University RAIT, Navi Mumbai 400 706, Maharashtra, India

[7]Department of Mechanical Engineering, Dr. D. Y. Patil Institute of Technology, Sant Tukaram Nagar, Pimpri, Pune – 411 018, Maharashtra, India

[8]Dnyaan Prasad Global University (DPGU), School of Technology and Research – Dr. D. Y. Patil Unitech Society, Sant Tukaram Nagar, Pimpri, Pune 411 018, Maharashtra, India

Received 9 January 2026; revised 13 May 2026; accepted 20 May 2026

The purpose of the study was to develop an efficient machine-learning model for early detection of diabetes employing Random Forest algorithm based on a systematic workflow KDD. The findings indicate the model has good overall diagnostic utility and consistent predictive performance. The outcomes showed a reliable performance of case classification 85% accuracy, 78% specificity, 75% recall and 86% AUC for the proposal random forest model, which is good in distinguishing diabetic & nondiabetic cases. This higher specificity translates to fewer false-positive predictions, which is necessary for good clinical screening. The high AUC value is also evidence of consistent classification ability across various threshold levels. The model addressed clinical data variability successfully with an efficient machine learning solution where complex feature engineering or advanced data balance techniques were neither necessary nor segment specific, which makes it appropriate for application in real-world healthcare. The proposed framework and tools will lay the foundation for future developments, such as validation on larger, more diverse datasets, sensitivity improvements using optimization techniques and deployment through APIs for real-time healthcare systems. These results underscore the promise of structured machine learning approaches for early diabetes risk evaluation and clinical decision support.

Keywords: Clinical decision support systems, Data mining, Diabetes prediction, Predictive analytics, Random forest

Introduction

Diabetes is considered a worldwide public-health problem by the World Health Organization (WHO). It affects as many as 9.3% of the global adult population, although elevating trends in low and middle-income areas have been noted. There is a small gender difference in the disorder (9.6% for men and 9.0% for women). Age is a continuous risk factor for ERLD, with the most significant incremental increase occurring in people over 35 years.[1] Diabetes leads to multiple lethal complications, including cardiovascular diseases, kidney disorders, and visual loss or limb amputation imposing major medical & socioeconomic burdens. Aged and with ongoing worldwide incidence, early diagnosis is key to avoid complications. The present study provides evidence in favour of early diagnoses, leading to a reduction in complications and a greater resolution of the general condition of patients. Early detection is also critical for the public health control of this disease.[2] Diabetes mellitus is a chronic metabolic disorder that is characterised by an increase in blood glucose concentration resulting from dysfunction of insulin production or inadequate responsiveness to insulin. The most common test for the diagnosis of diabetes is fasting blood glucose, GTT and HbA1c.[3] Associated factors are genetic predisposition, excess body weight, sedentary lifestyle and an unhealthy diet especially in those who have family members with the disease. It influences one's

*Authors for Correspondence
E-mail: a.kurhade@gmail.com, sinu.nambiar@mitwpu.edu.in

quality of daily life and is financially costly for the healthcare systems everywhere.[4]

This research contributes to a scientific and medical knowledge, too. Theoretically, it explores the potential ML (Random Forest [RF] model) to predict future risk of diabetes using clinical and demographic features. From a clinical perspective, this study may inform the efficiency of healthcare workers when can be able to use an accurate tool to early identify high-risk people and promote earlier diagnostic process as well as planning preventive care. Research methodology work is based on the KDD (Knowledge Discovery in Databases) process which is an ordered procedure for gaining more knowledge from big data sets. Through successive stages including data screening, cleaning, transforming, mining and interpretation, the investigation preserves a systematic and reliable method of model establishment which is fit for real clinical scenario. The aim is to develop a prediction model for diagnosis of Diabetes at an early stage with more accuracy and precision using Random Forest algorithm by learning how patient's data varies when they get diagnosed or vice versa.

Diabetes mellitus is a chronic metabolic disorder marked by persistent hyperglycaemia, leading to complications affecting the heart, kidneys, eyes, and peripheral nerves if not controlled. Pancreatic beta cells produce the hormone insulin which helps in uptake of glucose, disruption of this process is due to lack of insulin in Type I diabetes and reduced sensitivity to insulin at cellular level in Type II diabetes. There are two main types of diabetes mellitus: Type I diabetes is caused by autoimmune β-cell destruction, whereas type II diabetes develops because of insulin resistance combined with progressive beta-cell failure. Random Forest, a machine learning algorithm that integrates multiple classifiers and is based on decision trees using bagging, enhances prediction accuracy and deals with complex clinical information well.[5–7]

A study using 1,416 diabetic patients achieved 100% test accuracy for diabetic retinopathy and 76% accuracy for VTDR prediction.[8] Random Forest also reached 99.84% accuracy on imbalanced data, 96.75% on balanced data, and a 99% AUC in another investigation.[9,10] Using a Kaggle dataset of 768 samples, RF and XGBoost improved their accuracies to 0.8612 and 0.8351 after applying PSO and Genetic Algorithm feature selection techniques.[11–14] In a cohort of 10,663 individuals aged 40–70, RF identified fasting glucose, cholesterol, and creatinine as key T2DM predictors.[15] Environmental influences on T2DM risk were confirmed in 14,829 AMIGO participants, with LASSO showing the best predictive performance.[16–18] A hybrid SMKSVM–Random Forest model achieved 73.37% accuracy for T2DM classification.[19] For women over 25, Range-based Random Forest models improved diagnostic accuracy.[20,21] Survival models using RF outperformed Cox regression for predicting progression from single to multiple autoantibody positivity and then to clinical T1D.[22–24] Six resampling approaches were tested on the PIMA dataset and these best yielded to the performance of XGBoost under SMOTE-Tomek and RF under SMOTEENN.[25–27] Combined RF-KNN model even made better early detection of diabetes.[28,29] Other research also emphasized on the model to identify important features, improve early prediction, assist in risk assessment and provide support for clinical decision making in diabetes studies.[30–35]

The primary contributions of this work are to develop clinically interpretable Random Forest lines based on a comprehensive KDD practice or process that includes data selection, preprocessing and transformation, mining and interpretation stages for early diabetes prediction. This work differs from many studies which aim primarily at optimized algorithms, by focusing on reproducibility and applicability of the model to clinical practice. The model evaluation also emphasizes specificity, MCC, and ROC–AUC more than others due to the reduction in false-positive predictions, which is crucial for diabetes screening in real life. Moreover, simplified decision-tree visualization and feature-importance analysis would enhance explainability and help clinicians to comprehend the role of important prediction-related clinical parameters immediately. Random Forest was chosen as it mostly handles complexity and feature interactions without the need to set parameters.

The novelty of this work is the construction of a novel, clinically interpretable Random Forest framework for early diabetes risk prediction guided by a structured Knowledge Discovery in Databases (KDD) methodology. In contrast to existing literature that primarily attempts to maximize accuracy, we focus on balanced clinical performance with specificity, MCC, ROC-AUC and recall analyses. In addition, the SAGE study offers further insights into

how to interpret model activity using decision tree model representations and feature importance analysis combined with a sensitivity analysis that reinforces transparency and a clearer understanding of applications in practice. Moreover, the research emphasizes reproducible data pre-processing, exhaustive evaluation, and applicability with future clinical decision-support systems.

Methodology

In this work, the Knowledge Discovery in Databases (KDD) methodology is appropriate to be employed since it is one of the oldest methodologies for data mining & machine learning. It facilitates the retrieval of informative data from large and diverse data sets, which renders it appropriate for data-driven investigations. KDD process is usually composed of the systematic stages as shown in Fig. 1. It is an inter-disciplinary process which combines statistical methods, machine learning techniques and exploratory data analysis to discover meaning structures of the data collected as depicted in the architecture in Fig. 2. The detailed dataset statistics is shown in Table 1. The information has been selected from a Kaggle dataset online. The records in the dataset include anonymous adult individuals with or without diabetes and are publicly available on the Kaggle Data Set. Full clinical and symptom-based features were included, while those with incomplete or inconsistent records were excluded. This corroborates the issue of reproducibility in terms of openness and transparent processing, while limited population aspects and lack of laboratory markers preclude wider clinical generalization.

This cohort was tested and extracted from a publicly available Kaggle diabetes dataset, which comprised 520 patients with corresponding 16 clinical and demographic features associated with both diabetic and non-diabetic conditions. The most important attributes were Polyuria, Polydipsia, Obesity, Weakness, and Age. To evaluate the models, the dataset was split into an 80:20 ratio into a training and testing set. The generalisability of the study was limited as the data were collected retrospectively, and there was a lack of diversity in terms of demographics and specific laboratory parameters. These results may impact generalizability and clinical translatability to more heterogeneous populations.

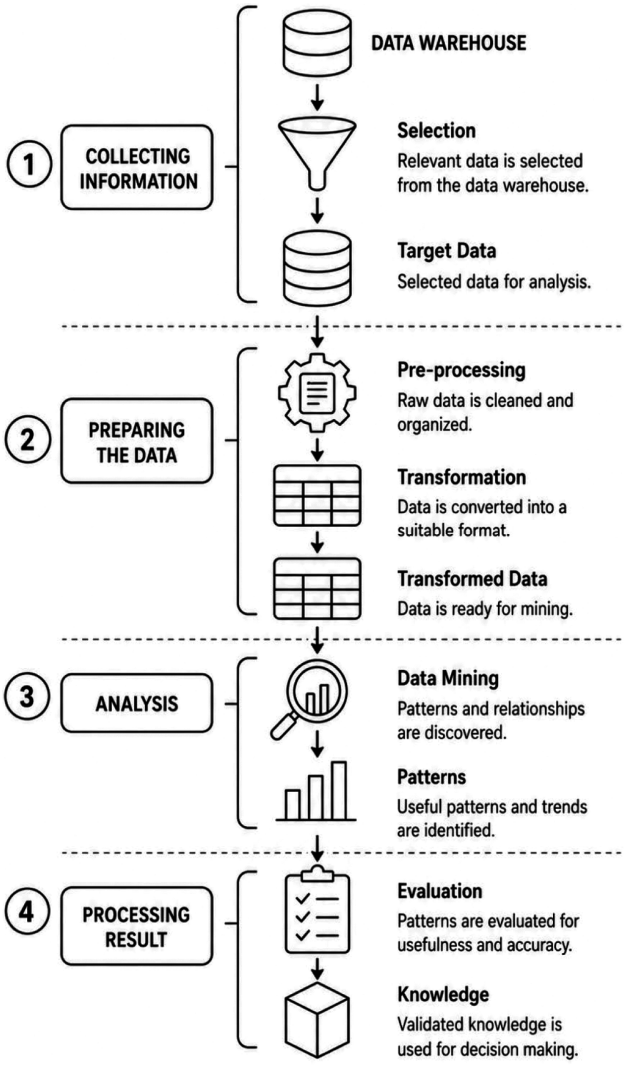

Fig. 1 — Knowledge discovery method

Preprocessing

Pre-processing commonly involves replacement of missing values and removal of outliers. In one advantage, the dataset did not include noise or missing entries that would affect behaviour of the model in Fig. 3.

Transformation

It is possible for the model to be trained on most of the data and then be tested on unseen records. The class names “Positive” and “Negative” are also mapped into numbers using 1 for Positive and 0 for Negative, to make them compatible with the algorithm, keeping all features in both sets. This preparation has explained in Fig. 4 and Fig. 5 shows some of the statistical properties that can be highlighted by looking at the distribution of numerical features.

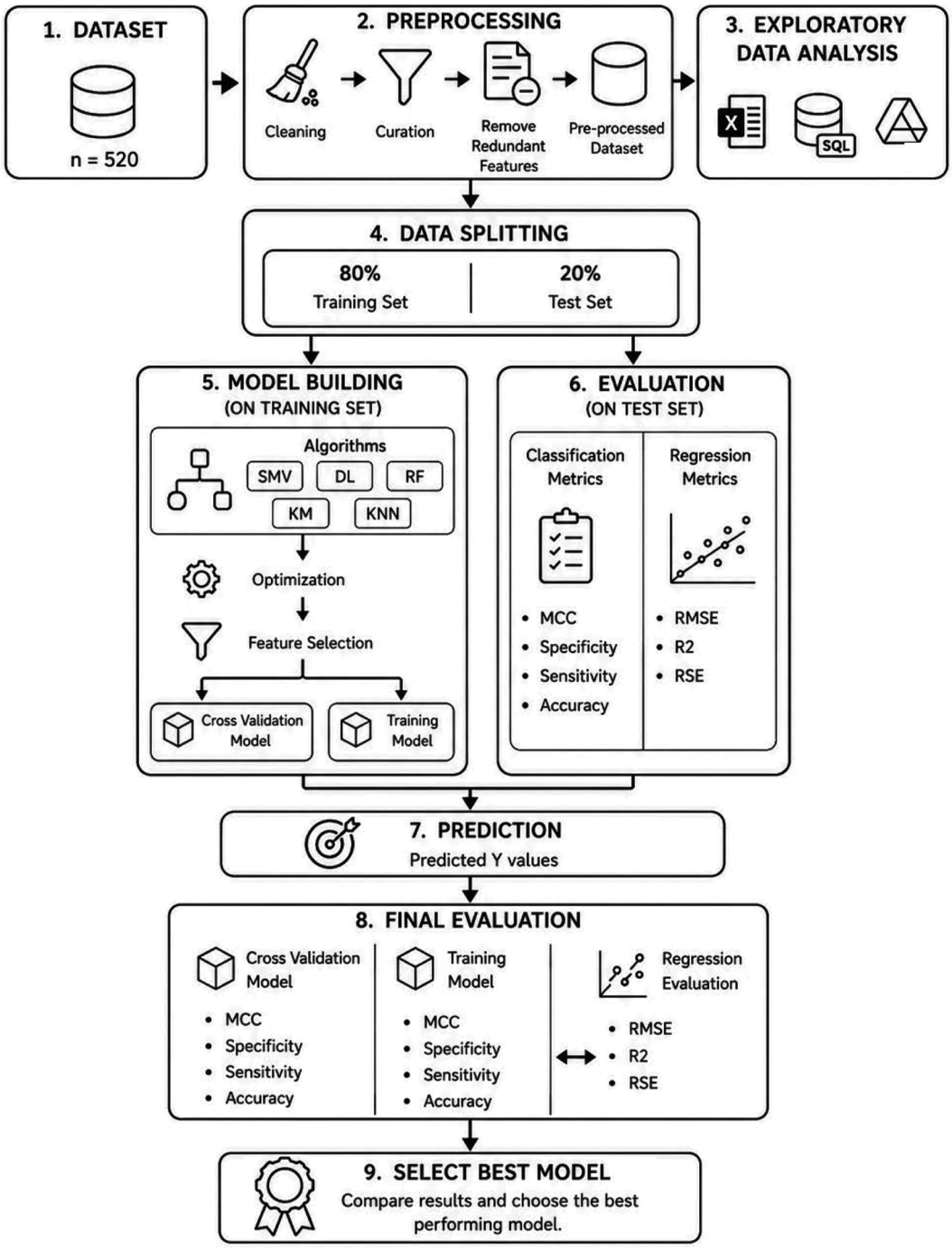

Fig. 2 — Machine learning architecture

Data Mining

The two visual outputs that help to explain the Random Forest model has explained in Fig. 6. The first is a single decision tree (restricted to depth 3), which shows how the model splits the data using influential features. It gives a clear view of the hierarchical decision rules used for classification. The second is a violin plot representing the distribution of feature importance values produced through simulation. This plot shows both the relative weight of each attribute and the variation in their importance. The dataset does not include potential institutional differences in clinical practices, laboratory measurement procedures and heterogeneity of the populations. Thus, the results demonstrate model performance under ideal data conditions rather than in clinical settings.

Table 1 — Clinical and demographic features used for diabetes prediction and their medical interpretation

Feature	Clinical relevance and medical interpretation
Age	Advancing age contributes to insulin resistance and decline in pancreatic β-cell activity
Gender	Hormonal and lifestyle differences influence diabetes prevalence
Polyuria	High blood glucose causes frequent urination through osmotic diuresis
Polydipsia	Increased thirst develops due to dehydration linked with polyuria
Sudden weight loss	Indicates impaired glucose utilization and tissue breakdown
Weakness	Fatigue caused by reduced cellular glucose uptake
Polyphagia	Increased hunger despite adequate intake, linked to insulin dysfunction
Genital thrush	Recurrent fungal infections due to elevated glucose and reduced immunity
Visual blurring	Temporary vision changes due to glucose-induced lens swelling
Itching	Skin irritation related to poor circulation and infections
Irritability	Mood changes due to glycaemic fluctuations
Delayed healing	Poor wound healing caused by microvascular damage
Partial paresis	Early neuromuscular impairment linked to diabetic neuropathy
Muscle stiffness	Associated with metabolic imbalance and musculoskeletal effects
Alopecia	Hair loss linked to impaired circulation and metabolic stress
Obesity	Major risk factor contributing to insulin resistance
Class	Outcome variable indicating diabetic or non-diabetic status

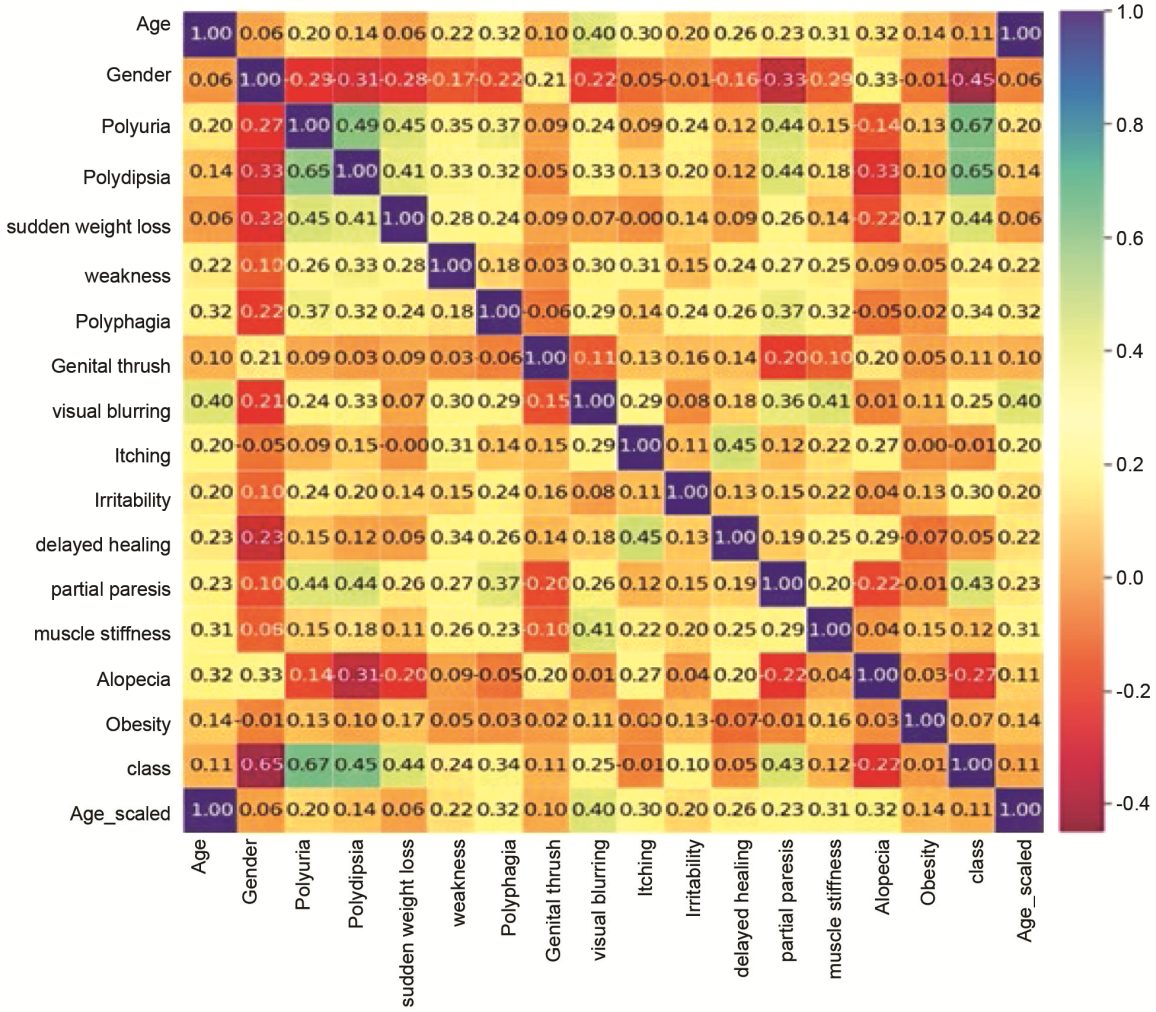

Fig. 3 — Data processing

Population bias is partly mitigated by utilizing a public Kaggle dataset of only adult participants and discarding incomplete records to preserve data consistency for the analysis. Model generalisability is indirectly supported by lower performance variability across multiple metrics, implying it can accommodate diversity in clinical characteristics in the same dataset. However, no subgroup analysis or external validation is conducted, making application to other demographic/clinical groups not directly demonstrated and a study limitation.

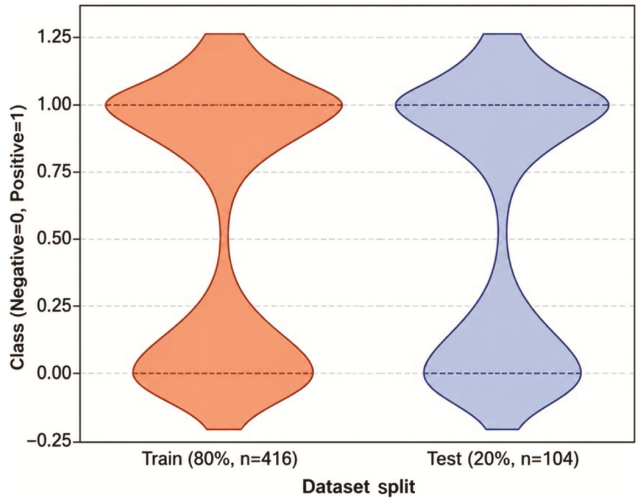

Fig. 4 — Class distribution across training and testing data partitions

Random Forest Mathematical Formulas

The Random Forest algorithm can be mathematically described as an ensemble of decision trees built on bootstrap sampled data points, with random selection of features. Each tree is fitted on a bootstrap sample of the original dataset, which helps reducing variance and increasing generalization. At each split, a random sample of features are considered, and the best split is chosen by minimising some impurity measure, such as Gini impurity or entropy. In classification, each tree makes a class prediction or gives the probability of assigned to each class and the overall Random Forest output is selected by majority voting or estimating average probabilities over all trees. This rule, inspired on the decision of a ensemble-based method classifier, allows nonlinear relations among variables to be modelled and for

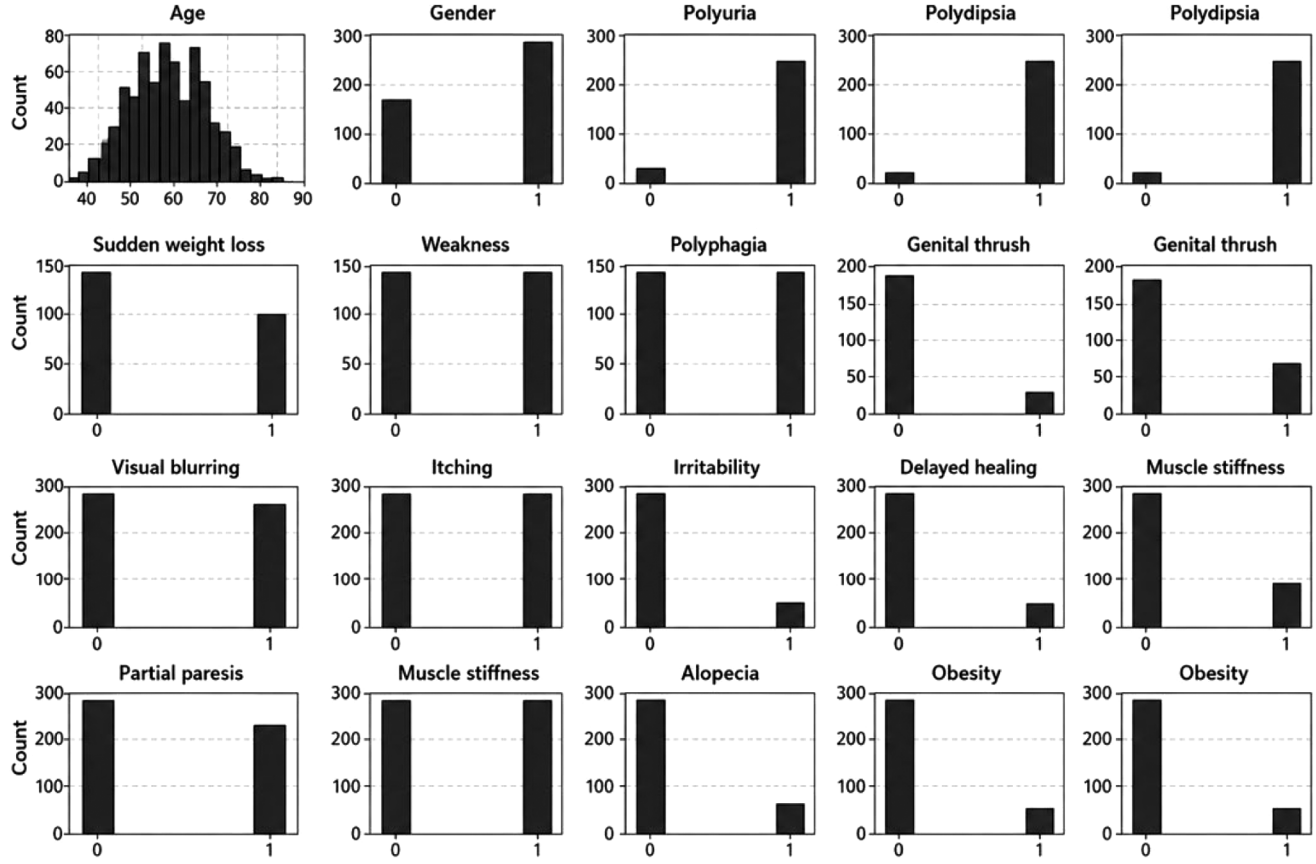

Fig. 5 — Distribution of numerical variables

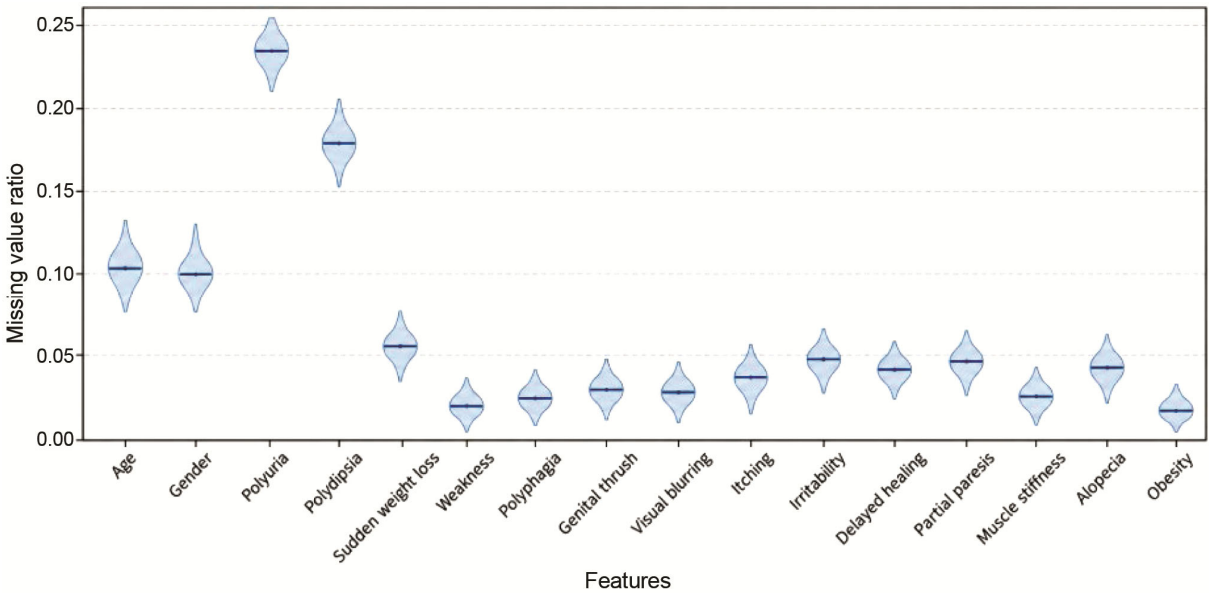

Fig. 6 — Artistic visualization of importance- violin plot

overfitting reduction while still leading to stable predictive performance across diverse datasets.

Bootstrap Sampling (Bagging)

$$D_t = \text{Bootstrap}(D),\ t = 1, 2, \ldots, T \quad \ldots(1)$$

where, D is the original dataset, D_t is the bootstrapped sample for tree t, T is the number of trees, and sampling is done with replacement. This step creates multiple varied subsets that reduce variance.

Random Feature Selection at Each Split

$$\theta_t = \text{argmin}_{\{\theta \in m\}} L(D_t, \theta) \quad \ldots(2)$$

where, m is the random subset of features, θ is the candidate split, L represents the impurity measure, and D_t is the data at the node. This reduces correlation among trees.

Gini Impurity

$$G(p) = 1 - \Sigma\, p_k^2 \quad \ldots (3)$$

where, p_k is the fraction of samples belonging to class k. Gini measures node impurity.

Gini-Based Split

$$G_{split} = (N_L/N)\, G_L + (N_R/N)\, G_R \quad \ldots (4)$$

where, N is parent node samples, N_L and N_R are child node samples, G_L and G_R are their impurities.

Entropy

$$H(p) = -\Sigma\, p_k \log_2(p_k) \quad \ldots (5)$$

where, p_k is the probability of class k. Entropy expresses uncertainty.

Information Gain

$$IG = H_{parent} - [\, (N_L/N)\, H_L + (N_R/N)\, H_R \,] \quad \ldots (6)$$

where, H_{parent} is parent entropy, H_L and H_R are child entropies.

Decision Function of a Single Tree

$$h_t(x) = \text{argmax}_k\, P(y=k \mid x, \text{tree } t) \quad \ldots (7)$$

where, $h_t(x)$ is class predicted by the t-th tree.

Random Forest Final Prediction (Majority Voting)

$$\hat{y} = \text{mode}\{\, h_1(x), h_2(x), \ldots, h_T(x) \,\} \quad \ldots (8)$$

where, $\hat{y}$ is the final predicted class.

Probability Averaging

$$P(y=k \mid x) = (1/T)\, \Sigma\, P_t(y=k \mid x) \quad \ldots (9)$$

where, P_t is the probability from tree t.

OOB Prediction

$$\hat{y}_i^{OOB} = \text{mode}\{\, h_t(x_i) : t \in B_i \,\} \quad \ldots (10)$$

where, B_i is the set of trees that did not see sample i.

OOB Error

$$\text{OOB Error} = (1/N)\, \Sigma\, I(\hat{y}_i^{OOB} \neq y_i) \quad \ldots (11)$$

where, I() is an indicator function.

Feature Importance $FI_j = (1/T) \Sigma \Sigma \Delta G_{t,s}$ … (12)

where, FI_j is importance of feature j, $\Delta G_{t,s}$ is impurity reduction.

Impurity Reduction

$\Delta G = G_{parent} - [(N_L/N)G_L + (N_R/N)G_R]$ … (13)

where, G_{parent} is impurity before split.

Accuracy

$Accuracy = (TP + TN) / (TP + TN + FP + FN)$ … (14)

where, TP, TN, FP, FN are confusion matrix components.

Precision

$Precision = TP / (TP + FP)$ … (15)

where, TP is true positive, FP is false positive.

Recall

$Recall = TP / (TP + FN)$ … (16)

where, FN is false negative.

Specificity

$Specificity = TN / (TN + FP)$ … (17)

where, TN is true negative.

MCC

$MCC = (TP \cdot TN - FP \cdot FN) / \sqrt{((TP + FP)(TP + FN)(TN + FP)(TN + FN))}$ … (18)

where, MCC incorporates all confusion matrix elements.

The Random Forest model was systematically optimized over hyperparameter space for improved predictive performance and reproducibility. Some techniques used here are Grid Search, Random Search & Bayesian Optimization which helps to find the best combination of parameters including number of trees, max. depth, min. samples split and feature selection strategy. In this study, the Random Forest hyperparameters were configured based on previously tested experimentally validated values to include 100 trees, Gini impurity criterion and bootstrap sampling which provided balanced accuracy and stable classification performance for early prediction of diabetes.

The hyper parameter settings in Table 2 correspond to how the Random Forest model was instantiated for both experiments. The number of trees in the ensemble was set to 100, such that predictions were stable, yet unexpectedly high computational requirements remained low. The depth of the tree was not specified, so that nonlinear relationships among clinical features could be learned. The criterion for the Gini impurity and bootstrap sampling of the features with sqrt {p}feature selection strategy were applied to decrease correlation among trees, as well as to improve generalization. Additionally, no class weight was used to bring the model more in line with our focus on high specificity early screening rather than overly aggressive sensitivity tuning.

Table 2 — Summary of Random Forest hyperparameters used in the study

Hyperparameter	Final value used
Number of trees (T)	100
Maximum tree depth	Not restricted
Minimum samples to split	2
Minimum samples per leaf	1
Feature selection strategy	√p (p = total features)
Split criterion	Gini impurity
Bootstrap sampling	Enabled
Class weighting	Not applied

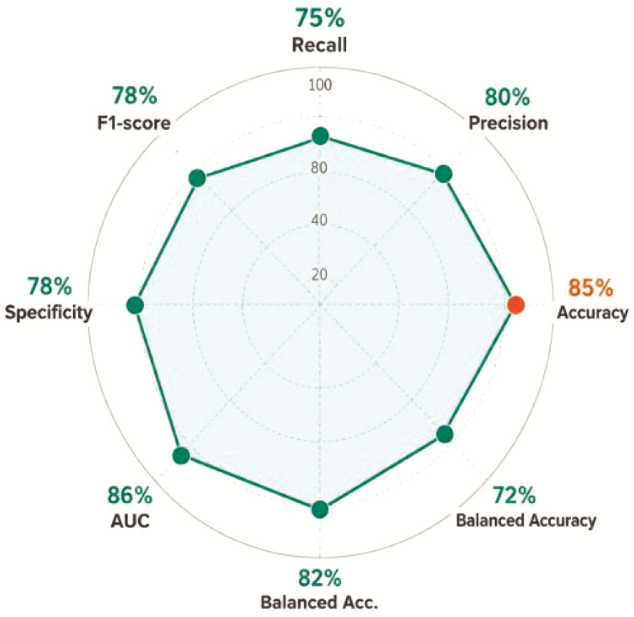

Fig. 7 — Performance metrics analysis

Results

The radar plot shown in Fig. 7 explains the overall performance of classification model on eight measurement considerations. Higher values are seen

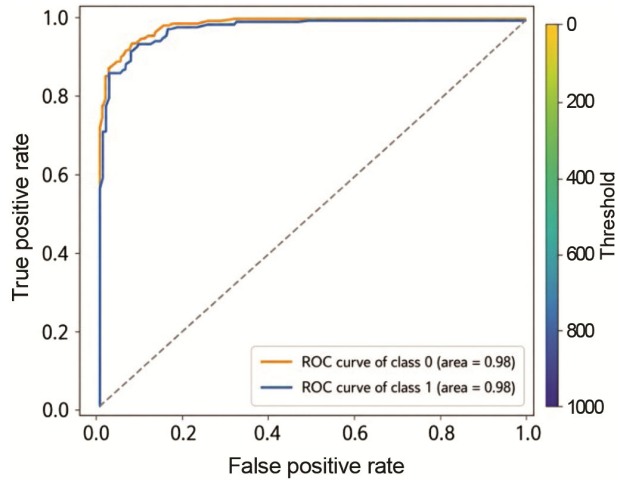

Fig. 8 — ROC curve analysis

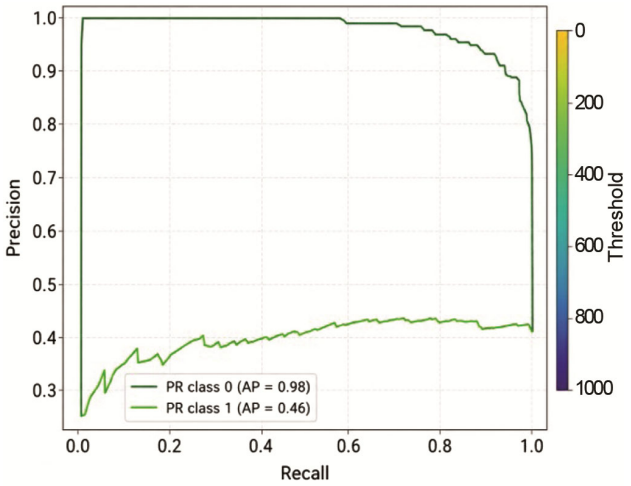

Fig. 9 — Precision - recall curve

for specificity (78%), AUC (86%) and accuracy (85%), which reflects a good capability to correctly detect negative samples, and stabilize its overall performance. Metrics like recall and MCC were relatively lower (75% and 72%), which however reflected the poor performance to detect all positives. This implies that false negatives exist, contributing to the decrease of sensitivity for the model. Overall, the performance indicators indicate a good overall trait balance of the model; however, sensitivity needs to be improved and agreement with actual class labels boosted. The ROC curves in Fig. 8 represent performance of the Random Forest classifier class-wise. The ROC–AUC for both class 0 and class 1 is up to 0.98, with a strong discriminative power to distinguish between the two categories. The random prediction performance is indicated in the diagonal reference line. Since the ROC curve for both classes is far above this line, the classifier does much better than random guessing. The provided colour-bar is only meant to ease the visual interpretation and not affecting the results.

This conclusion is echoed by the precision–recall (PR) curve in Fig. 9, which indicates the model performance regarding positive class. The AP of class 0 was 0.98 for the model, and a simulated curve in another area had an AP of 0.46 as comparison value. The curves show that the model retains its precision across high levels of recall, meaning it can identify positive cases while producing only a few false-positive predictions. PR curves are particularly useful when class imbalance exists, as they directly display the model's strength in capturing positive cases. The bar on the right serves only an aesthetic purpose and does not influence the interpretation. The combined evidence confirms that the RF model performs reliably in terms of precision and recall.

The relatively low recall reported in the proposed model is largely due to the class distribution of the dataset and symptom-oriented characterisation of our dataset. As most instances are non-diabetic, the Random Forest classifier, when minimising impurity values tends to try and get the larger class right in which case it has higher specificity. Thus, during the search operation, a diabetic patient who presents with mild or incomplete symptoms are classified in error, which results low recall. Class-imbalance measures such as re-sampling or cost-sensitive learning were not used, since this study presents the model as an early-stage screening and decision-assisted tool with greater emphasis on minimizing false-positive predictions and unnecessary further investigations. Although the achieved accuracy, specificity and AUC values demonstrate reliable predictive ability in the tested dataset overall, such scores alone cannot warrant a real-time clinical application. Validation on external sets of hospital or region-specific samples is required to ensure robustness and clinical applicability.

A Random Forest model yields high ROC–AUC values, indicating overall strong discriminative power over a range of classification threshold settings. Simultaneously, the moderate recall and specificity scores imply that this is a general behavior of the model at the operating threshold measured during evaluation. ROC–AUC evaluates the global discrimination of the classifier, whereas recall and

Table 3 — Systematic comparison of random forest with other machine learning algorithms

Algorithm	Working principle	Strengths	Limitations	Suitability for diabetes prediction	Comparative observation with Random Forest
Logistic Regression (LR)	Statistical classification model based on probability estimation using a sigmoid function	Simple, interpretable, computationally efficient, suitable for linear relationships	Limited ability to capture nonlinear interactions and complex feature dependencies	Effective for baseline diabetes classification using structured clinical data	Random Forest generally provides higher predictive accuracy and better handling of nonlinear medical features compared to LR
Support Vector Machine (SVM)	Separates classes using optimal hyperplanes with kernel functions	High classification capability in high-dimensional datasets, effective with nonlinear kernels	Computationally expensive for large datasets, sensitive to parameter tuning	Useful for diabetes datasets with complex boundaries	Random Forest requires less parameter tuning and offers better interpretability for clinical applications
Decision Tree (DT)	Recursive tree-based splitting using impurity measures such as Gini index or entropy	Easy visualization, transparent decision-making, interpretable	High risk of overfitting and unstable performance with noisy data	Suitable for understanding clinical decision rules	Random Forest improves stability and accuracy by combining multiple decision trees through ensemble learning
Gradient Boosting (GB)	Sequential ensemble learning where weak learners correct previous errors	High predictive performance, strong handling of complex patterns	Longer training time, sensitive to hyperparameters, risk of overfitting	Effective for high-accuracy diabetes prediction tasks	Random Forest provides more balanced performance with lower computational complexity and reduced overfitting risk
Random Forest (RF)	Ensemble of multiple decision trees using bagging and random feature selection	Robust against overfitting, handles nonlinear relationships, stable performance, supports feature importance analysis	Reduced interpretability compared to a single tree, larger computational cost than LR	Highly suitable for clinical risk prediction and structured healthcare datasets	Demonstrated balanced accuracy, specificity, AUC, and robustness for early diabetes risk assessment

specificity correspond to a threshold-dependent performance metric in clinical prediction problems since they reflect binary confusion. This relatively lower specificity correlates with a moderate recall, the latter confirms low number of actual diabetic cases missed by the model whilst still picking at least 50% additional false-positive predictions. This is clinically significant in early diabetes screening as it helps to achieve minimization with equal performance of high-risk and low-risk subjects; minimize the unsatisfactory proportionality for false alarms. A comparison between various machine learning algorithms has been explained in Table 3**.**

A limitation is with comparisons against baseline classifiers on the same data partition and evaluation metric. But we are unable to quantitatively show the relative gain of Random Forest over simpler classifiers without these comparisons in the same conditions. The selection of Random Forest is motivated in part by previous findings, as well as its consistent performance on all our evaluation measures. We will expand our evaluation in future works by comparing Random Forest with the baseline classifiers (Logistic Regression, Decision Tree, and KNN) on the same dataset, pre-processing steps and validation protocol to have a complete view about model choice. The main restriction of this study is lack of hospital validation. Because this is a model that was trained and tested on retrospective public data, its utility in real clinical decision support still needs to be established via prospective studies and population-specific validation.

Conclusions

This study provides a new data-driven methodology for predicting the risk of diabetes at an early stage using the Random Forest algorithm as part of KDD. The model was created based on a publicly available Kaggle dataset with several clinical and demographic features related to diabetes. Note this is a screening and clinical decision-assistance system, not a standalone diagnostic model. The Random Forest classifier demonstrated balanced predictive performance with an overall accuracy of 85%, specificity of 78%, recall (sensitivity) of 75%, area under the ROC curve (AUC) of 86% and Matthews correlation coefficient (MCC) score of 72% in predicting diabetes against non-diabetes. The increase in specificity shows a decrease in the number of false positives, which is crucial for those applications that

require large-scale screening. This framework could potentially be used to detect risk at an early stage, develop targeted screening programs and plan healthcare. The early results were promising, but this study has the limitation of using a single retrospective dataset. In the future, larger and more diversified clinical datasets must be validated in real-time healthcare systems and explainable AI frameworks to provide transparency and clinical applicability.

References

1 Shetty S & Joshi S, A tool for diabetes prediction and monitoring using data mining technique, *Int J Inf Technol Comput Sci*, **8(11)** (2016) 26, doi.org/10.5815/ ijitcs.2016.11.04.

2 Rashid T A, Saman S & Rezhna, Decision support system for diabetes mellitus through machine learning techniques, *Int J Adv Comput Sci Appl*, **7(7)** (2016), doi.org/10.14569/ ijacsa.2016.070724.

3 Sa'di S, Maleki A, Hashemi R, Panbechi Z & Chalabi K, Comparison of data mining algorithms in the diagnosis of type II diabetes, *Int J Comput Sci Appl*, **5(5)** (2015) 1, doi.org/10.5121/ijcsa.2015.5501.

4 Zheng L, Wang Y, Hao S, Shin A Y, Jin B, Ngo A, Jackson-Browne MS, Feller D, Fu T, Zhang K, Zhou X, Zhu C, Dai D, Yu Y, Gang Z, Li Y, McElhinney D B, Culver D S, Alfreds S T, Stearns F, Sylvester K G, Widen E & Ling X B, Web-based real-time case finding for the population health management of patients with diabetes mellitus: A prospective validation of the natural language processing–based algorithm with statewide electronic medical records, *JMIR Med Inform*, **4(4)** (2016), doi.org/10.2196/ medinform.6328.

5 Casanova R, Saldana S, Simpson S L, Lacy M E, Subauste A, Blackshear C, Wagenknecht L E & Bertoni A G, Prediction of incident diabetes in the Jackson Heart Study using high-dimensional machine learning, *PLoS One*, **11(10)** (2016), doi.org/10.1371/journal.pone.0163942.

6 Daghistani T & Alshammari R, Diagnosis of diabetes by applying data mining classification techniques, *Int J Adv Comput Sci Appl*, **7(7)** (2016), doi.org/10.14569/ ijacsa.2016.070747.

7 Anderson A, Kerr W T, Thames A D, Li T, Xiao J & Cohen M S, Electronic health record phenotyping improves detection and screening of type 2 diabetes in the general United States population: A cross-sectional, unselected, retrospective study, *arXiv*, 2015, http://arxiv.org/abs /1501.02402.

8 Zheng T, Xie W, Xu L, He X, Zhang Y, You M, Yang G & Chen Y, A machine learning-based framework to identify type 2 diabetes through electronic health records, *Int J Med Inform*, **97** (2016) 120, doi.org/10.1016/ j.ijmedinf. 2016.09.014.

9 Seyhan A A, López Y, Xie H, Yi F, Mathews C E, Pasarica M & Pratley R E, Pancreas-enriched miRNAs are altered in the circulation of subjects with diabetes: A pilot cross-sectional study, *Sci Rep*, **6(1)** (2016), doi.org/10.1038/ srep31479.

10 Nagarajan S & Chandrasekaran R, Design and implementation of expert clinical system for diagnosing diabetes using data mining techniques, *Indian J Sci Technol*, **8(8)** (2015) 771, doi.org/10.17485/ijst/2015/v8i8/69272.

11 Gill N S & Mittal P, A novel hybrid model for diabetic prediction using hidden Markov model, fuzzy based rule approach and neural network, *Indian J Sci Technol*, **9(35)** (2016), doi.org/10.17485/ijst/2016/ v9i35/99146.

12 Kälsch J, Bechmann L P, Heider D, Best J, Manka P, Kälsch H, Sowa J, Moebus S, Slomiany U, Jöckel K, Erbel R, Gerken G & Canbay A, Normal liver enzymes are correlated with severity of metabolic syndrome in a large population based cohort, *Sci Rep*, **5(1)** (2015), doi.org/10.1038/ srep13058.

13 Claesen M, Smet F D, Gillard P, Mathieu C & Moor B D, Building classifiers to predict the start of glucose-lowering pharmacotherapy using Belgian health expenditure data, *arXiv*, 2015, http://arxiv.org/abs/1504.07389.

14 Coffman E & Richmond-Bryant J, Multiple biomarker models for improved risk estimation of specific cardiovascular diseases related to metabolic syndrome: A cross-sectional study, *Popul Health Metr*, **13(1)** (2015), doi.org/10.1186/s12963-015-0041-5.

15 Ruiz E C, García-Sáez G, Rigla M, Villaplana M, Pons B & Hernando M E, Automatic classification of glycaemia measurements to enhance data interpretation in an expert system for gestational diabetes, *Expert Syst Appl*, **63** (2016) 386, doi.org/10.1016/j.eswa.2016.07.019.

16 Golubnitschaja O, Debald M, Kühn W, Yeghiazaryan K, Bubnov R, Goncharenko V M, Lushchyk U B, Grech G & Konieczka K, Flammer syndrome and potential formation of pre-metastatic niches: A multi-centred study on phenotyping, patient stratification, prediction and potential prevention of aggressive breast cancer and metastatic disease, 2016, doi.org/10.1186/s13167-016-0054-6.

17 Hu X, Reaven P D, Saremi A, Liu N, Abbasi M A, Liu H & Migrino R Q, Machine learning to predict rapid progression of carotid atherosclerosis in patients with impaired glucose tolerance, *EURASIP J Bioinform Syst Biol*, **1** (2016), doi.org/10.1186/s13637-016-0049-6.

18 Cichosz S L, A prediction model for insulin associated weight gain in patients with type 2 diabetes, *Europe PMC*, **31** (2016), http://europepmc.org/articles/pmc7088195.

19 Daskalaki E, Diem P & Mougiakakou S, Model-free machine learning in biomedicine: Feasibility study in type 1 diabetes, *PLoS One*, **11(7)** (2016), doi.org/10.1371/journal.pone. 0158722.

20 Brahim N B, Integrated system for an automated insulin therapy for type 1 diabetes: Real time evaluation of the effect of physical exercise and adjustment of the insulin dosage, HAL, 2016, https://tel.archives-ouvertes.fr/tel-01496776.

21 Francescato M P, Stel G, Stenner E & Geat M, Prolonged exercise in type 1 diabetes: Performance of a customizable algorithm to estimate the carbohydrate supplements to minimize glycemic imbalances, *PLoS One*, **10(4)** (2015), doi.org/10.1371/journal.pone.0125220.

22 Zarkogianni K, Litsa E, Mitsis K, Wu P Y, Kaddi C, Cheng C, Wang M D & Nikita K S, A review of emerging technologies for the management of diabetes mellitus, *IEEE Trans Biomed Eng*, **62(12)** (2015), 2735, doi.org/10.1109/ TBME.2015.2470521.

23 Hochberg I, Feraru G, Kozdoba M, Mannor S, Tennenholtz M & Yom-Tov E, A reinforcement learning system to

encourage physical activity in diabetes patients, *arXiv*, 2016, doi.org/10.2196/jmir.7994.

24 Wang Y, Wu P, Liu Y, Weng C & Zeng D, Learning optimal individualized treatment rules from electronic health record data, **65** (2016), doi.org/10.1109/ichi.2016.13.

25 Kleinberger J W & Pollin T I, Personalized medicine in diabetes mellitus: Current opportunities and future prospects, *Ann N Y Acad Sci*, **1346(1)** (2015) 45, doi.org/10.1111/nyas.12757.

26 Böhm A, Weigert C, Staiger H & Häring H, Exercise and diabetes: Relevance and causes for response variability, *Endocrine*, **51(3)** (2015) 390, doi.org/10.1007/s12020-015-0792-6.

27 Agboola S, Jethwani K, López L, Searl M, O'Keefe S M & Kvedar J C, Text to move: A randomized controlled trial of a text-messaging program to improve physical activity behaviors in patients with type 2 diabetes mellitus, *J Med Internet Res*, **18(11)** (2016), doi.org/10.2196/jmir.6439.

28 Dijk J van & Loon L, Exercise strategies to optimize glycemic control in type 2 diabetes: A continuing glucose monitoring perspective, Diabetes *Spectr*, **28(1)** (2015) 24, doi.org/10.2337/diaspect.28.1.24.

29 Kleinberger J W, Maloney K A & Pollin T I, The genetic architecture of diabetes in pregnancy: Implications for clinical practice, *Am J Perinatol*, **33(13)** (2016), 1319, doi.org/10.1055/s-0036-1592078.

30 Sridhar G, Diabetes and data in many forms, *Int J Diabetes Dev Ctries*, **36(4)** (2016) 381, doi.org/10.1007/s13410-016-0540-3.

31 Willemsen G, Ward K, Bell C G, Christensen K, Bowden J L, Dalgård C, Harris J R, Kaprio J, Lyle R, Magnusson P K E, Mather K A, Ordoñana J R, Pérez-Riquelme F, Pedersen N L, Pietiläinen K H, Sachdev P S, Boomsma DI & Spector T D, The concordance and heritability of type 2 diabetes in 34,166 twin pairs from international twin registers: The discordant twin (DISCOTWIN) consortium, *Twin Res Hum Genet*, **18(6)** (2015) 762, doi.org/10.1017/thg.2015.83.

32 Feinberg A P & Fallin M D, Epigenetics at the crossroads of genes and the environment, *JAMA*, **314(11)** (2015), 1129, doi.org/10.1001/jama.2015.10414.

33 Choubey D K & Paul S, G A_MLP N N: A hybrid intelligent system for diabetes disease diagnosis, *Int J Intell Syst Appl*, **8(1)** (2016) 49, doi.org/10.5815/ijisa.2016.01.06.

34 Agarwal A, Baechle C, Behara R S & Rao V, Multi-method approach to wellness predictive modeling, *J Big Data*, **3(1)** (2016), doi.org/10.1186/s40537-016-0049-0.

35 Iyer A, Jeyalatha S & Sumbaly R, Diagnosis of diabetes using classification mining techniques, *Int J Data Min Knowl Manag Process*, **5(1)** (2015) 1, doi.org/10.5121/ijdkp.2015.5101.