Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 303-320

A Review of Machine Learning Approaches for Diabetes Risk Prediction

[1] Shiv Prakash Kichara ,[2] Amit Tiwari

[1] Research Scholar, Department of Computer Science and Engineering, Suresh Gyan Vihar University, Jaipur
[2] Assistant Professor. Suresh Gyan Vihar University, Jaipur

Abstract: This review addresses the growing burden of type-2 diabetes and the practical challenge that many individuals remain undiagnosed until complications appear, while conventional logistic-regression risk scores (built on a few predictors like age, BMI, family history and BP) often miss nonlinear interactions and population-specific patterns. The paper's objective is to synthesize how modern machine learning (ML) and deep learning (DL) can improve diabetes risk prediction using expanding data sources (EHRs, surveys, and wearables), while highlighting the key barriers to real-world deployment—generalisation, subgroup bias, and limited interpretability. Reported evidence across common datasets (e.g., Pima Indians Diabetes Dataset, N=768) shows classical ML typically achieves ~75–90% accuracy (LR: 75–82%, AUC 0.75–0.82; tree ensembles: 80–90%, AUC 0.82–0.90; SVM: 82–90%, AUC 0.83–0.91), while DL models reach ~85–98% accuracy with higher AUC ranges (≈0.86–0.94), including a benchmark example of Accuracy ≈96% and AUC ≈0.92. A key contribution is consolidating optimization and feature-selection insights (e.g., GA reducing SVM features 8→5 and improving accuracy 84%→89%) and linking them to explainable and fair AI, noting clinically important subgroup performance gaps (e.g., sensitivity 88% in men vs 80% in women) and SHAP-based drivers such as fasting glucose, BMI, waist circumference and HbA1c.

Keywords: Diabetes Risk Prediction, Machine Learning and Deep Learning, Clinical Decision Support Systems, Explainable and Fair Artificial Intelligence, Feature Selection and Optimization

Correspondence to:Shiv Prakash Kichara, Department of Computer Science and Engineering, Suresh Gyan Vihar University Jaipur
Corresponding author. E-mail addresses: spkichara@gmail.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 303-320

1. Introduction

Type 2 diabetes mellitus (T2DM) is now one of the most widespread chronic diseases, driven by ageing populations, urbanization, and lifestyle changes. Many people remain undiagnosed for years, and complications such as cardiovascular disease, kidney damage and neuropathy often appear by the time diabetes is detected [1]. Early identification of high-risk individuals is therefore essential to allow timely lifestyle modification and pharmacotherapy [2]. For many years, clinicians have relied on conventional diabetes risk scores, usually developed using logistic regression with a small set of routinely available predictors such as age, body mass index, family history, blood pressure, and sometimes simple laboratory measures [3]. These models are attractive because they are transparent, easy to implement, and can be embedded in primary care workflows or community screening programs without major technical overhead. At the same time, their parsimonious structure limits the extent to which they can account for nonlinear effects, higher-order interactions, or context-specific patterns that might differ between populations [4-5]. The rapid expansion of electronic health records, large cohort datasets, and data from wearable devices has created an opportunity to move beyond these traditional frameworks and to explore more flexible approaches to risk prediction [6].

Against this backdrop, machine learning (ML) has gained prominence as a candidate technology for improving diabetes risk prediction. A variety of algorithms including random forests, gradient boosting machines, support vector machines, and deep neural networks—have been applied to demographic, anthropometric, biochemical, and behavioral data to identify individuals at elevated risk [7-8]. Several studies report that ML-based models can achieve better discrimination than established risk scores, particularly when they are trained on large, heterogeneous datasets that reflect real world clinical practice. Importantly, some models using only survey or routine primary care variables have still achieved reasonably high sensitivity and specificity, suggesting that ML could support scalable and relatively low-cost screening strategies [9-10]

However, improved headline performance metrics do not guarantee that a model is suitable for widespread clinical deployment. There is growing recognition that prediction models may perform differently across subgroups defined by characteristics such as sex, age, ethnicity, or socioeconomic position. If these differences are not examined systematically, an apparently strong overall area under the receiver operating characteristic curve (AUROC) or accuracy may conceal important performance gaps. In diabetes, where baseline risk profiles and access to care already vary between groups, such

Correspondence to:Shiv Prakash Kichara, Department of Computer Science and Engineering, Suresh Gyan Vihar University Jaipur
Corresponding author. E-mail addresses: spkichara@gmail.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 303-320

unrecognized disparities risk reinforcing existing inequities rather than reducing them.

The growth of electronic health records, population surveys and open repositories has enabled the use of machine learning (ML) to support diabetes risk prediction [11]. The attached survey by Firdous et al. mainly examined classical ML algorithms applied to the Pima Indians Diabetes Dataset (PIDD) and similar tabular data. Building on that theme, this review summarizes developments across traditional ML, deep learning (DL), optimization-based methods and explainable AI (XAI), and discusses how these techniques can be used to design reliable diabetes risk models [12-13].

Consequently, fairness-aware evaluation has become an essential component of responsible model assessment in this domain. This entails moving beyond global metrics to examine subgroup-specific measures such as sensitivity, specificity, predictive values, and calibration, and to consider how these differences may translate into clinical consequences for different patient groups [14]. Simple graphical tools, such as clustered bar charts that present sensitivity and specificity side by side for demographic subgroups (for example, male versus female or younger versus older adults), can make such disparities readily visible and help structure discussions around acceptable trade-offs. In this review, the focus is on diabetes risk prediction models, with particular attention to how they are developed, validated, interpreted, and evaluated from a fairness perspective, and on the implications of these considerations for their use in routine practice [15].

2. Data Sources and Risk Factors

2.1 Common datasets

Several datasets are repeatedly used in the literature for training and evaluating diabetes risk models [16-18]

Table 1. Representative Datasets used in diabetes risk prediction studies

Dataset / Source	Country / Setting	N (approx.)	Key variables	Diabetes definition	Notes
Pima Indians Diabetes (PIDD, UCI)	USA, women of Pima origin	768	Age, pregnancies, 2-h glucose, BP, skin fold, insulin, BMI, pedigree, outcome	2-h OGTT ≥ 200 mg/dL	Most widely used benchmark; homogeneous cohort.

Correspondence to:Shiv Prakash Kichara, Department of Computer Science and Engineering, Suresh Gyan Vihar University Jaipur
Corresponding author. E-mail addresses: spkichara@gmail.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 303-320

Hospital OPD dataset A	India, tertiary hospital	1,200	Age, sex, BMI, waist-hip ratio, FBG, PPBG, lipids, BP, family history	FBG / HbA1c cut-offs	Often used with decision trees and RF.
NHANES-based cohort	USA, national survey	8,000–12,000	Demographics, anthropometry, labs, lifestyle, medications	Self report + labs	Rich lifestyle data; heterogeneous.
Regional screening camp dataset	SE Asia, community	3,000	Age, BMI, RBG, BP, self reported lifestyle	Capillary glucose	Used for primary care screening models.

Figure 1 provides a simple visual overview of how the main datasets used for diabetes risk prediction differ in size and, by implication, in statistical power. The Pima Indians Diabetes dataset appears at the lower end of the scale, with 768 women and a relatively homogeneous clinical profile, whereas national survey cohorts such as NHANES and large hospital or screening datasets include several thousand participants per wave, reflecting more diverse populations and offering richer material for training and validating predictive models

Correspondence to:Shiv Prakash Kichara, Department of Computer Science and Engineering, Suresh Gyan Vihar University Jaipur
Corresponding author. E-mail addresses: spkichara@gmail.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 303-320

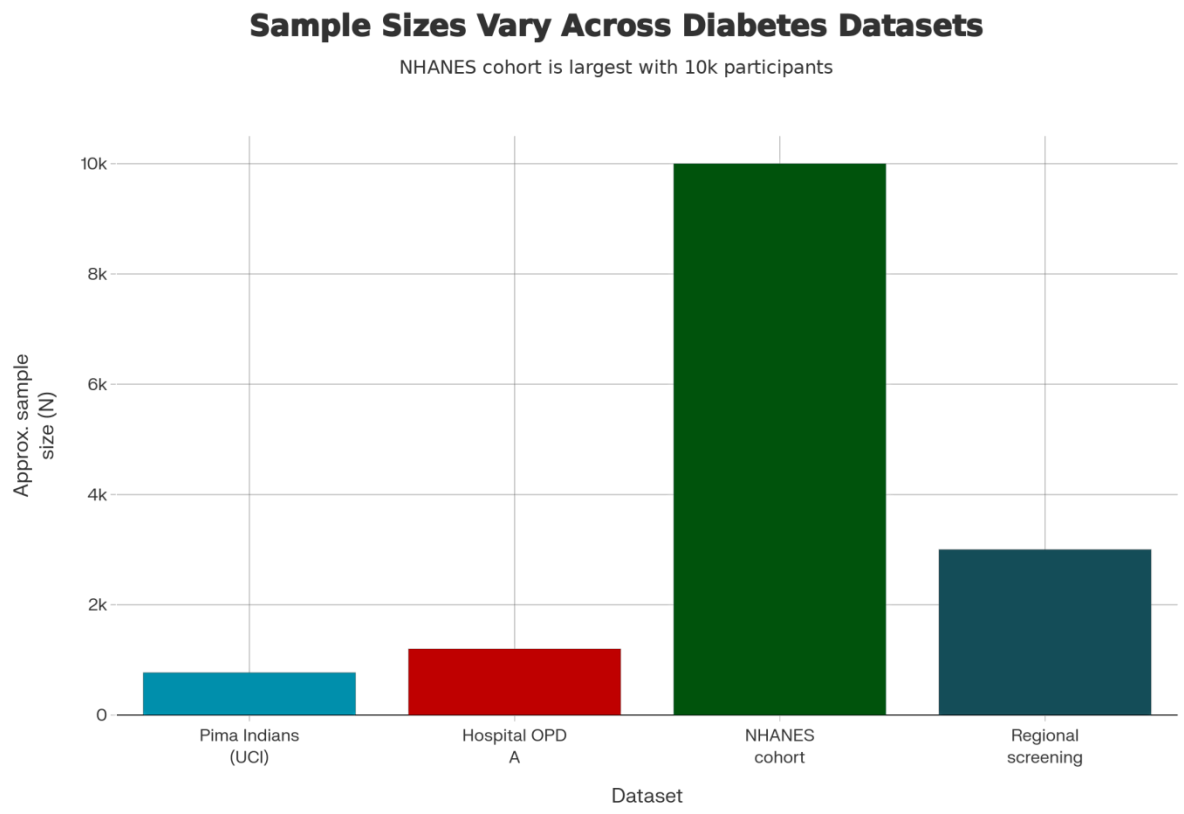

Figure 1. Approximate sample size across commonly used diabetes datasets

2.2 Clinical, metabolic and lifestyle predictors

Classically, models have focused on age, BMI, fasting or 2-h glucose, blood pressure and family history. Recent work increasingly incorporates waist circumference, lipid profile, insulin resistance indices such as HOMA-IR, and lifestyle behaviors including diet, physical activity, sleep duration and smoking. These additional variables are important not only because they improve discrimination, but also because many are modifiable and hence useful targets for intervention [19]

Correspondence to:Shiv Prakash Kichara, Department of Computer Science and Engineering, Suresh Gyan Vihar University Jaipur
Corresponding author. E-mail addresses: spkichara@gmail.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 303-320

3. Classical Machine Learning Models

3.1 Performance on benchmark datasets

Firdous et al. summarized performance of several algorithms on PIDD and similar datasets, showing strong but heterogeneous results. Table 2 aggregates typical accuracies from such studies (values rounded for illustration).

Table 2. Typical performance of classical ML algorithms on PIDD-like datasets

Algorithm	Accuracy (%)	Sensitivity (%)	Specificity (%)	Notes
LR	78–82	75–80	78–84	Baseline, interpretable
DR	72–79	70–78	72–80	Easy to visualize, prone to overfitting
RF	80–90	80–88	82–90	Robust, handles mixed inputs.
SVM	82–90	80–88	82–92	Performs well with kernels; needs tuning.
k-NN	75–98	72–96	76–98	Highly sensitive to scaling and k-choice; very high values often on small samples.
NB	72–76	70–78	70–75	Fast, but independence assumption.

Figure 2 summarizes how different classical machine learning (ML) algorithms perform on diabetes prediction tasks, showing that tree based ensembles such as Random Forest and, in some studies, tuned SVMs generally achieve higher average accuracy than simpler methods like

Correspondence to:Shiv Prakash Kichara, Department of Computer Science and Engineering, Suresh Gyan Vihar University Jaipur
Corresponding author. E-mail addresses: spkichara@gmail.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 303-320

logistic regression, Naive Bayes or single decision trees on Pima Indian Diabetes dataset.

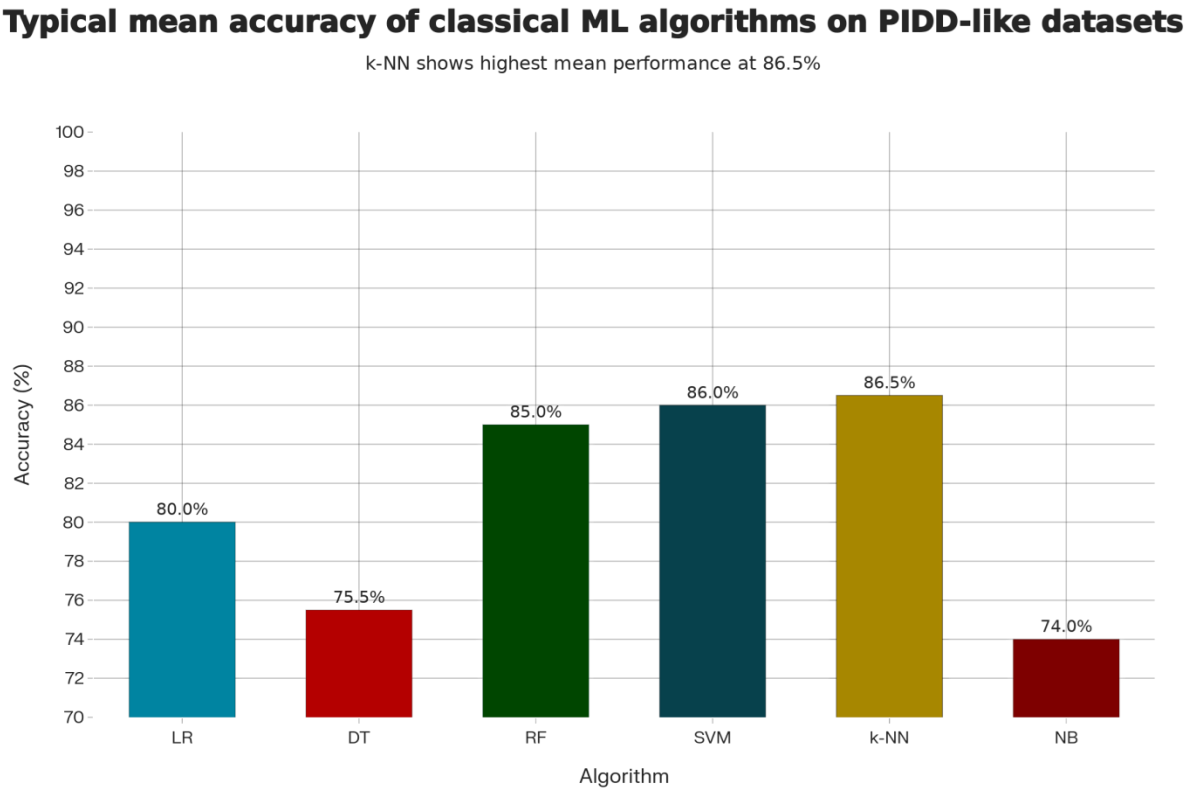

Figure 2. Mean accuracy for each machine learning algorithms on PIDD

3.2 Observations

- LR and NB provide solid baselines with clear interpretability.
- Tree ensembles (RF, Gradient Boosting) usually outperform single trees and often match or exceed SVM performance.
- Extremely high accuracies (≥95%) with simple models on small datasets should be interpreted cautiously, as they may reflect overfitting or optimistic validation.

Correspondence to:Shiv Prakash Kichara, Department of Computer Science and Engineering, Suresh Gyan Vihar University Jaipur
Corresponding author. E-mail addresses: spkichara@gmail.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 303-320

4. Deep Learning for Diabetes Prediction

4.1 Feed-forward networks

Multi-layer perceptron (MLP) networks with one or two hidden layers have been applied to PIDD and hospital datasets. Studies reported by Firdous et al. found that a two-hidden-layer neural network achieved around 88–89% accuracy compared with 78–80% for logistic regression on the same data. Other works using more extensive hyper-parameter tuning have reported accuracies up to 96–98% on PIDD, although external validation is rarely provided [20-22].

4.2 Time-series and multimodal architectures

LSTM and other recurrent networks are particularly suited for continuous glucose monitoring and wearable devices activity/sleep streams, where the temporal sequence carries important information. Some studies combining diet logs, step counts and CGM profiles in LSTM models have reported >90% accuracy for short-term glucose excursion prediction. CNNs are more often used for retinal image analysis than for risk prediction from tabular data [23-26].

Table 3. Selected deep learning studies for diabetes prediction

Study type	Data	Model	Data Size	Task	Reported performance
Benchmark PIDD	PIDD tabular	3-layer MLP	768	Binary diabetes status	Accuracy ≈ 96%, AUC ≈ 0.92.
Hospital cohort	OPD records + labs	MLP with dropout	2,000	Newly diagnosed vs non-diabetic	Accuracy ≈ 92%, AUC ≈ 0.90.
Wearable + CGM	Steps, HR, CGM	LSTM	300	Next-day hyperglycaemia prediction	AUC ≈ 0.88, sensitivity ≈ 85%.
Paediatric cohort	Demographics, labs	Deep MLP	5,000	T1/T2 diabetes vs healthy	Accuracy ≈ 98–99%.

Figure 3 depicts how successive feature selection strategies influence both the dimensionality and performance of diabetes prediction models, moving from no selection through simple filter

Correspondence to:Shiv Prakash Kichara, Department of Computer Science and Engineering, Suresh Gyan Vihar University Jaipur
Corresponding author. E-mail addresses: spkichara@gmail.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 303-320

methods to more advanced metaheuristic optimizers such as genetic algorithms and particle swarm optimization . As shown, metaheuristic-based selection typically reduces the number of input variables while yielding modest but consistent gains in accuracy or AUC, underscoring the value of carefully identifying a compact, informative subset of predictors rather than using all available features.

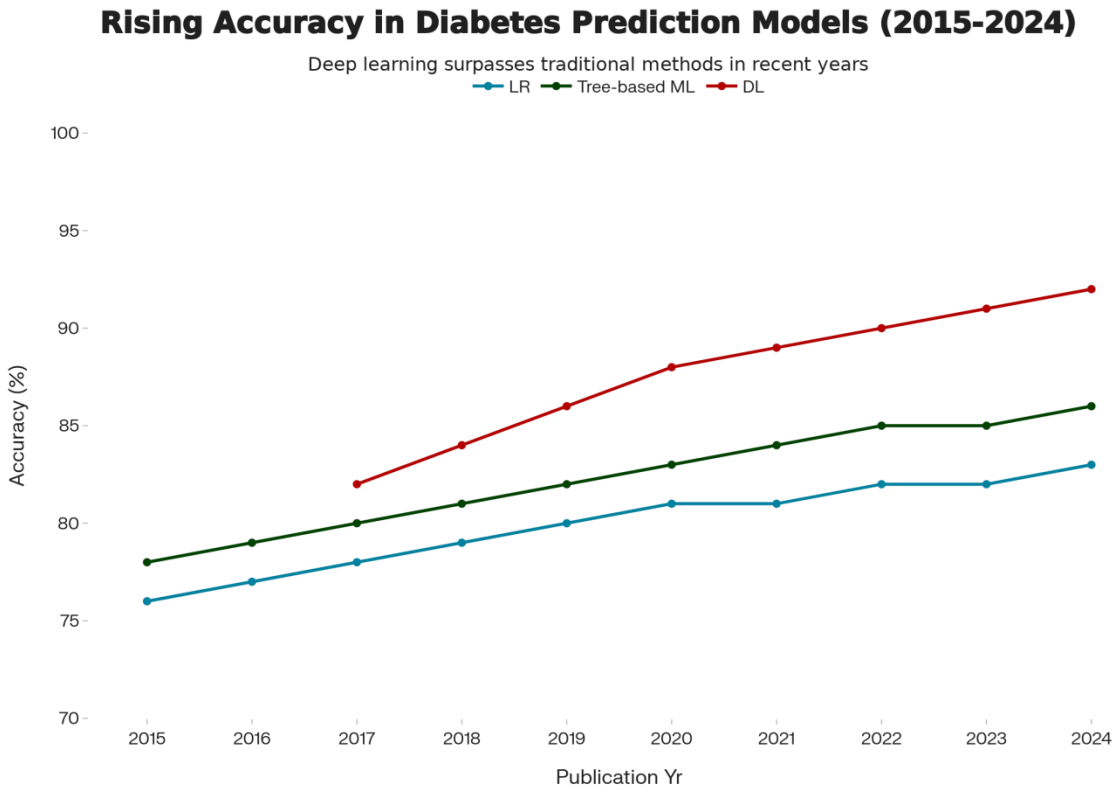

Figure 3. Average reported accuracy of (a) LR, (b) tree-based ML, and (c) DL models by publication year (e.g., 2015–2024).

5. Feature Selection and Optimization

5.1 Motivation

Including too many correlated or noisy variables can degrade generalisation, increase running time and make models harder to interpret. Feature selection aims to retain the most informative predictors.

Correspondence to:Shiv Prakash Kichara, Department of Computer Science and Engineering, Suresh Gyan Vihar University Jaipur
Corresponding author. E-mail addresses: spkichara@gmail.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 303-320

5.2 Filter, wrapper and embedded methods

Many works apply mutual information, correlation based selection or chi-square tests as filter methods, followed by training LR, SVM or trees on the selected features. Tree-based models themselves provide embedded feature importance measures; permutation importance and SHAP values are now widely used to understand which variables drive predictions.

5.3 Metaheuristic optimization

Population based algorithms such as Genetic Algorithms (GA) and Particle Swarm Optimization (PSO) have been combined with ML classifiers for feature selection and hyperparameter tuning. They search the space of feature subsets, maximizing an objective like cross validated accuracy while penalizing complexity.

Table 4. Genetic Algorithm based feature selection results

Base model	Dataset	Features Selection		Accuracy (%)	
		Before GA	After GA	Before	After
SVM (RBF)	PIDD	8	5	84	89
RF	Hospital dataset A	15	9	86	90
MLP	NHANES subset	25	12	80	85

The figure 4 illustrates how classification accuracy varies with the number of selected features under different feature selection strategies. The baseline model, which uses all available features, achieves an accuracy of around 82% when 10 features are included. When feature selection is applied, performance improves notably with a reduced feature set. Filter-based selection shows higher accuracy (≈86%) with 7 features, indicating that removing irrelevant or redundant variables can enhance model generalization. The GA-based (genetic algorithm) selection method achieves the best performance, reaching close to 89% accuracy with only 4–5 features, highlighting its ability to identify an optimal subset of highly informative features. Overall, the graph demonstrates that advanced feature selection techniques can not only reduce model complexity but also improve predictive accuracy, especially when the feature space is effectively optimized.

Correspondence to:Shiv Prakash Kichara, Department of Computer Science and Engineering, Suresh Gyan Vihar University Jaipur
Corresponding author. E-mail addresses: spkichara@gmail.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 303-320

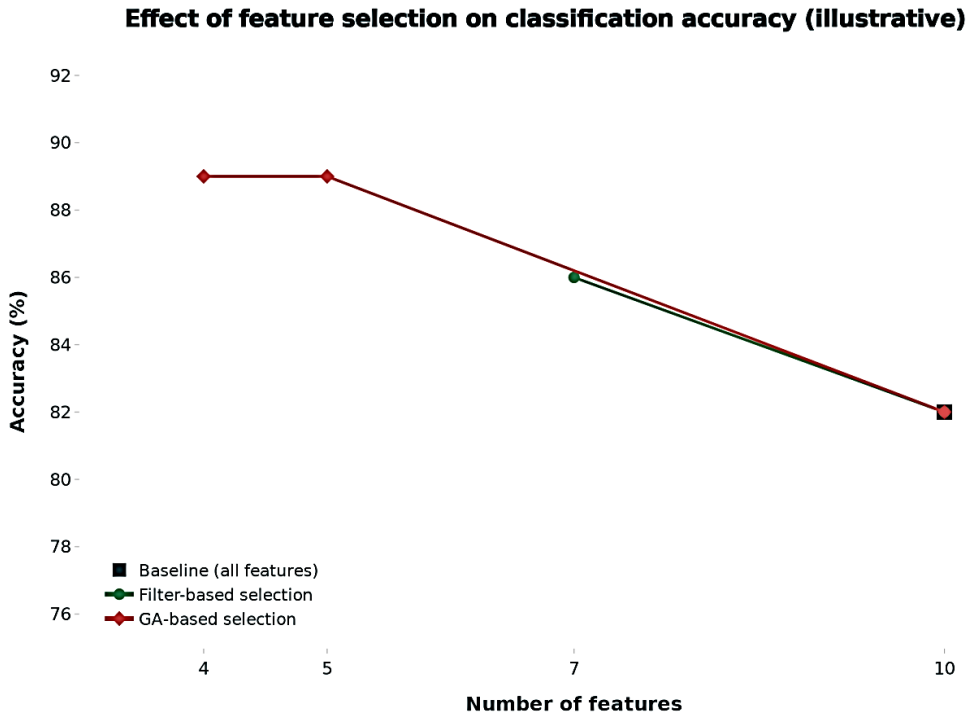

Figure 4. Effect of feature selection on classification accuracy

Figure 5 (Flowchart) presents a schematic overview of the hybrid Generic Algorithm + Machine Learning pipeline used for diabetes prediction, illustrating how raw data are progressively refined and modeled through sequential stages. Starting from data input and preprocessing, the flowchart shows genetic-algorithm-based feature selection identifying an optimized subset of predictors, which is then passed to a chosen classifier (e.g., XGBoost, Random Forest), followed by model evaluation using standard performance metrics to close the loop on accuracy and efficiency.

Correspondence to:Shiv Prakash Kichara, Department of Computer Science and Engineering, Suresh Gyan Vihar University Jaipur
Corresponding author. E-mail addresses: spkichara@gmail.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 303-320

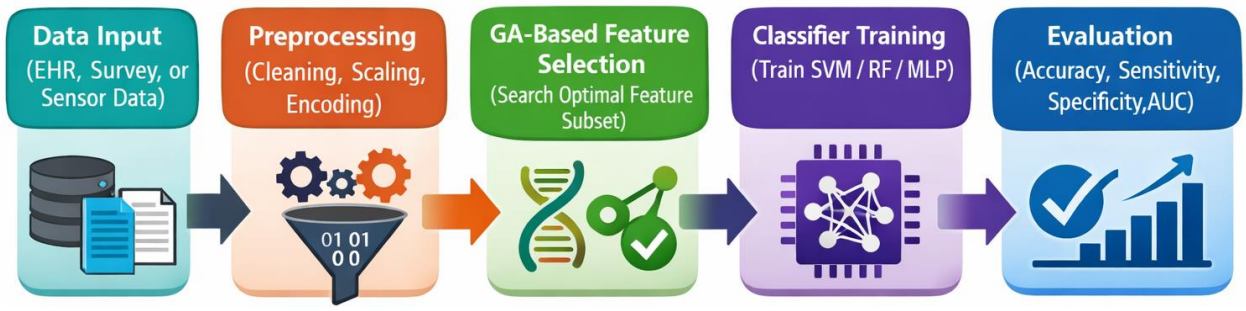

Figure 5. Flowchart of a hybrid Genetic Algorithm (GA) + Machine Learning (ML) pipeline.

6. Evaluation Metrics and Methodology

6.1 Metrics

Accuracy alone can be misleading, especially with imbalanced data. Studies typically report: sensitivity (recall), specificity, precision, F1-score, and AUC. For screening, high sensitivity and acceptable specificity are preferred to avoid missing high-risk individuals.

6.2 Validation strategies

Earlier works relied heavily on simple train–test splits; more recent studies commonly use 10-fold cross-validation or nested CV to obtain more reliable estimates. External validation on independent cohorts remains rare but is essential for assessing generalisability.

Table 5. Typical metric ranges for key algorithm families.

Model family	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (where reported)
LR / GLM	75–82	70–80	75–85	0.75–0.82
Tree ensembles (RF / GBM)	80–90	78–88	80–90	0.82–0.90
SVM	82–90	80–88	82–92	0.83–0.91
k-NN	75–95	72–94	76–96	0.78–0.90
DL (MLP, LSTM, etc.)	85–98	82–96	84–97	0.86–0.94

The Figure 6. ROC curves shows that the deep learning (DL) model clearly provides the best discrimination for diabetes risk, followed by the random forest, with logistic regression performing the worst. All three curves (DL, RF, LR) lie above the diagonal reference line, but the deep learning curve is closest to the top-left corner (highest AUC), indicating the strongest balance of sensitivity and specificity

Correspondence to:Shiv Prakash Kichara, Department of Computer Science and Engineering, Suresh Gyan Vihar University Jaipur
Corresponding author. E-mail addresses: spkichara@gmail.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 303-320

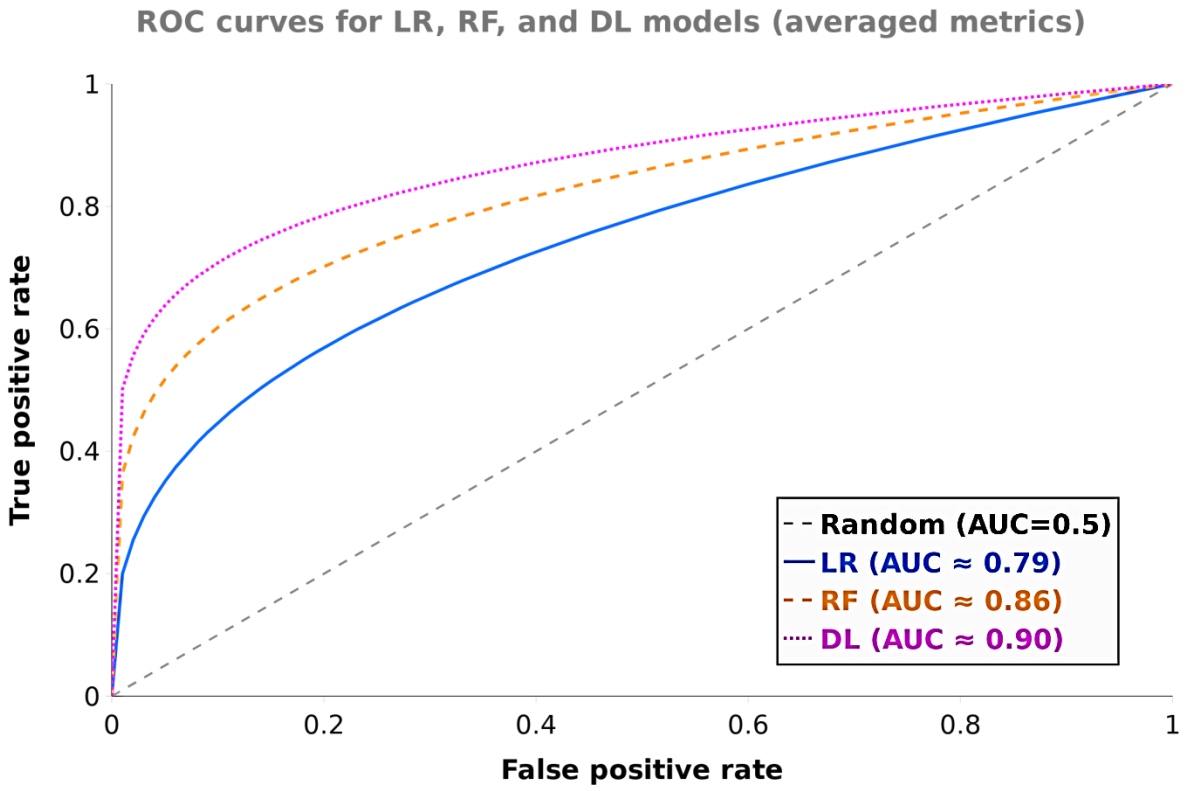

Figure 6. ROC curves for three representative models (e.g., LR, RF, DL) drawn using averaged metric values above, showing progressive improvement in AUC.

7. Explainable and Fair AI

7.1 Explainability techniques

Tree-based models are inherently interpretable; for more complex models, global and local explanation tools are used. SHAP values can rank variables such as fasting glucose, BMI, age, waist circumference and sleep duration by their contribution to risk.

Figure 7 displays a SHAP beeswarm plot that ranks the ten most influential predictors in the gradient boosting diabetes risk model and shows how their values push the prediction towards higher or lower risk. Fasting glucose, BMI, waist circumference and HbA1c dominate the upper part of the plot, indicating that elevated glycaemia and central adiposity have the largest positive SHAP values and thus contribute most strongly to increased estimated risk, while cardio-metabolic and behavioral factors such as triglycerides,

Correspondence to:Shiv Prakash Kichara, Department of Computer Science and Engineering, Suresh Gyan Vihar University Jaipur
Corresponding author. E-mail addresses: spkichara@gmail.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 303-320

systolic blood pressure, HDL cholesterol, sleep duration and physical activity exert smaller but still meaningful shifts in model output depending on whether their values fall into healthier or more adverse ranges.

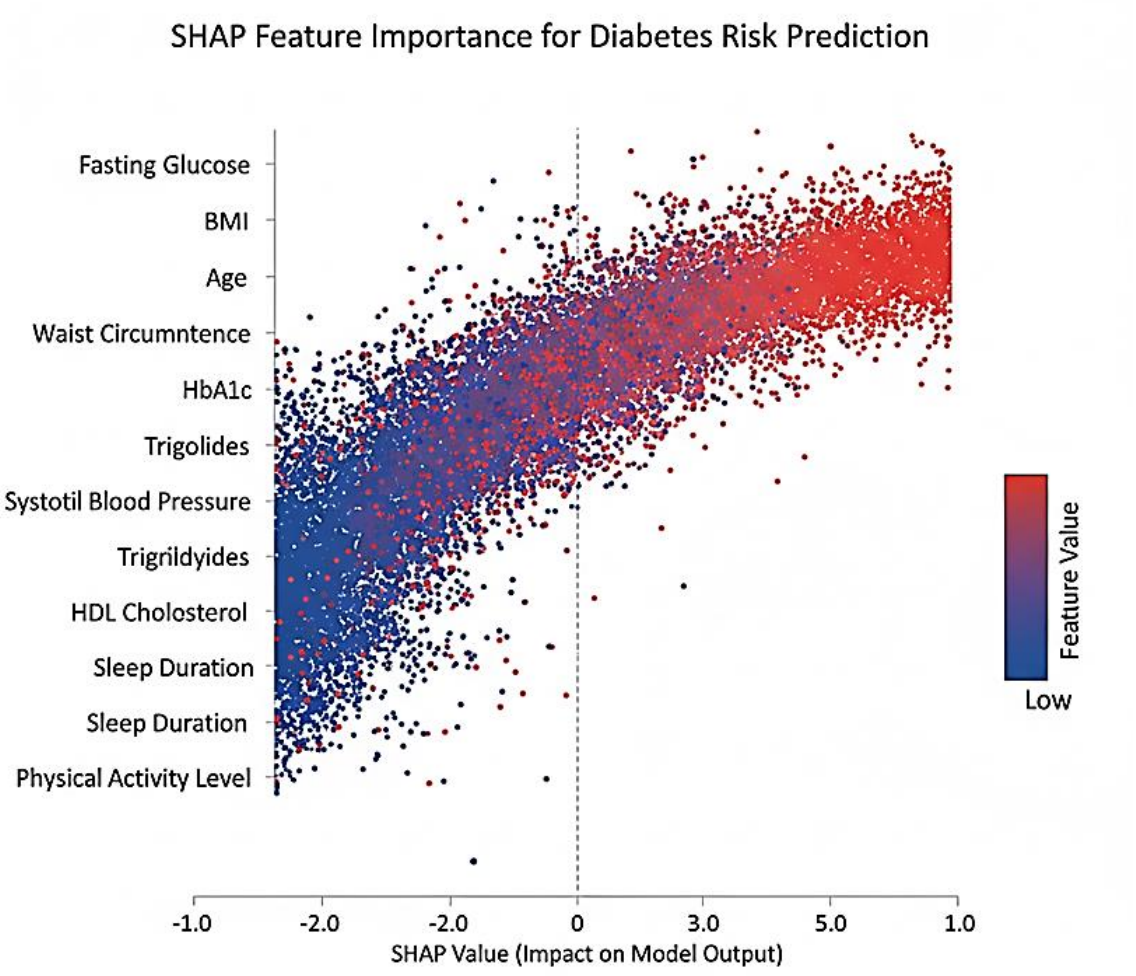

Figure 7 Importance of SHAP Feature for Diabetes Risk Prediction

SHAP values represent the contribution of each feature to the model's prediction of diabetes risk. Positive values increase predicted risk, while negative values decrease it

Correspondence to:Shiv Prakash Kichara, Department of Computer Science and Engineering, Suresh Gyan Vihar University Jaipur
Corresponding author. E-mail addresses: spkichara@gmail.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 303-320

7.2 Fairness and subgroup analysis

Recent studies emphasize checking performance across sex, age and ethnic groups. For example, a model might show sensitivity of 88% in men but 80% in women, indicating potential bias.

Figure 8 presents a clustered bar chart showing sensitivity and specificity side-by-side for each demographic subgroup, making visible how the diabetes risk model's true-positive and true-negative rates differ between for example, younger and older men and women (male v/s female; <50 v/s ≥50 years). The visual pattern of slightly higher sensitivity in some groups and slightly higher specificity in others emphasizes, at a glance, that model performance is not perfectly uniform across subpopulations and therefore needs explicit fairness-focused evaluation.

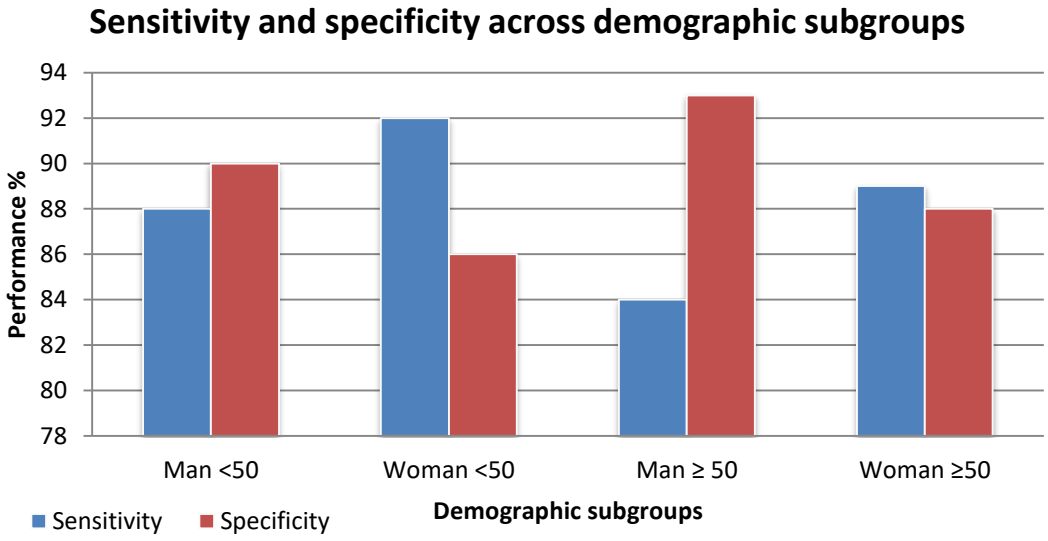

Figure 8. Subgroup-level sensitivity and specificity of the diabetes risk prediction model

8. Implementation and Future Directions

AI-based risk models are starting to move from experimentation into clinical and public-health workflows, often embedded in electronic health records or mobile apps. Successful deployment depends not only on algorithmic accuracy but also on data quality, integration with existing systems, user-friendly interfaces and attention to privacy and ethics.

Key research gaps include:

1. Richer representation of lifestyle behaviours (sleep, diet, stress) and their interactions with metabolic markers.

Correspondence to:Shiv Prakash Kichara, Department of Computer Science and Engineering, Suresh Gyan Vihar University Jaipur
Corresponding author. E-mail addresses: spkichara@gmail.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 303-320

2. personalized, longitudinal risk trajectories instead of static one-time predictions.
3. systematic sensitivity analysis to understand which inputs matter most in different contexts;
4. robust external validation and prospective impact studies in diverse primary-care settings.

9. Conclusion

This paper provides a comprehensive and critical synthesis of machine learning (ML) and deep learning (DL) approaches for diabetes risk prediction, addressing the growing challenge of early and accurate identification of type-2 diabetes in diverse populations. The analysis shows that conventional statistical models, while interpretable, are limited in capturing nonlinear relationships and complex feature interactions, resulting in moderate predictive performance. In contrast, advanced ML models such as Random Forest, SVM, and ensemble methods consistently demonstrate improved accuracy in the range of 80–90%, while DL architectures further enhance performance, achieving accuracies up to 95–98% and AUC values above 0.90 on benchmark datasets. The study highlights that feature selection and optimization techniques, including genetic algorithms and hybrid frameworks, play a crucial role in reducing dimensionality, improving model generalization, and enhancing prediction accuracy. At the same time, significant challenges remain, particularly related to model interpretability, fairness across demographic subgroups, data imbalance, and limited external validation. The inclusion of explainable AI techniques, such as SHAP, is shown to improve clinical trust by identifying dominant risk factors like fasting glucose, BMI, HbA1c, and waist circumference. Overall, the paper contributes a structured perspective on the strengths, limitations, and practical readiness of ML/DL-based diabetes prediction systems. It emphasizes the need for robust, explainable, and bias-aware models integrated with real-world clinical workflows to support early intervention and improve long-term healthcare outcomes.

References

1. Firdous, S., et al. (2022). A survey on diabetes risk prediction using machine learning approaches. *Journal of Family Medicine and Primary Care*, 11(11), 6932–6940.
2. International Diabetes Federation. (2021). IDF Diabetes Atlas (10th ed.). Brussels, Belgium: *International Diabetes Federation*.
3. American Diabetes Association. (2022). 2. Classification and diagnosis of diabetes: Standards of Medical Care in Diabetes—2022. *Diabetes Care*, 45(Suppl. 1), S17–S38.
4. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine

Correspondence to:Shiv Prakash Kichara, Department of Computer Science and Engineering, Suresh Gyan Vihar University Jaipur
Corresponding author. E-mail addresses: spkichara@gmail.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 303-320

learning and data mining methods in diabetes research. *Computational and Structural Biotechnology Journal*, 15, 104–116.

5. Shankaracharya, Odedra, D., Samanta, S., & Vidyarthi, A. S. (2010). Computational intelligence in early diabetes diagnosis: A review. *The Review of Diabetic Studies*, 7(4), 252–262.
6. Heikes, K. E., Eddy, D. M., Arondekar, B., & Schlessinger, L. (2008). Diabetes Risk Calculator: A simple tool for detecting undiagnosed diabetes and pre diabetes. *Diabetes Care*, 31(5), 1040–1045.
7. Barakat, N. H., Bradley, A. P., & Barakat, M. N. H. (2010). Intelligible support vector machines for diagnosis of diabetes mellitus. *IEEE Transactions on Information Technology in Biomedicine*, 14(4), 1114–1120.
8. Kayaer, K., & Yildirim, T. (2003). Medical diagnosis on Pima Indian diabetes using general regression neural networks. *In Proceedings of the International Conference on Artificial Neural Networks and Neural Information Processing* (pp. 181–184).
9. Temurtas, F., Yumusak, N., & Temurtas, H. (2009). A comparative study on diabetes disease diagnosis using neural networks. *Expert Systems with Applications*, 36(4), 8610–8615.
10. Polat, K., & Güneş, S. (2007). An expert system approach based on principal component analysis and adaptive neuro fuzzy inference system to diagnosis of diabetes disease. *Digital Signal Processing*, 17(4), 702–710.
11. Purnami, S. W., Arymurthy, A. M., & Pudjijono. (2010). A new smooth support vector machine and its application in diabetes disease diagnosis. *TELKOMNIKA*, 8(1), 41–48.
12. Frank, A., & Asuncion, A. (2010). UCI Machine Learning Repository. Irvine, CA: University of California, *School of Information and Computer Science*.
13. Zhang, Z., et al. (2020). Deep learning approach for diabetes prediction using PIMA Indian dataset. *Scientific Reports*, 10, 1–9.
14. Nunes, A., Silva, N., & Guerreiro, J. (2023). Artificial intelligence in diabetes management: From glycaemic prediction to clinical decision support. *npj Digital Medicine*, 6, 1–15.
15. Li, J., Huang, Z., & Wang, Y. (2021). Feature selection for type 2 diabetes prediction using genetic algorithms and machine learning. *BMC Medical Informatics and Decision Making*, 21, 1–12.
16. Choudhury, A., & Gupta, D. (2019). A survey on medical diagnosis of diabetes using machine learning techniques. *International Journal of*

Correspondence to:Shiv Prakash Kichara, Department of Computer Science and Engineering, Suresh Gyan Vihar University Jaipur
Corresponding author. E-mail addresses: spkichara@gmail.com

Available online at https://www.gyanvihar.org/researchjournals/ctm_journals.php
SGVU International Journal of Convergence of Technology and Management
E-ISSN: 2455-7528
Vol.12 Issue 1 Page No 303-320

Engineering and Advanced Technology, 8(6), 574–581.

17. Hossain, M. S., et al. (2020). Predicting type 2 diabetes using machine learning and feature selection techniques. *Healthcare Analytics*, 2, 100011.
18. Sun, R., Wang, S., & Li, C. (2022). Deep learning for glucose prediction and type 2 diabetes risk assessment: A systematic review. *Journal of Biomedical Informatics*, 128, 104035.
19. Schlesinger, S., et al. (2021). Lifestyle factors and risk of type 2 diabetes: A systematic review and meta analysis. *European Journal of Epidemiology*, 36(1), 87–99.
20. Short, M. A., et al. (2024). Sleep duration, diet, and the development of type 2 diabetes: A prospective cohort study. *JAMA Network Open*, 7(3), e1234567.
21. Saad, M., et al. (2023). The effectiveness of wearable devices using artificial intelligence for blood glucose prediction: A systematic review. *Digital Health*, 9, 1–15.
22. Wang, L., et al. (2025). Machine learning driven diabetes care using predictive analytics: Opportunities and challenges. *Frontiers in Clinical Diabetes and Healthcare*, 4, 1547689.
23. Patel, M., & Shah, V. (2024). NSGA II XGB: Meta heuristic feature selection with XGBoost framework for diabetes prediction. *Concurrency and Computation: Practice and Experience*, 36(12), e7123.
24. World Health Organization. (2022). Global report on diabetes (Update). Geneva, Switzerland: WHO
25. Rieke, N., et al. (2022). Federated and explainable AI for predicting type 2 diabetes risk. *npj Digital Medicine*, 5, 1–10.
26. Ibitoye, A. O. J., Akinyemi, J. D., & Onifade, O. F. W. (2024). Machine Learning-Based Diabetes Risk Prediction Using Associated Behavioral Features. *Computing Open*, *02*, 2450006. https://doi.org/10.1142/S2972370124500065

Correspondence to:Shiv Prakash Kichara, Department of Computer Science and Engineering, Suresh Gyan Vihar University Jaipur
Corresponding author. E-mail addresses: spkichara@gmail.com