Med & Health Aug 2024; 19(2): 380-399 https://doi.org/10.17576/MH.2024.1902.03

REVIEW ARTICLE

A Scoping Review of Supervised Machine Learning Techniques in Predicting the Prevalence of Type 2 Diabetes Mellitus

MOHD RIZAL MF[1], ABDUL MAULUD KN[1], GANASEGERAN K[2,3], ABDUL MANAF MR[2], SAFIAN N[2], MUSTAPHA FI[4], WALLER LA[5]

[1]Earth Observation Centre, Institute of Climate Change, Universiti Kebangsaan Malaysia, 43600 Selangor Darul Ehsan, Malaysia
[2]Department of Public Health Medicine, Faculty of Medicine, Universiti Kebangsaan Malaysia, 56000 Kuala Lumpur, Malaysia
[3]Occupational Health and Safety Unit, Seberang Jaya Hospital, Ministry of Health Malaysia, 13700 Seberang Perai, Penang, Malaysia
[4]Perak State Health Department, Ministry of Health Malaysia, 30000 Perak, Malaysia
[5]Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA 30322, USA

Received: 05 Mar 2024 / Accepted: 03 Apr 2024

ABSTRAK

Adalah penting bagi pihak perubatan untuk mengemudi perawatan daripada bergantung semata-mata pada pendekatan analisis data konvensional untuk saringan penyakit, diagnostik dan rawatan kepada keputusan yang dikonfigurasikan dengan pantas melalui analisis data besar daripada algoritma kecerdasan buatan (AI). Seni bina pengkomputeran "kabus" dan "tepi" yang dibina dalam sistem pangkalan data penjagaan kesihatan dan pengawasan yang besar membolehkan aplikasi algoritma pembelajaran mesin (ML) untuk ramalan penyakit dan kapasiti ramalan. Semakan tinjauan ini menilai penggunaan pelbagai kaedah ML untuk meramal T2DM. Enjin carian yang digunakan ialah IEEE Xplore, JSTOR, PubMed, Sage, Scopus, Wiley dan WOS. Kriteria kemasukan termasuk artikel yang diterbitkan dalam tempoh enam tahun lalu, akses terbuka dan kajian memfokuskan pada Diabetis Melitus Jenis 2 (T2DM) sahaja. Daripada 41 kajian, kaedah ML yang paling banyak digunakan ialah Random Forest (n=33) dan juga model ML terbaik yang paling banyak dijalankan (n=13). Kaedah ML Customised Ensemble yang disesuaikan dengan set data didapati menunjukkan ketepatan tertinggi. Walau bagaimanapun, kawasan kajian dan sampel tidak mencukupi di negara Asia Tenggara, kerana terdapat perbezaan

Address for correspondence and reprint requests: Associate Professor Dr. Khairul Nizam Abdul Maulud. Earth Observation Centre, Institute of Climate Change, Universiti Kebangsaan Malaysia, 43600 Selangor Darul Ehsan, Malaysia. Tel: +603-8921 6767 Email: knam@ukm.edu.my

dalam demografi dan budaya yang mempengaruhi faktor risiko T2DM di mana sumber pengiraan dan pembangunan sistem adalah terhad. Kami menyimpulkan kaedah ML mampu meramal T2DM, dari perspektif sistem kebolehoperasian intranya yang berdaya maju untuk digunakan dalam sistem penjagaan kesihatan.

Kata kunci: Diabetis melitus jenis 2; pembelajaran mesin; ramalan

ABSTRACT

It is crucial for medical practice to navigate from solely dependent on conventional data-analytical approaches for disease screening, diagnostics, and treatment plans to decisions that are configured rapidly through big data analytics from artificial intelligence algorithms. The fog- and edge-computing architectures built within the huge healthcare database systems would allow the applications of machine learning (ML) algorithms for disease predictions and forecasting capacities. This scoping review appraised the use of multiple ML methods for type 2 diabetes mellitus (T2DM) prediction. Search engines used were IEEE Xplore, JSTOR, PubMed, Sage, Scopus, Wiley, and WOS. Inclusion criteria included articles published within the past six years, open access and studies that focused on T2DM only. Out of 41 studies included, the most used ML method was Random Forest (n=33) and the most occurred best ML model (n=13). Customised Ensemble ML method adapted to the dataset was found to show the highest accuracy. However, there were insufficient study areas and samples in Southeast Asia countries, as there were differences in demographics and culture that affect the T2DM risk factors where computational resource and systems development were limited. We conclude ML methods can predict T2DM, from the system's perspective its intra-operability is viable for use in healthcare systems.

Keywords: Prediction; supervised machine learning; type 2 diabetes mellitus

INTRODUCTION

Diabetes is a key non-communicable disease (NCD) to populations worldwide, posing global socioeconomic and healthcare burden. Global estimates reported approximately 463 million adults aged 20 to 79 years old being afflicted with diabetes as of 2019, and this number is projected to rise to 700 million by 2045 (Galicia-Garcia et al. 2020). Type 2 diabetes mellitus (T2DM) is a condition of insulin insufficiency, and accounts for 90-95% of all diabetes cases (Tripathi & Srivastava 2006). Interventions to control the burden and complications of T2DM requires on-going health consultations, frequent blood glucose and podiatry monitoring, complex medication regimes, and lifestyle modifications;

all of which pose greater burden to the country's healthcare cost and the national budget (Rask-Madsen & King 2013). When populations are at an increased risk to NCDs many national health systems would struggle to configure appropriate interventions to control the burden. However, crafting and rapidly executing appropriate policies or strategies would be difficult if surveillance data on disease burden are too limited, complex, or chaotic to execute systematic analysis in understanding the trends and pattern of the disease occurrence.

The rise of Big Data within health systems has allowed scientists to apply computational methods and machine learning (ML) algorithms for time series analytics for synthesis of disease prediction, nowcasting, and forecasting. ML algorithms are used for analysing complex datasets, extracting patterns, and generating accurate predictions. In the context of predicting T2DM, ML has been proven to assist scientists and clinicians in shortening the time for analysis and pattern prediction compared to traditional statistical approaches (Ngiam & Khor 2019). In recent years, ML techniques have been applied extensively in epidemiology and public health research to predict various disease outcomes including T2DM. ML methods offer a promising approach to integrating environmental factors and uncovering their predictive power. By leveraging advanced algorithms such as Decision Trees (DT), Random Forests (RF), Support Vector Machines (SVM), Artificial Neural Network (ANN), k-nearest neighbors (KNN), Naive Bayes (NB), Gradient Boosting (GB) and Logistic Regression (LR), ML models can capture complex interactions between various risks factors with T2DM outcomes.

Supervised Machine Learning

Three primary categories of ML include supervised ML, unsupervised ML and reinforcement learning. In this review, supervised ML is the primary type of ML used because it well suits two problems of interest which are regression and classification problems. These two approaches often form the root of prediction models. In a supervised ML, a labelled dataset undergoes training and testing phases to determine the performance of the ML technique deployed. Using the trained dataset, it is then fed into a real-world dataset for prediction purposes. The training dataset has a "Diabetic" and "No Diabetic" classifier that separates the "True" and "False" label. The "True" and "False" label is crucial to determine the confusion matrix used to evaluate the ML performance. A confusion matrix is a table summarising the performance of a classification model. It is a useful tool for understanding how well the model can distinguish between different classes. The confusion matrix is a square table with two dimensions, representing the actual and predicted classes. The rows of the table represent the actual classes, while the columns represent the predicted classes. The table is divided into four quadrants: True Positive (TP) - The model correctly predicts that the sample belongs to

the positive class; True Negative (TN) - The model correctly predicts that the sample belongs to the negative class; False Positive (FP) - The model incorrectly predicts that the sample belongs to the positive class; False Negative (FN) - The model incorrectly predicts that the sample belongs to the negative class.

The confusion matrix can be used to calculate several performance metrics, such as accuracy, precision, recall, and F1 score. Accuracy is the percentage of samples that are classified correctly. It is calculated by dividing the sum of TP and TN by the total number of samples. The precision is the percentage of samples that are predicted to be positive that are actually positive. It is calculated by dividing TP by TP + FP. The recall is the percentage of samples that are actually positive that are predicted to be positive. It is calculated by dividing TP by TP + FN. The F1 score is a weighted average of precision and recall. It is calculated by dividing 2 * (precision * recall) by precision + recall.

The confusion matrix is a powerful tool for evaluating the performance of a ML model. It can be used to identify the types of errors that the model is making, and it can be used to improve the performance of the model. The advantage of using the confusion matrix based on metrics of performance is that they provide a simple and intuitive way to evaluate the performance of a classification model. The confusion matrix can be used to identify the types of errors that the model is making, can evaluate the performance of a model for multiple classes, and can be used to suggest improvements to the performance of the model.

The eight most used ML techniques are Random Forest, Gradient Boosting, Logistic Regression, Support Vector Machines, Decision Tree, Naive Bayes, Artificial Neural Network, and k-Nearest Neighbors. We briefly reviewed each ML and the comparison was shown in Table 1.

This scoping review aimed to comprehensively evaluate and synthesise existing literatures on the application of ML methods for predicting T2DM based on various risk factors. The review identified available ML methods commonly used for T2DM prediction. This review also aimed to evaluate the predictive accuracy of ML algorithms along with the risk factor when specifically applied to T2DM. Finally, we discussed gaps and limitations of this review to suggest improvements for future works.

MATERIALS AND METHODS

Scoping reviews aims to provide an overview of the existing literature, identify needs-gaps in knowledge, and understand the current landscape of works conducted on a particular topic. This scoping review followed the five-stage framework proposed by Arksey and O'Malley (2005). This scoping review was also conducted using Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) Checklist (Tricco et al. 2018).

TABLE 1: Comparison of machine learning techniques

Type of Model	Description	Advantage	Disadvantage
Random Forest (RF) (Wang et al. 2021; Brieman 2001)	An ensemble classifier that consists of many Decisions Tree to make predictions	Very accurate algorithm that is not sensitive to outliers and imperfect data.	Sensitive to the choice of hyperparameter tuning of the ML model and can be computationally expensive to train.
Gradient Boosting Variants	Gradient boosting machine which can be used for both regression and classification tasks.	Iteratively adding new trees to the model, each of which is trained to reduce residual errors of previous trees.	Computationally intensive, complexity, overfitting and sensitive to hyperparameters.
Logistic Regression (LR) (Joshi et al. 2021; Khanam & Foo 2021; Li et al. 2023)	Statistical model used to predict the probability of a binary outcome	Simple and easy-to-understand ML model that is relatively easy to fit to the data.	Only is used to predict binary outcomes and can be sensitive to outliers.
Support Vector Machine (SVM) (Abbas et al. 2019; Firdous et al. 2022; Joachims 1998; Khanam & Foo 2021)	Find a hyperplane that best separates the data points of people with T2DM and people without T2DM	Handle both linear and non-linear relationships between the features and the target variable and are also are not sensitive to outliers.	Computationally expensive to train and SVM are difficult to interpret.
Decision Tree (DT) (Quinlan 1986)	Supervised ML algorithms that create a model that predicts whether a person has T2DM based on their medical history and other risk factor.	They can be used to predict both binary and multiclass outcomes.	Sensitive to overfitting, and they can be difficult to scale to larger datasets
Naïve Beyes (Lindley 1958).	Bayes' theorem is a mathematical formula that relates the probability of an event to the probability of its causes	Very fast, easy to train and can be used to handle both categorical and continuous features.	Sensitivity to outliers, and it generally assumes that features are independent of each other, which is often not true in real world.
Artificial Neural Network (ANN) (McCulloch & Pitts 1943)	Inspired by the human brain, and they can learn complex relationships between the features and the target variable.	Non-linearity, capable of parallel processing and generalization.	It is overfitting, has training complexity and challenging to interpret why they make specific predictions or decisions.
K-Nearest Neighbors (KNN) (Cover & Hart 1967)	Works by finding the k most similar data points to a new data point and then predicting the label of the new data point based on the labels of the k-Nearest Neighbors	Simplicity, versatility, robust to outliers and no training time needed.	Computational complexity, sensitive to irrelevant features has an impact to imbalanced data leading to biased prediction.

First Stage: Identifying the Research Question

Two research questions were developed to attain the relevant information after appraising the topic of discussion i.e. (i) What are the most used ML methods in predicting T2DM?; (ii) How was the effectiveness of ML techniques evaluated in the prediction of T2DM?

Second Stage: Identifying Relevant Studies

The inclusion criteria used in this review was based on the Population-Concept-Context (PCC) recommendations developed by Joanna Briggs Institute (Joanne Briggs Institute 2015), described in Table 2. This review excluded studies focusing on populations with Type 1 Diabetes or Gestational Diabetes. Only T2DM studies were focused. This review also excluded studies outside of the 6-year time frame (1 January 2018 until 31 December 2023). This was because the rise of popularity of ML usage with disease studies emerges especially with T2DM and this allowed the scoping review to focus on recent research and developments of the topic. This field which involves technology methodology often sees rapid changes. Therefore, a short time frame is needed to reflect the current landscape.

A comprehensive search strategy using seven electronic databases: IEEE Xplore, JSTOR, PubMed, Sage, Scopus, Wiley and WOS were conducted. Studies published from 1 January 2018 until 31 December 2023 were included. Keyword searches were applied using Boolean operators (AND, OR, and NOT) which combine or exclude keywords in a search, resulting in more focused and precised ((“type 2 diabetes mellitus” OR T2DM) AND (“Machine Learning”)). The review workflow was conducted following the PRISMA guidelines (Figure 1).

Third Stage: Study Selection

Titles and abstracts were screened to ascertain the suitability of the articles. All records were transferred into a spreadsheet software program where irrelevant articles were excluded, and duplicates removed. Two reviewers determined the articles’ eligibility based on the title, abstract, and full text.

Fourth Stage: Charting the Data

The selected data were extracted and sorted from the chosen articles. Data extracted include the author’s name

TABLE 2: Inclusion criteria

Population	Concept	Context
T2DM patients from original country of residence; patients registered within their local country database.	Any machine learning methods in predicting T2DM prevalence from studies published between 2018 to 2023	Research articles from any countries, any setting; original research articles (excluding reviews or meta-analysis) that were written in English.

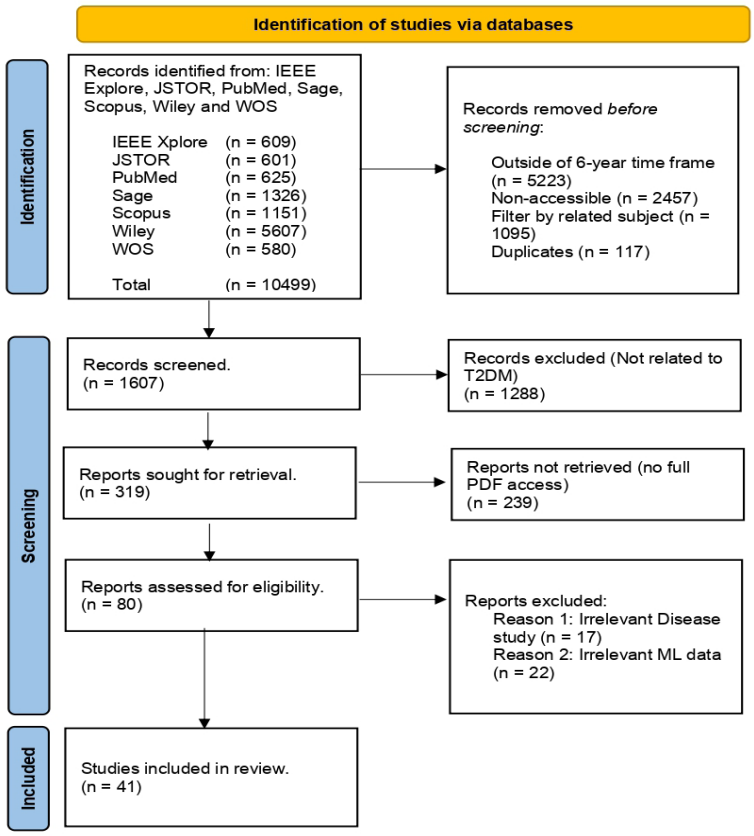

FIGURE 1: PRISMA 2020 flow diagram for scoping review database search

and year, author's country, study area, study size, database used, ML models involved, best DM predictors, best ML model outcome and performance. The best outcome was defined by the highest percentage of performance metrics of confusion matrix used in the studies. This included accuracy, precision, and F1 score that measures the corresponding ML models involved. But most of authors showed only accuracy or area under curve (AUC) for the ML performance metrics.

Fifth Stage: Collating, Summarising, and Reporting the Results

The information extracted was gathered, and consolidated on a descriptive table with numerical values, allowing interpretations through thematic summaries.

RESULTS

Of 10,499 results retrieved from the initial search Boolean from seven journal article search engines, 1607 articles were identified for screening

after others were removed because of non-accessible, duplicates, outside of the 6-year time frame and filtered by related subjects. Then, 80 articles were one-by-one further assessed for eligibility, answering the study's aim. Finally, 41 articles were included in the final analyses (Table 3). The main

TABLE 3: Result of 41 included articles for ML analyses on T2DM prediction

Author	Year	Author's country	Study Area	Database/source	Study size
Agliata et al.	2023	Italy	USA	The National Center for Health Statistics (NHANES) biennial survey, MIMIC-III and MIMIC-IV.	13687
Aguilera-Venegas et al.	2023	Spain	Spain	Nation-wide cohort diabetes study	5072
Al Sadi & Balachandran	2023	UK	Oman and USA	Al Shifa health system of South Al Batinah Province (Oman) and Pima Indian Diabetes Dataset	921 Oman, 768 Pima
Bernandini et al.	2020	Italy	Italy	Federazione Italiana Medici di Medicina Generale dataset	2433
Bhaskar et al.	2023	India	India	District Community Medical Center, India	152
Chang et al.	2022	UK	USA	Pima Indian Diabetes Dataset	768
Cheng et al.	2023	Taiwan	Taiwan	Kaohsiung Hospitals	647
Deberneh & Kim	2021	South Korea	South Korea	KNHANES dataset (Korea)	8454
Dritsas & Trigka	2022	Greece	Bangladesh	Sylhet Diabetes Hospital in Sylhet, Bangladesh	520
Dutta et al.	2022	Bangladesh	Bangladesh	Diabetes Diseases Classification (DDC) dataset from the northeastern part of South Asia (Bangladesh). BDHS-2011 and 2017–2018 BDHS surveys.	(5223) BDHS-2011 and (12119) 2017–2018 BDHS
Esmaily et al.	2018	Iran	Iran	MASHHAD database	9528 (1361 DM)
Fazakis et al.	2021	Greece	UK	English Longitudinal Study of Ageing (ELSA)	2009
Ganie et al.	2022	India	India	Population of Kishtwar and Rajouri geographical regions of Jammu and Kashmir	1,552 (DM 780)
Ginting et al.	2023	Indonesia	Indonesia	Secondary surveillance data from Puskesmas Johar Baru, Jakarta (Indonesia)	65,533 (2,766 DM)
Hahn et al.	2022	South Korea	South Korea	The Korean Genome and Epidemiology Study (KoGES) Ansan-Ansung cohort	1,425
Haneef et al.	2021	France	France	CONSTANCES cohort	44,659 (81 DM)
Iparraguirre-Villanueva et al.	2023	Peru	USA	Pima Indian Diabetes Dataset	768
Islam et al.	2020	Qatar	Texas	San Antonio Heart Study	1,791

Islam et al.	2023	Bangladesh	USA	NHANES 2009-2010, 2011-2012	4,922 (2009-2010), 4,936 (2011-2012)
Jiang[a] et al.	2023	China	China	Grassroots community service management information system in Haizhu District, Guangzhou	252,176
Jiang[b] et al.	2023	Japan	Japan	The Ministry of Healthcare, Labor, and Welfare (MHLW) (Japan)	28,292
Kopitar et al.	2020	Slovenia	Slovenia	10 Slovenian primary healthcare institutions	3,723
Li et al.	2023	China	China	National physical examination (NPE) project in 2020, China	4,075,431 (301,347 DM)
Liu et al.	2022	China	China	Screening records in Wuhan, China	127031
Mao et al.	2023	China	China	Chronic disease research database of Wuyishan City, Fujian Province	3687
Marzouk et al.	2022	Egypt	USA	Kaggle (Synthetic database), Pima Indian Diabetes Dataset	7691 (Kaggle), 768 (Pima)
Nuankaew et al.	2021	Thailand	USA and Mexico	Pima Indian Diabetes Dataset & Mendeley diabetes data (Iraq)	392 pima, 392 Mendeley
Ordonez-Guillen et al.	2023	Mexico	USA and Iraq	NHANES (USA) and Mexican National Health and Nutrition Surveys, ENSANUT (Mexico)	10077
Perveen et al.	2019	Pakistan	Canada	Canadian Primary Care Sentinel Surveillance Network (CPCSSN)	4403
Qin et al.	2022	China South	USA	Lifestyle data from NHANES database	17883
Shin et al.	2022	Korea	South Korea Synthetic	Health Promotion Center of Seoul St. Mary's Hospital (Korea)	138643
Stolfi et al.	2020	Italy	simulation	M-T2D simulation data	46170
Syed et al.	2020	Saudi Arabia	Saudi Arabia	Questionnaire (9Q) on Saudi Arabia population	4896
Tasin et al.	2023	Bangladesh	Bangladesh and USA	Local textile factory (Bangladesh), and Pima Indian Diabetes Dataset	768 Pima, 203 Local textile
Uddin et al.	2023	Bangladesh	Bangladesh	Questionnaires on Bangladeshi population	508
Ullah et al.	2022	Saudi Arabia	USA	CDC's Behavioral Risk Factor Surveillance System (BRFSS) in 2015	253680
Wang et al.	2023	China	China	Dongguan residents' questionnaires (China)	8013 (4023 DM)
Yilmaz	2022	Turkey	USA	University of California Irvine (UCI) database	520
Zhang et al.	2020	China	China	Henan Rural Cohort Study	36652
Zhang et al.	2021	China	China	Henan Rural Cohort Study	37730

reasons for removal from the scoping review were multiple diseases studies, not relevant to the research questions and no outcome of interest were presented. A total of 22 countries represented the authors' home institutions: China (n=9), Bangladesh (n=4), Italy and South Korea (n=3) each, Greece, India, Saudi Arabia, and United Kingdom (n=2) each, Egypt, France, Indonesia, Iran, Japan, Mexico, Pakistan, Peru, Qatar, Slovenia, Spain, Taiwan, Thailand, Turkey with one each. However, the study setting often was not consistent with the author's country. We found a total of eight different study settings from the articles: USA (n=13), Bangladesh (n=4), China (n=8), South Korea (n=3), India (n=2), Canada, France, Indonesia, Iran, Italy, Iraq, Japan, Oman, Saudi Arabia, Slovenia, Spain, Taiwan, Texas, United Kingdom, Mexico with each (n=1). Publication years included 2018 (n=2), 2019 (n=1), 2020 (n=6), 2021 (n=5), 2022 (n=11) and 2023 (n=16).

A total of 62 different ML models and 225 ML models all together tested for diabetes throughout all 41 included studies. The 11 most frequently used which was occurrence more than five times, includes: Random Forest (RF) (n=33), Logistic Regression (LR) (n=23), Decision Trees (DT) (n=19), Support Vector Machine (SVM) (n=17), Extreme Gradient Boost (XGB) (n=14), k-Nearest Neighbors (KNN) (n=14), Naive Bayes (NB) (n=11), Artificial Neural Network (ANN) (n=8), Multilayer Perceptron (MLP) (n=7), Gradient Boosting (GB) (n=6), and Ensemble learning algorithms representing a combination of multiple ML techniques into one model (n=5). Other ML models that not mentioned were presented in Table 4.

Average accuracy and AUC of the best ML model in all study were more than 70% or 0.7. There were some of them that reported below 70% or 0.7 but when compared to other ML within the study, the ML model still bested another ML model respectively. Table 4 also showed the specific best T2DM predictors for the concurrent best ML model that successfully analysed.

DISCUSSION

We evaluated 62 different supervised ML techniques from all 41 articles

TABLE 4: Result of 41 included articles of best DM predictors, ML model used, best ML model and performance

Author	Best DM predictors (high risk)*	ML model used**	Best ML model**	Best model performance % **
Agliata et al. (2023)	Gender, Age, HDL, HbA1c, BP, TG, BMI	ADAM, SGD, RMSPROP, LM	Ensemble + ADAM	Accuracy 86, ROC 0.934
Aguilera-Venegas et al. (2023)	Not specified	DT, RF, KNN, NN	RF	Accuracy 92.91
Al Sadi & Balachandran (2023)	Gender, Age, BP, BMI, WC, HDL, FPG, HbA1c	KNN, SVM, NB, RF, DT, ANN, LDA	RF and DT	Accuracy 98.38

Bernardini et al. (2020)	BP, Age, Arterial hypertension, BMI	SB-SVM, RF, DT, KNN, SVM Lin, LR, Gauss, MLP, DBN	SB-SVM	Case 2, 3 = Recall 74.64 and 65.33, AUC 81.43 and 68.90
Bhaskar et al. (2023)	Concentrated Acetone in breath	CORNN-MLP, CORNN-SVM, SVD-SVM, PCA-KNN, PCA-SVM, Shallow CNN, CNN-MLP, CNN-RF, CNN-SVM	CORNN-SVM	Accuracy 98.02
Chang et al. (2022)	HbA1c, BMI, Age, Insulin and Skin thickness	NB, RF, DT	RF (all features), NB (2 and 3 features)	RF (Accuracy 79.57), NB (Accuracy 79.13)
Cheng et al. (2023)	Depression and anxiety, HbA1c	KNN, RF	RF	Accuracy 84, AUC 95
Deberneh & Kim (2021)	FPG, HbA1c, Gamma-glutamyl transferase, BMI	LR, RF, SVM, XGB, Ensemble	Ensemble	Accuracy 71-73
Dritsas & Trigka (2022)	Polyuria, Polydipsia, Sudden weight loss, Age, Gender, Partial paresis	BayesNet, NB, SVM, LR, ANN, KNN, J48, LMT, RF, RT, RepTree, RotF, AdaBoost, SGD, Stacking	KNN and RF	Accuracy 98.58 (99.22 with SMOTE)
Dutta et al. (2022)	BMI, Age, BP, Occupation	GNB, BNB, RF, DT, XGB, LGBM	Ensemble (DT + RF + XGB + LGBM)	Accuracy 73.5, ROC 83.2
Esmaily et al. (2018)	TG, FHD, hs-CRP, BP, BMI	DT, RF	RF	Accuracy 71.1, ROC 77.3
Fazakis et al. (2021)	Not specified	NB, DT, ANN, DNN, RF, LR, Ensemble (Weighted voting, Voting, Stacking of LR, RF)	Ensemble (Weighted Voting LR RF)	AUC 0.884
Ganie et al. (2022)	Age, Gender, FHD, Extreme thirst, Urination, Drinking, Smoking, BMI	KNN, LR, SVM, NB, DT, RF, GB	GB	Accuracy 97.24
Ginting et al. (2023)	Age, Gender, FHD, HPT, HbA1c, BMI	RF	RF	Accuracy 84
Hahn et al. (2022)	Genome-wide Polygenic Risk Score	LR, RF	RF	Accuracy 85.4
Haneef et al. (2021)	Age, Glucose blood test, HbA1c, Alkaline phosphatase test, Gamma glutamyle transferase test, Transaminases blood test, Uric acid test, Creatinine level blood	LR, FDA, DT, LDA	LDA	Accuracy 67
Iparraguirre-Villanueva et al. (2023)	HbA1c, BP, Skin thickness, Insulin, BMI, Age, FHD	KNN, BNB, DT, LR, SVM	KNN	Accuracy 72.1

Islam et al. (2020)	Age, Education level, Martial status, BP, Smoking, BMI, PA, HDL, Ethnicity	LR, NB, J48, MLP, RF	RF	Accuracy 84.9, AUC 0.677
Islam et al. (2023)	Not specified	SVM, RF, Bagging, Boosting, NB, A1DE, A1DE, Ensemble	Ensemble for top 30 feature	Accuracy 95.94%, Sensitivity 100%, Specificity 91.5%, AUC 96.3%
Jiang[a] et al. (2023)	BMI, Age, PA, Drinking, BP, Food intake	RF, XGB, KNN, Ensemble	RF	Accuracy 91.24, ROC 91.15
Jiang[b] et al. (2023)	Age, Gender, BMI, Hyperlipidemia, HPT, Public pension, Health awareness level	LA (logistic analysis), LR, LDA, Hayashi's Quantification method 2 (q2), RF, XGB	Average across all	No clear result
Kopitar et al. (2020)	Hyperglycemia, Age, HDL, Triglycerides,	LR, Glmnet, RF, XGB, LGBM	LBGM	Balanced and stable in many tests.
Li et al. (2023)	HPT, FPG, Age, Coronary heart disease, Ethnicity, FHD, TG, WC, HDL, BMI	CART, LGBM, RF, XBG, MLP, LR, TabNet (NN)	XGB	AUC 0.9122
Liu et al. (2022)	FBG, Education level, Exercise activity, Gender, WC	LR, DT, RF, XGB	XGB	AUC 0.7805, Sensitivity 0.6452, Specificity 0.7577, Accuracy 0.7503
Mao et al. (2023)	Age, FHD, IFG, IGT, HPT, TG, Alanine Aminotransferase and Gamma glutamyl transpeptidase.	XGB, RF, LBGM, AdaBoost, MLP, GNB	RF	AUC 0.855
Marzouk et al. (2022)	Diabetes pedigree function, HbA1c, BMI, BP, Age	DT, SVM, RF, GB, MLP, ANN, KNN, LR, NB	ANN (for Pima), DT and GB (for synthetic Kaggle)	(ANN Pima) Precision 82, Accuracy 81.6993 (DT) Precision 90, Accuracy 87.37 (GB) Precision 88, Accuracy 87.49
Nuankaew et al. (2021)	Increased level of HDL particularly in women	AWOD, KNN, SVM, RF, DL	AWOD	Accuracy 93.22% (Pima), Accuracy 98.95% (Mendeley data)
Ordonez-Guillen et al. (2023)	Age, BMI, HbA1c, FPG, Insulin	SVM, KNN, MLP, SNNN	SVM and MLP	Accuracy 98.8% and 98.9%
Perveen et al. (2019	FPG	DT, NB	NB with K-medoids under sampling	Average ROC 86%
Qin et al. (2022)	Sleep time, Energy, Age	CATBoost, XGB, RF, LR, SVM	CATBoost	Accuracy 82.1%, AUC 0.83
Shin et al. (2022)	FPG, BMI, gamma-GTP, Age, WC	GB, RF	GB	Accuracy 0.815
Stolfi et al. (2020)	BMI, FPG, TNF	LR, Polynomial degree, Multivariate RF	RF	MSEin 0.01991, MSEout 0.02769

Syed et al. (2020)	Region, Gender, BMI, Healthy diet, BP, Smoking	LR, Average perceptron, NB , ANN, SVM, LD SVM, Decision jungle, Decision forest, Boosted DT	Decision Forest	Accuracy 0.821, Precision 0.776, Recall 0.89, AUC 0.867, F1 score 0.829
Tasin et al. (2023)	HbA1c, BMI, Age, Skin thickness	DT, SVM, RF, LR, KNN, Bagging, Adaboost, XGB, Voting	XGB	Accuracy 81%, F1 coefficient 0.81, AUC 0.84
Uddin et al. (2023)	Age, Extreme thirst, FHM	DT, LR, SVM, GB, XGB, RF, Ensemble	Ensemble	Accuracy 0.876
Ullah et al. (2022)	BMI, Age, Insulin resistance, HPT	KNN, RF, XGB, Bagging, AB ensemble	KNN	Accuracy 98.38%, ROC 0.98
Wang et al. (2021)	Potato consumption, Fish consumption, TC, FPG, HDL	LR, SVM, BPNN, CART, C4.5 (DT), DNN	BPNN	Accuracy 93.7%, Precision 94.6%, Recall 92.8%, AUC 97.7%
Yilmaz (2022)	Polyuria, Polydipsia, Sudden weight loss, Partial paresis	NB, LR, DT, RF, SVM, XGB, Proposed Hybrid XGB, KNN	Proposed Hybrid XGB	Accuracy 97.26% and 95.16%
Zhang et al. (2020)	Urine glucose, Sweet flavor, FDM, Waist to hip ratio, Age, HPT, HR, Creatine	LR, CART, ANN, SVM, RF, GB	GB	AUC 0.872
Zhang et al. (2021)	Age (older), Gender (woman), Education level (less), FHD (yes), waist to hip ratio (high), HR (high), BP (high)	LR, ANN, XGB, RF, GB, JBM	JBM	AUC 0.885, Recall 0.847
Zou et al.	Not specified	DT, NN, RF	RF	Accuracy 80.84%

*Note: High-Density Lipoprotein (HDL), Triglycerides (TG), Hemoglobin A1C Glucose Blood Level (HbA1c), Blood Pressure (BP), Physical Activity (PA), Fasting Plasma Glucose (FPG), Body Mass Index (BMI), Waist Circumference (WC), Hypertension (HPT), Impaired Glucose Tolerance (IGT), Impaired Fasting Glycaemia (IFG), Total Cholesterol (TC), Family History Diabetes (FHD), Heart Rate (HR), Tumor Necrosis Factor (TNF), High sensitivity C-reactive Protein (hs-CRP)

**Note: Adaptive Boosting (AB/AdaBoost), Adaptive Moment Estimation (ADAM), Artificial Neural Network (ANN), Neural Network (NN), Average Weighted Objective Distance (AWOD), Bayesian Network (BayesNet), Bernouli Naïve Bayes (BNB), Back Propagation Neural Network (BPNN), Categorical Boosting (CATBoost), Classification and Regression Tree (CART), Correlational Neural Network (CORNN), Convolutional Neural Network (CNN), Deep Belief Network (DBN), Deep Learning (DL), Deep Neural Network (DNN), Decision Tree (DT), Flexible Discriminant Analysis (FDA), Gradient Boosting (GB), Regularised generalised linear model (Glemnet), Gaussian Naïve Bayes (GNB), Decision Tree based on Iterative Dichtomiser 3 (J48), Joint Bagging-Boosting (JBM), k-Nearest Neighbor (KNN), Localised multiple kernel-SVM (LD-SVM), Logistic Analysis (LA), Linear Discriminant Analysis (LDA), Light Gradient Boosting Machine (LGBM), Levenberg-Marquardt (LM), Logistic Model Tree (LMT), Logistic Regression (LR), Multilayer Perceptron (MLP), Naïve Bayes (NB), Principal Component Analysis (PCA), Reduce Error Pruning Tree (RepTree), Ensemble of Naïve Bayes (A1DE). Random Forest (RF), Root Mean Squared Propagation (RMSPROP), Rotation Forest (RotF), Suppport Vector Machine (SVM), Sparse Balanced-SVM (SB-SVM), Stochastic gradient descent (SGD), Self-Normalized Neural Network (SNNN), Singular Value Decomposition-SVM (SVD-SVM), Extreme Gradient Boosting (XGB), Receiver operating characteristic (ROC), Area under the curve (AUC).

included in our review. Most studies originated from China, Bangladesh and South Korea. However, some of those studies did not project the study setting of their home country. USA had the highest occurrence of study setting (n=13), followed by China (n=8). This showed that the bibliographies retrieved do not consistently report works from authors home country. So, there is no equal contribution to the corresponding T2DM analyses on behalf of the author's country. Plausible explanations are authors from these countries would utilise countries that have higher risk or prevalence for diabetes, as ML algorithms require huge datasets to be executed via computational methods. Next, high-income countries (e.g. the USA and China) invest largely on big data analytics, hence are capable to building huge databases with open data sharing policies to scientists from different countries that allows prediction of T2DM at the global or multi-country perspective. This also showed that datasets from the USA are abundant and readily used by everyone to tinker and train their own proposed ML model such as PIMA Indian Diabetes database which is the most used database in this scoping review (n=6). However, there are risks of bias as there are definite differences in demographic and social aspects in local risk factors within different countries (Celi et al. 2022). There are insufficient datasets from Southeast Asia countries despite regional high prevalence of T2DM cases, due to a shared common ground of rapid urbanisation, lifestyle changes, obesity rates, genetics, and healthcare access which affect local risk factors for T2DM (ASEAN Sustainable Urbanisation 2022; Ramachandran & Snehalatha 2010). Despite accelerated risks, these countries have weaker surveillance systems and lack of large, integrated databases to develop artificial intelligence capacities.

Improvements for ML Techniques

Ensemble methods are a type of ML algorithm that combines the predictions of multiple base learners to improve the overall performance of the model. They are often used to improve the accuracy, robustness, and generalisation of ML models. Most popular ensemble models are Bagging, Boosting and Random Forest. Ensembles can also help to reduce the variance of the predictions, which can make the predictions more reliable.

But all in all, to achieve the best performance of disease prediction, the specific objective, problem, and data available, with an extra help from additional ML data preprocessor such as SMOTE (Synthetic Minority Oversampling Technique) will tremendously increase the accuracy (Alghamdi et al. 2017). ML methods also can enhance performance further using Feature Selection techniques, data pre-processing (e.g., SMOTE), hyperparameter tuning using "Grid Search."

In nine out of 41 study analysed, specially created and modified ML that suits the dataset of each researcher got the best model in their respective study. These are ADAM Ensemble

model (Agliata et al. 2023), Average Weighted Objective Distance (AWOD) model (Nuankaew et al. 2021), Back Propagation Neural Network (BPNN) model (Wang et al. 2023), CATBoost (Qin et al. 2022), Correlational Neural Network (CORNN-SVM) model (Bhaskar et al. 2023), Decision Forest (Syed et al. 2020), Joint Bagging-Boosting (Zhang et al. 2021), proposed hybrid XGB (Yilmaz 2022) and Sparse Balanced-SVM (Bernardini et al. 2020). These custom models were catered for the specific task that is to achieve the highest accuracy and AUC compared to other model that were tested together considering the T2DM risk factor that has been feature selected beforehand. Combination of feature selection method to carefully allowing the best outcome to the proposed ML methodology achieve the best predictive capability.

Improvements for Current Study Limitations

Many of the reviewed articles focused on basic demographic risk factors of T2DM; there was no "external risk factors" such as built environment in T2DM analysis. Risk factors of most studies used basic census data such as age, gender, occupation, educational level, family history (diabetic), smoking, drinking alcohol, physical activity, etc. As well as anthropometry and biochemical factors such as blood pressure, blood test, cholesterol level, BMI, etc.

A study by Qin et al. (2022) showed a unique risk factor comprised of lifestyle variables correlates with T2DM prediction. The study utilises the NHANES database, USA which includes all basic demographics, alcohol intake, smoking status, sleeping hours, dietary, laboratory, and physical examination. The best ML model was CATBoost with accuracy 82.1% and AUC of 0.83. Combination of complete and balanced sets of biological data with lifestyle data can further improve the analysis towards 'real-time' and closer to human behaviour analysis for diabetes prediction.

A study by Jiang et al. (2023) in Japan proved that they have found a new DM predictor which are public pension and health awareness level that correlated to prevalence of T2DM. Different demographic properties of different socioeconomic attributes could lead to the possibility of action that pushed into DM risk factor territory.

A study by Bhaskar et al. (2023) was carried out in India where they proposed a methodology to detect Diabetes Mellitus through sensing chemical content when tested through breathing with the clinical apparatus. This method might be useful because of the novelty and detection of acetone content in breathing has been proved to correlate with prevalence of T2DM.

A study by Stolfi et al. (2020) has one uniqueness which is their database is fully simulated and fabricated (virtual subjects). This means that the study does not required the fuss of collecting clinical data. But the drawback is that the data is fully synthetic and not reflecting any real situation even if the data is simulated. There will be bias to the result because of human

randomness is something that cannot be simulated easily.

A study by Ganie et al. (2022) also has similar risk factor variables which are lifestyle biological features. These lifestyle features include age, sex, family history of diabetes, smoking, alcohol intake, thirst level, urination, BMI, fatigue level and diabetes status. Analysis that uses lifestyle parameters may provide better results on external prevention ideas to reduce diabetes prevalence. The database was from the Kishtwar and Rajouri geographical regions of Jammu and Kashmir, India. The best ML model was GB with accuracy of 97.2%. Another unique risk factor study by Hahn et al. (2022) included genome-wide polygenic risk score and metabolic profile as their risk factor to predict T2DM. The database used was from The Korean Genome and Epidemiology Study (KoGES) Ansan-Ansung cohort. The best ML model was RF with accuracy of 85.4%.

Limitation

Future studies associated with T2DM should test additional possible risk factors that may be associated with the prevalence of T2DM worldwide which may be beneficial to policy makers. We also recommend further studies in Southeast Asian countries as these may bring additional, local risk factors to light. Each ML technique has its own strengths and weaknesses. Although Random Forest has the highest occurrence in the best ML models in the 41 included articles reviewed, that does not mean that it is the best technique for most used cases. Because different risk factors play different roles in determining which ML method is most suited for the prediction task, the critical task of properly selecting ML techniques must be taken seriously.

There are some limitations to the study. We cannot rule out the possibility that we did not find any relevant studies despite using broad search terms and we would not have found any newer studies described only in conference proceedings or unpublished studies. Papers that were correctly excluded (according to our criteria) may still be useful for T2DM prediction, and further review of these may suggest new methodologies for generating T2DM predictions. The review also cannot be used as a definitive guide to prediction approaches with higher predictive skill, because settings and methodologies varied greatly. Because of this diversity, we approached the review as a scoping rather than a systematic review. Furthermore, the goal was not to provide detailed critiques of ML methodologies. Such an evaluation would be useful, but we believe that a broader review of prediction applications provides the context for this.

There was also limitation on the ML algorithm - there is either too much sample data or too little data. A small sample may not be adequately representing the underlying distribution of the data. Thus, this can lead to biased model training because of the higher variability, making it challenging to discern true patterns. Small samples also are more sensitive to influence of outliers which can have

a disproportionate impact on model training, leading to skewed predictions. Another side of the problem is that if the sample size is too large. This will impact on the computational resources and will also be time consuming. Beyond a certain point, adding more data might not significantly improve model performance. Researchers must find the perfect law of diminishing returns to get the optimal sample size so that we decrease the risk of drawing incorrect conclusions about the prediction of our ML model.

CONCLUSION

Based on the studies reviewed, we conclude that a correct choice of ML technique, combined with additional supplementary enhancements on top of the ML model can help boost the prediction performance. Generally, an ensemble method or a "specially calculated model" is going to generate higher accuracy because of the complexity of the combined ML algorithm, much like Random Forest (an ensemble of Decision Trees). Our study also reveals that many ML applications to the study of T2DM are conducted on data outside the authors' country of residence. The limited number of studies meeting these criteria suggest a need for increased effort toward developing in-country data resources to support ML for T2DM and increase chances for use for the development and evaluation of local prevention and treatment options. The study also emphasises the necessity of utilising ML techniques in healthcare practices to enhance time management and workload, potentially leading to reduction of healthcare burden. The medical community can make educated judgements and put preventive measures into place thanks to this knowledge, which ultimately may help to delay the spread of disease. The findings of this scoping review can contribute to the growing body of knowledge on ML applications in T2DM prediction. By elucidating the relationship between environmental factors and T2DM risk, this review has the potential to inform public health initiatives, policy-making, and clinical decision-making. Furthermore, the review will serve as a valuable resource for researchers, practitioners, and policymakers working at the intersection of epidemiology and ML.

ACKNOWLEDGEMENT

This work was supported by the Ministry of Higher Education (MOHE) Malaysia Fundamental Research Grant Scheme (FRGS/1/2022/SKK04/UKM/01/1) and Universiti Kebangsaan Malaysia for research management.

AUTHOR CONTRIBUTIONS

M.F.M.R., M.R.A.M., K.N.A.M. involved in the conceptualisation, methodology, extensive search of articles, critical review of articles, result synthesis and original draft write-up. N.S., K.G., F.I.M. and L.A.W reviewed the final manuscript.

ETHIC DECLARATION

Not applicable.

FUNDING

The authors received funding from the Ministry of Higher Education (MOHE) Malaysia Fundamental Research Grant Scheme with the reference number of FRGS/1/2022/SKK04/UKM/01/1.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding authors upon request.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Abbas, H.T., Alic, L., Erraquntla, M., Ji, J.X., Abdul-Ghani, M., Abbasi, Q.H., Qaraqe, M.K. 2019. Predicting long-term type 2 diabetes with support vector machine using oral glucose tolerance test. *PLoS One* **14**(12): e0219636.

Agliata, A., Giordano, D., Bardozzo, F., Bottiglieri, S., Facchiano, A., Tagliaferri, R. 2023. Machine learning as a support for the diagnosis of type 2 diabetes. *Int J Mol Sci* **24**(7): 6775.

Aguilera-Venegas, G., López-Molina, A., Rojo-Martínez, G., Galán-García, J.L. 2023. Comparing and tuning machine learning algorithms to predict type 2 diabetes mellitus. *J Comput Appl Math* **427**: 115115.

Alghamdi, M., Al-Mallah, M., Keteyian, S., Brawner, C., Ehrman, J., Sakr, S. 2017. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford exercise testing (FIT) project. *PLoS One* **12**(7): e0179805.

Al Sadi, K., Balachandran, W. 2023. Prediction model of type 2 diabetes mellitus for oman prediabetes patients using artificial neural network and six machine learning classifiers. *Appl Sci* **13**(4): 2344.

Arksey, H., O'Malley, L. 2005. Scoping studies: Towards a methodological framework. *Int. J. Social Res. Methodol. Theory Pract* **8**: 19-2.

ASEAN Sustainable Urbanization Report 2022. 2022. Sustainable Cities towards 2025 and Beyond. ISBN 978-623-5429-16-8 (EPUB).

Bernardini, M., Romeo, L., Misericordia, P., Frontoni, E. 2020. Discovering the type 2 diabetes in electronic health records using the sparse balanced support vector machine. *IEEE J Biomed Health Inform* **24**(1): 235-46.

Bhaskar, N., Bairagi, V., Boonchieng, E., Munot, M.V. 2023. Automated detection of diabetes from exhaled human breath using deep hybrid architecture. *IEEE Access* **11**: 51712-22.

Breiman, L. 2001. Random forests. *Mach Learn* **45**(1): 5-32.

Celi, L.A., Cellini, J., Charpignon, M-L., Dee, E.C., Dernoncourt, F., Eber, R., Mitchell, W.G., Moukheiber, L., Schirmer, J., Situ, J., Paguio, J., Park, J., Wawira, J.G., Yao, S. 2022. Sources of bias in artificial intelligence that perpetuate healthcare disparities - A global review. *PLOS Digit Health* **1**(3): e0000022.

Chang, V., Bailey, J., Xu, Q.A., Sun, Z. 2022. Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. *Neural Comput Applic* **35**: 16157-73.

Cheng, Y.L., Wu, Y.R., Lin, K.D., Lin, C.R., Lin, I.M. 2023. Using machine learning for the risk factors classification of glycemic control in type 2 diabetes mellitus. *Healthcare (Basel)* **11**(8): 1141.

Cover, T., Hart. P. 1967. Nearest neighbor pattern classification. *IEEE Trans Inf Theory* **13**(1): 21-7.

Deberneh, H.M., Kim, I. 2021. Prediction of type 2 diabetes based on machine learning algorithm. *Int J Environ Res Public Health* **18**(6): 3317.

Dritsas, E., Trigka, M. 2022. Data-driven machine-learning methods for diabetes risk prediction. *Sensors* **22**(14): 5304.

Dutta, A., Hasan, M.K., Ahmad, M., Awal, M.A., Islam, M.A., Masud, M., Meshref, H. 2022. Early prediction of diabetes using an ensemble of machine learning models. *Int J Environ Res Public Health* **19**(19): 12378.

Esmaily, H., Tayefi, M., Doosti, H., Ghayour-Mobarhan, M., Nezami, H., Amirabadizadeh, A. 2018. A comparison between decision tree and random forest in determining the risk factors associated with type 2 diabetes. *J Res Health Sci* **18**(2): 412.

Fazakis, N., Kocsis, O., Dritsas, E., Alexiou, S., Fakotakis, N., Moustakas, K. 2021. Machine learning tools for long-term type 2 diabetes risk prediction. *IEEE Access* **9**: 103737-57.

Firdous, S., Wagai, G.A., Sharma, K. 2022. A survey on diabetes risk prediction using machine learning approaches. *J Family Med Prim Care* **11**(11): 6929-34.

Galicia-Garcia, U., Benito-Vicente, A., Jebari, S., Larrea-Sebal, A., Siddiqi, H., Uribe, K.B., Ostolaza, H., Martin, C. 2020. Pathophysiology of type 2 diabetes mellitus. *Int J Mol Sci* **21**: 6275.

Ganie, S.M., Malik, M.B., Arif, T. 2022. Performance analysis and prediction of type 2 diabetes mellitus based on lifestyle data using machine learning approaches. *J Diabetes Metab Disord* **21**(1): 339-52.

Ginting, J.B., Suci, T., Ginting, C.N., Girsang, E. 2023. Early detection system of risk factors for diabetes mellitus type 2 utilization of machine learning-random forest. *J Family Community Med* **30**(3): 171-9.

Hahn, S.J., Kim, S., Choi, Y.S., Lee, J., Kang, J. 2022. Prediction of type 2 diabetes using genome-wide polygenic risk score and metabolic profiles: A machine learning analysis of population-based 10-year prospective cohort study. *EBioMedicine* **86**: 104383.

Haneef, R., Fuentes, S., Fosse-Edorh, S., Hrzic, R., Kab, S., Cosson, E., Gallay, A. 2021. Use of artificial intelligence for public health surveillance: A case study to develop a Machine Learning-algorithm to estimate the incidence of diabetes mellitus in France. *Arch Public Health* **79**(1): 168.

Hosmer, D.W., Lemeshow, S., Sturdivant, R.X. 2013. Applied logistic regression. *John Wiley & Sons Inc.*

Iparraguirre-Villanueva, O., Espinola-Linares, K., Flores Castaneda, R.O., Cabanillas-Carbonell, M. 2023. Application of machine learning models for early detection and accurate classification of type 2 diabetes. *Diagnostics* **13**(14): 2383.

Islam, M.M., Rahman, M.J., Menhazul Abedin, M., Ahammed, B., Ali, M., Ahmed, N., Maniruzzaman, M. 2023. Identification of the risk factors of type 2 diabetes and its prediction using machine learning techniques. *Health Syst* **12**(2): 243-54.

Islam, M.S., Qaraqe, M.K., Belhaouari, S.B., Abdul-Ghani, M.A. 2020. Advanced techniques for predicting the future progression of type 2 diabetes. *IEEE Access* **8**: 120537-47.

Jiang[a], L., Xia, Z., Zhu, R., Gong, H., Wang, J., Li, J., Wang, L. 2023. Diabetes risk prediction model based on community follow-up data using machine learning. *Prev Med Rep* **35**: 102358.

Jiang[b], P., Suzuki, H., Obi, T. 2023. Interpretable machine learning analysis to identify risk factors for diabetes using the anonymous living census data of Japan. *Health Technol* **13**(1): 119-31.

Joachims, T. 1998. Making large-scale SVM learning practical. *SFB 475: Komplexitätsreduktion Multivariaten Datenstrukturen, Univ. Dortmund, Dortmund, Tech. Rep*. No. 1998: 28.

Joanne Briggs Institute. 2015. *Methodology for JBI Scoping Reviews*; Joanne Briggs Institute: Adelaide, Australia; 1-24.

Joshi, R.D., Dhakal, C.K. 2021. Predicting type 2 diabetes using logistic regression and machine learning approaches. *Int J Environ Res Public Health* **18**(14): 7346.

Khanam, J.J., Foo, S.Y. 2021. A comparison of machine learning algorithms for diabetes prediction. *ICT Express* **7**(4): 432-9.

Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A., Stiglic, G. 2020. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Sci Rep* **10**(1): 11981.

Li, L., Cheng, Y., Ji, W., Liu, M., Hu, Z., Yang, Y., Wang, Y., Zhou, Y. 2023. Machine learning for predicting diabetes risk in Western China adults. *Diabetol Metab Syndr* **15(1):** 165.

Lindley, D.V. 1958. Fiducial distributions and Bayes' theorem. *J Royal Stat Soc Series B (Methodological).* **1**: 102-7.

Liu, Q., Zhang, M., He, Y., Zhang, L., Zou, J., Yan, Y., Guo, Y. 2022. Predicting the risk of incident type 2 diabetes mellitus in Chinese elderly using machine learning techniques. *J Pers Med* **12**(6): 905.

Mao, Y., Zhu, Z., Pan, S., Lin, W., Liang, J., Huang, H., Li, L., Wen, J., Chen, G. 2023. Value of machine learning algorithms for predicting diabetes risk: A subset analysis from a real-world retrospective cohort study. *J Diabetes Investig* **14**(2): 309-20.

Marzouk, R., Alluhaidan, A.S., El Rahman, S.A. 2022. An analytical predictive models and secure web-based personalized diabetes monitoring system. *IEEE Access* **10**: 105657-73.

McCulloch, W.S., Pitts, W. 1943. A logical calculus of the ideas immanent in nervous activity. *Bull Math Biophys* **5**(4): 115-33.

Ngiam, K.Y., Khor, I.W. 2019. Big data and machine learning algorithms for health-care delivery. *Lancet Oncol* **20**(5): e262–e273.

Nuankaew, P., Chaising, S., Temdee, P. 2021. Average weighted objective distance-based method for type 2 diabetes prediction. *IEEE Access* **9**: 137015-28.

Ordonez-Guillen, N.E., Gonzalez-Compean, J.L., Lopez-Arevalo, I., Contreras-Murillo, M., Aldana-Bobadilla, E. 2023. Machine learning based study for the classification of Type 2 diabetes mellitus subtypes. *BioData Min* **16**(1): 24.

Perveen, S., Shahbaz, M., Keshavjee, K., Guergachi, A. 2019. Metabolic syndrome and development of diabetes mellitus: Predictive modeling based on machine learning techniques. *IEEE Access* **7**: 1365-75.

Qin, Y., Wu, J., Xiao, W., Wang, K., Huang, A., Liu, B., Yu, J., Li, C., Yu, F., Ren, Z. 2022. Machine learning models for data-driven prediction of diabetes by lifestyle type. *Int J Environ Res Public Health* **19**(22): 15027.

Quinlan, J.R. 1986. Induction of decision trees. *Mach Learn* **1**(1): 81-106.

Ramachandran, A, Snehalatha, C. 2010. Rising

burden of obesity in Asia. *J Obes* 2010: 868573.

Rask-Madsen, C., King, G.L. 2013. Vascular complications of diabetes: Mechanisms of injury and protective factors. *Cell Metab* **17**: 20-33.

Shin, J., Kim, J., Lee, C., Yoon, J.Y., Kim, S., Song, S., Kim, H.S. 2022. Development of various diabetes prediction models using machine learning techniques. *Diabetes Metab J* **46**(4): 650-7.

Stolfi, P., Valentini, I., Palumbo, M.C., Tieri, P., Grignolio, A., Castiglione, F. 2020. Potential predictors of type-2 diabetes risk: Machine learning, synthetic data and wearable health devices. *BMC Bioinformatics* **21**(17): 508.

Syed, A.H., Khan, T. 2020. Machine learning-based application for predicting risk of type 2 diabetes mellitus (T2DM) in Saudi Arabia: A retrospective cross-sectional study. *IEEE Access* **8***:* 199539-61.

Tasin, I., Nabil, T.U., Islam, S., Khan, R. 2023. Diabetes prediction using machine learning and explainable AI techniques. *Healthc Technol Lett* **10**(1-2): 1-10.

Tricco, A.C., Lillie, E., Zarin W, et al. 2018. PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Ann Intern Med* **169**: 467-73.

Tripathi, B.K., Srivastava, A.K. 2006. Diabetes mellitus: Complications and therapeutics. *Med Sci Monit* **12***:* RA130-47.

Uddin, M.J., Ahamad, M.M., Hoque, M.N., Walid, M.A.A., Aktar, S., Alotaibi, N., Alyami, S.A., Kabir, M.A., Moni, M.A. 2023. A comparison of machine learning techniques for the detection of type-2 diabetes mellitus: Experiences from Bangladesh. *Information* **14**(7): 376.

Ullah, Z., Saleem, F., Jamjoom, M., Fakieh, B., Kateb, F., Ali, A.M., Shah, B. 2022. Detecting high-risk factors and early diagnosis of diabetes using machine learning methods. *Comput Intell Neurosci* **29**: 2557795.

Wang, S., Chen, R., Wang, S., Kong, D., Cao, R., Lin, C., Luo, L., Huang, J., Zhang, Q., Yu, H., Ding, Y.L. 2023. Comparative study on risk prediction model of type 2 diabetes based on machine learning theory: A cross-sectional study. *BMJ Open* **13**(8): e069018.

Wang, X., Zhai, M., Ren, Z., R, H., Li, M., Quan, D., Chen, L., Qiu, L. 2021. Exploratory study on classification of diabetes mellitus through a combined Random Forest Classifier. *BMC Med Inform Decis Mak* **21**: 105.

Yilmaz, A. 2022. Prediction of type 2 diabetes mellitus using feature selection-based machine learning algorithms. *Health Problems of Civilization* **16**(2): 128-39.

Zhang, L., Wang, Y., Niu, M., Wang, C., Wang, Z. 2020. Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: the Henan rural cohort study. *Sci Rep* **10**(1): 4406.

Zhang, L., Wang, Y., Niu, M., Wang, C., Wang, Z. 2021. Nonlaboratory-based risk assessment model for type 2 diabetes mellitus screening in chinese rural population: A joint bagging-boosting model. *IEEE J Biomed Health Inform* **25**(10): 4005-4016.

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., Tang, H. 2018. Predicting diabetes mellitus with machine learning techniques. *Front Genet* **9**: 515.