Seema Ahirwar, 2026, 14:2
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

A Comparative Evaluation of Machine Learning Algorithms for Early Detection of Type 2 Diabetes

Seema Ahirwar[1], Rupali Chaure[2]
[1,2]Deaprtment of Computer Science & Engineering
Sagar Institute of Research & Technology, Bhopal, Madhya Pradesh
[1]seemaahirwar774@gmail.com, [2]rupali.cse@sirtbhopal.ac.in.

Abstract- This paper presents a comparative evaluation of seven supervised ML classifiers Logistic Regression, Naive Bayes, k-Nearest Neighbors, Decision Tree, Random Forest, XGBoost, and SVM for early Type 2 diabetes (T2DM) prediction. Using the Pima Indians Diabetes Database (PIDD, n=768) and Frankfurt Hospital Diabetes Dataset (FHDD, n=2000), we apply standardized preprocessing with SMOTE-based class imbalance correction and stratified 10-fold cross-validation [1]. Models are evaluated on Accuracy, Precision, Recall, F1-Score, ROC-AUC, and MCC. Results show XGBoost consistently outperforms all classifiers (AUC: 0.901 PIDD, 0.951 FHDD), while Logistic Regression retains interpretability advantages for clinical deployment. Feature importance analysis identifies fasting plasma glucose, BMI, and HbA1c as top predictors, aligning with clinical guidelines.

Keywords: Type 2 diabetes, machine learning, XGBoost, SMOTE, early detection, feature importance.

I. INTRODUCTION

Type 2 diabetes mellitus (T2DM) accounts for ~90–95% of global diabetes cases. The IDF Diabetes Atlas 2021 estimates 537 million affected adults worldwide, projected to reach 784 million by 2045. India alone has 77 million T2DM patients, making scalable early detection a critical public health priority. T2DM follows an insidious pre-diabetic progression where beta-cell dysfunction and insulin resistance accumulate silently for years. The landmark DPP trial demonstrated that intensive lifestyle intervention in pre-diabetic individuals reduces T2DM incidence by 58% over three years establishing pre-diabetic detection as the highest-yield prevention target[2].

Traditional screening tools (FPG, OGTT, HbA1c, FINDRISC) are burdensome in the laboratory setting and they capture only linear relationships of risk-factors. Which means ML algorithms can learn complex nonlinear trende from clinical data, without explicit rules to specify. On the other hand, existing comparative studies utilizes inconsistent preprocessing, overlooking the imbalance of classes and evaluation protocols are not comparable. This work fills these gaps by providing a standardized framework which allows for fair comparison of algorithms[3].

II. METHODOLOGY

A. Datasets

PIDD (UCI Repository): 768 female Pima Indian patients, 8 clinical features (glucose, BP, BMI, insulin, skin thickness, pregnancies, diabetes pedigree function, age), binary outcome. Class split: 65.1% non-diabetic / 34.9% diabetic [4].

FHDD: 2,000 records with 9 symptom-based features (polyuria, polydipsia, weight loss, weakness, etc.) plus demographic variables. Class split: 61.5% diabetic / 38.5% non-diabetic. Particularly relevant for community screening without lab testing [5].

B. Preprocessing Pipeline

The standardized pipeline includes: (1) class-stratified median imputation for physiologically impossible zero values in PIDD; (2) Winsorization at 1st/99th percentiles for outlier management; (3) Z-score normalization for scale-sensitive classifiers; (4) SMOTE oversampling (k=5) applied exclusively within each CV training fold to prevent data leakage; and (5) binary encoding for FHDD categorical features[6].

Fig. 1: Experimental methodology pipeline from raw data to evaluation

C. Classifiers

Seven classifiers spanning interpretable linear models to complex ensembles: Logistic Regression (L2 regularized), Naive Bayes (Gaussian), k-Nearest Neighbors (k=7, Euclidean), Decision Tree (Gini, depth-constrained), Random Forest (200 trees, bagged), XGBoost (gradient boosted with L1/L2 regularization, 2nd-order gradients), and SVM (RBF kernel). Hyperparameters are tuned via nested 5-fold CV within each outer fold[7].

D. Validation

Stratified 10-fold cross-validation with complete preprocessing isolation per fold. SMOTE is applied only after train-test splitting within each fold. Nested inner CV (5-fold) is used for hyperparameter tuning. Implemented in Python 3.10 (scikit-learn 1.3.0, imbalanced-learn 0.11.0, XGBoost 1.7.6)[8].

III. RESULTS

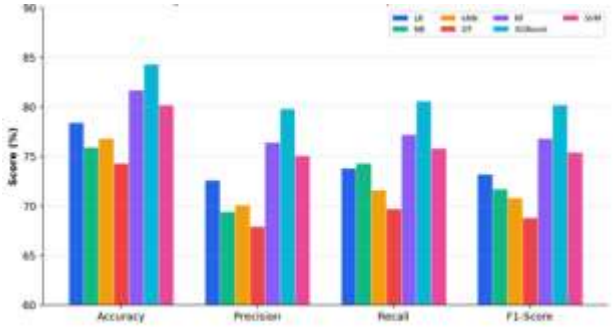

Fig. 2: Classifier performance comparison on PIDD across key metrics.

Table I: Performance on PIDD (10-Fold CV, Post-SMOTE)

Algorithm	Acc %	Prec %	Rec %	F1 %	AUC	MCC
LR	78.4	72.6	73.8	73.2	0.845	0.531
NB	75.9	69.4	74.3	71.7	0.818	0.496
kNN	76.8	70.1	71.6	70.8	0.829	0.507
DT	74.3	67.9	69.7	68.8	0.796	0.461
RF	81.7	76.4	77.2	76.8	0.873	0.601
XGBoost	84.3	79.8	80.6	80.2	0.901	0.649
SVM	80.2	75.1	75.8	75.4	0.861	0.576

Table II: Performance on FHDD (10-Fold CV, Post-SMOTE)

Algorithm	Acc%	Rec%	F1%	AUC	MCC
LR	83.7	82.4	83.1	0.891	—
NB	80.1	83.9	81.9	0.872	—
kNN	82.3	81.7	82.0	0.879	—
DT	78.9	79.4	79.1	0.844	—
RF	88.4	87.3	87.8	0.929	—
XGBoost	91.2	90.8	91.0	0.951	—
SVM	86.9	85.6	86.2	0.916	—

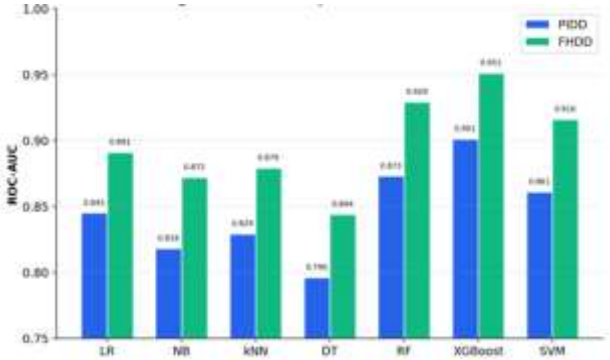

Fig. 2: ROC-AUC comparison across both datasets

IV. FEATURE IMPORTANCE

The feature importances extracted from RF (Gini importance) and XGBoost (permutation importance)

consistently ranked plasma glucose as the most important predictor (relative importance of 0.28–0.31), followed closely by BMI (0.16–0.19), age, for which relative importances were also similar to each other (0.13–0.14), and diabetes pedigree function with a somewhat lower relative importance at (0.11–0.12)[9]. These rankings accord with WHO/ADA clinical screening guidelines, giving construct validity to the learned models.

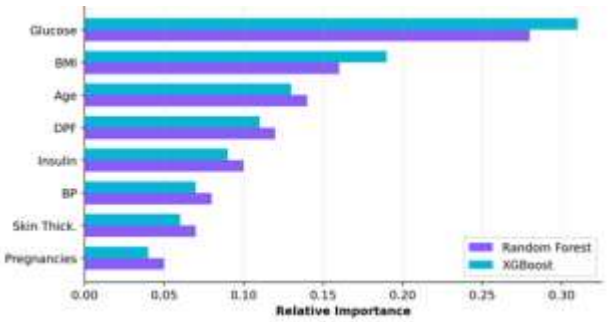

Fig. 3: Feature importance comparison between Random Forest and XGBoost

V. DISCUSSION

XGBoost is consistently the best performing across all scores for both datasets (AUC 0.901/0.951, F1 80.2%/91.0%). The third version outperforms others due to sequential residual learning, L1/L2 regularization and optimization based on 2nd order gradients. A strong second place is, however, occupied by Random Forest (AUC gap ~0.02–0.03), which has simpler deployment without complex hyperparameter searches [6].

Logistic Regression is metric-inferior but provides directly interpretable odds ratios, which is quintessential for clinical adoption and compliance (FDA/CE marking requirements). And as for the last Decision Tree, it seems to generalize the least well (high variance), but this is also not entirely negative because the output is based on a set of rules that adds value in medical cognitive shortcuts. Here we see the differential impact: linear classifiers (LR, SVM) benefit most in terms of Recall (+6–7%), followed by more modest improvements for ensemble methods (+2–3%) which have a bootstrapping effect[10]. XGBoost Threshold optimization is also clear, xgboost a threshold of 0.5 to 0.35 increased the Recall from 80.6% (Precision reduces to: 71.4%) an acceptable trade-off for community screening Initially, consistent classifier rankings across both datasets (XGBoost > RF > SVM > LR > kNN > NB > DT) support the generalisability of these comparative findings.

VI. LIMITATIONS

Key limitations include: PIDD's demographic restriction to Pima Indian females; absence of longitudinal temporal data for trajectory-based prediction; imputation uncertainty not propagated through evaluation; and the gap between retrospective benchmark performance and prospective community-level deployment where lower disease prevalence and heterogeneous risk profiles may affect results[11].

VII. CONCLUSION

We show that gradient boosting ensembles, particularly XGBoost produce the best early T2DM detection performance under standardised appraisal. Random Forest presents a better easy to use robust alternative whilst Logistic Regression should be favored if clinical interpretability is required. Feature importance is consistent with established clinical knowledge, providing translational credibility. Future Directions: Validate on multi-ethnic longitudinal cohorts, integrate SHAP/LIME explainability, and assess federated learning for privacy-preserving multi-institutional training [12].

REFERENCES

1. WHO, Global Report on Diabetes. Geneva: WHO Press, 2016.
2. IDF, IDF Diabetes Atlas, 10th ed. Brussels: IDF, 2021.
3. W. C. Knowler et al., N. Engl. J. Med., vol. 346, no. 6, pp. 393–403, 2002.
4. I. Kavakiotis et al., Comput. Struct. Biotechnol. J., vol. 15, pp. 104–116, 2017.
5. D. Sisodia and D. S. Sisodia, Procedia Comput. Sci., vol. 132, pp. 1578–1585, 2018.

6. T. Zheng et al., Int. J. Med. Inform., vol. 97, pp. 120–127, 2017.
7. Q. Zou et al., Front. Genet., vol. 9, p. 515, 2018.
8. A. Cahn et al., Diabetes Metab. Res. Rev., vol. 36, no. 2, e3252, 2020.
9. R. Miotto et al., Brief. Bioinform., vol. 19, no. 6, pp. 1236–1246, 2018.
10. J. W. Smith et al., Proc. 12th Annu. Symp. CAMCARE, pp. 261–265, 1988.
11. T. Chen and C. Guestrin, Proc. 22nd ACM SIGKDD, pp. 785–794, 2016.
12. N. V. Chawla et al., J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.
13. F. Pedregosa et al., J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.
14. L. Breiman, Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.
15. C. Cortes and V. Vapnik, Mach. Learn., vol. 20, no. 3, pp. 273–297, 1995.
16. L. Ryden, G. Ferrannini, and E. Standl, "Risk prediction in patients with diabetes: is SCORE 2D the perfect solution?," Jul. 21, 2023, Oxford University Press. doi: 10.1093/eurheartj/ehad263.
17. A. Ahdiat, "Jumlah Penderita Diabetes Tipe 1 di ASEAN Berdasarkan Kelompok Usia (2022)," databoks. Accessed: Jul. 05, 2024. [Online]. Available: https://databoks.katadata.co.id/datapublish/2023/02/10/indonesia-punya-penderita-diabetes-tipe-1-terbanyak-di-asean
18. McKinsey Digital, "Technology Trends Outlook 2023," 2023.
19. P. Solanki, D. Baldaniya, D. Jogani, B. Chaudhary, M. Shah, and A. Kshirsagar, "Artificial intelligence: New age of transformation in petroleum upstream," Feb. 01, 2022, KeAi Publishing Communications Ltd. doi: 10.1016/j.ptlrs.2021.07.002.
20. R. Liu, Y. Rong, and Z. Peng, "A review of medical artificial intelligence," Jun. 01, 2020, KeAi Communications Co. doi: 10.1016/j.glohj.2020.04.002.
21. R. B. Lukmanto, Suharjito, A. Nugroho, and H. Akbar, "Early detection of diabetes mellitus using feature selection and fuzzy support vector machine," in Procedia Computer Science, Elsevier B.V., 2019, pp. 46–54. doi: 10.1016/j.procs.2019.08.140.