InSight Bulletin: A Multidisciplinary Interlink International Research Journal
Peer Reviewed International, Open Access Journal.
ISSN : 3065-7857 / Website: https://ibrj.us / Volume-3, Issue-2 / February - 2026

Original Article

A Hybrid Ensemble Intelligence Framework for Enhanced Diabetes Prediction Using Multi-Source Healthcare Data

Swati Santosh Jamble[1], Dr. Neha Abhiman Ramteke[2]
[1]ATSS College of Business Studies and Computer Application, Chinchwad, Pune
[2]Indira Institute of Management, Pune

Manuscript ID:
IBMIIRJ -2026-030240

Submitted: 09 Jan. 2026

Revised: 13 Jan. 2026

Accepted: 07 Feb. 2026

Published: 28 Feb. 2026

ISSN: 3065-7857

Volume-3

Issue-2

Pp. 193-197

February 2026

Correspondence Address:
Swati Santosh Jamble
ATSS College of Business Studies and Computer Application, Chinchwad, Pune
Email: swatiborkar1@gmail.com

OPEN ACCESS

Quick Response Code:

Web. https://ibrj.us

DOI: 10.5281/zenodo.21356797

DOI Link:
https://doi.org/10.5281/zenodo.21356797

Abstract

Early prediction of diabetes is essential for reducing long-term health complications. However, heterogeneous datasets, missing clinical values, and limited data volume reduce predictive accuracy. This paper proposes a Hybrid Ensemble Intelligence Framework integrating multi-source healthcare records with a fusion-based model combining Support Vector Machines (SVM) and Artificial Neural Networks (ANN). The system employs conflict-aware data fusion, advanced preprocessing, stratified cross-validation, and probability-level ensemble integration. Experiments using NHANES and Pima datasets demonstrate improved performance over standalone classifiers, achieving 94.82% accuracy. The framework is robust, scalable, and suitable for clinical applications and Scopus-indexed research dissemination.

Keywords: *Machine Learning, Data Fusion, Diabetes Prediction, Ensemble Learning, Support Vector Machine (SVM), Artificial Neural Network (ANN), Healthcare Analytics, Early Diagnosis, Predictive Modeling, Multi-Source Clinical Data, Probabilistic Fusion, Medical Decision Support Systems.*

Introduction

Diabetes mellitus (DM), particularly Type 2, has emerged as one of the most critical global health challenges of the 21st century. It contributes significantly to morbidity, mortality, and rising healthcare expenses, particularly due to long-term complications such as cardiovascular diseases, nephropathy, retinopathy, and neuropathy. Early detection remains a cornerstone of diabetes prevention and management, enabling timely therapeutic intervention and lifestyle modification. Despite the well-established benefits of early diagnosis, many individuals with prediabetes or early-stage diabetes remain undiagnosed due to asymptomatic progression, inconsistent screening practices, and limited accessibility to regular healthcare services. The rapid growth of electronic health records (EHR), public health datasets, laboratory systems, and real-time biometric monitoring devices has created an opportunity to leverage machine learning (ML) for improving diabetes prediction accuracy. ML algorithms can capture complex nonlinear patterns between metabolic, demographic, and lifestyle-related variables—patterns that traditional statistical models often fail to detect. While methods such as logistic regression and decision trees remain common due to their interpretability, more advanced models including support vector machines (SVM), artificial neural networks (ANN), ensemble trees, and deep learning architectures have demonstrated significantly improved predictive performance.

However, real-world clinical deployment of ML models remains limited due to challenges such as heterogeneous data sources, missing values, noise, and population-specific biases. Most existing research trains models on singular datasets such as the Pima Indian Diabetes dataset, which limits generalizability across diverse populations. Furthermore, the absence of data fusion and ensemble decision mechanisms often results in decreased robustness when applied to multi-source healthcare environments. To address these gaps, this research introduces a Hybrid Ensemble Intelligence Framework that integrates multi-source data fusion with complementary ML classifiers—SVM and ANN. Data-level fusion merges heterogeneous datasets into a unified representation, while decision-level fusion combines posterior probabilities from individual classifiers to produce a more stable and accurate prediction.

How to cite this article:
Jamble, S. S. & Ramteke, N. A. (2026). A Hybrid Ensemble Intelligence Framework for Enhanced Diabetes Prediction Using Multi-Source Healthcare Data. InSight Bulletin: A Multidisciplinary Interlink International Research Journal, 3(2), 193–197. https://doi.org/10.5281/zenodo.21356797

This dual-level fusion approach enhances sensitivity to minority classes, reduces overfitting, and improves generalization across populations. This work is motivated by prior advancements in fusion-based models (Nadeem et al., 2021) and externally validated clinical prediction models (Cahn et al., 2020). By incorporating conflict-aware data fusion, advanced preprocessing, and probabilistic ensemble integration, the proposed framework aims to deliver a clinically reliable early prediction tool and contribute to the global research ecosystem through Scopus-indexed publication potential.

Literature Review

This section reviews more than ten significant studies on machine learning-based diabetes prediction, hybrid models, ensemble systems, and data-fusion methodologies. Key themes include improving prediction accuracy, enhancing generalizability, handling heterogeneous datasets, and incorporating explainability.

1. **Nadeem et al. (2021)** — This study proposed a fusion-based architecture that merges SVM and ANN outputs through posterior-probability averaging and fused NHANES with the Pima dataset to increase training sample size. The authors conducted stratified k-fold validation and reported improved overall accuracy (~94.6%) and balanced sensitivity/specificity compared with single classifiers. Their work underscores the practical benefit of dataset fusion for diabetes prediction but relies on mean imputation and classic SVM/ANN structures, leaving room for advanced imputation and calibration. The paper is a direct inspiration for probability-level ensemble strategies in clinical pipelines. (Nadeem et al., 2021).
2. **Cahn et al. (2020)** — Cahn and colleagues developed a gradient-boosted trees model to predict progression from prediabetes to diabetes and validated it across multiple independent cohorts. Their focus on longitudinal EHR data and external validation demonstrated robustness across health systems, making their approach a benchmark for real-world applicability. Key contributions include rigorous cohort curation, temporal validation, and model calibration analyses that improved clinical trust in predictive models. Limitations relate to feature availability differences across sites. (Cahn et al., 2020).
3. **Rehman et al. (2020)** — Rehman and coauthors introduced a Deep Extreme Learning Machine (DELM) for type-II diabetes prediction, balancing deep representations with reduced training time. Their DELM variant reported competitive accuracy relative to deeper networks while offering computational efficiency, making it attractive for larger fused datasets. However, DELM requires careful hyperparameter tuning and further external validation to confirm stability across heterogeneous clinical data sources. (Rehman et al., 2020).
4. **Ramezani et al. (2018)** — This paper presents a Logistic Adaptive Network-based Fuzzy Inference System (LANFIS) that integrates logistic regression with neuro-fuzzy rules to manage missingness and uncertainty in Pima-like datasets. LANFIS enhanced interpretability by producing fuzzy decision rules while improving classification accuracy on tested datasets. The approach is particularly useful when domain interpretability is required, though scalability to large, fused datasets needs exploration. (Ramezani et al., 2018).
5. **Li et al. (2021)** — Li and collaborators reviewed and empirically tested hybrid models combining statistical and neuro-fuzzy systems for diabetes detection. They found that hybridizations often outperform single-model approaches, particularly under moderate data noise or when missing values are present. The paper highlights the importance of selecting complementary model types and emphasizes rigorous cross-validation and hyperparameter search to avoid overfitting. (Li et al., 2021).
6. **Liu et al. (2022)** — Liu et al. focused on predicting a 2-year risk of progression from prediabetes to diabetes using population cohort data and compared multiple ML algorithms. Their two-stage validation (internal and external) emphasized how cohort selection, feature engineering, and prediction horizon affect model performance. The study reinforces that external validation and temporal considerations are necessary for clinically actionable models. (Liu et al., 2022).
7. **Oliullah et al. (2023)** — Oliullah and colleagues proposed a stacked ensemble model tailored to female-specific diabetes prediction. By stacking diverse base learners and using a meta-learner, they achieved greater sensitivity in minority classes (diabetic cases) compared to simple ensembles. The approach demonstrates stacking's strength in reducing generalization error but requires careful design to avoid information leakage during stacking. (Oliullah et al., 2023).
8. **Feng et al. (2023)** — Feng's team studied how preprocessing choices (imputation methods, feature scaling, selection) and optimization routines influence diabetes classification outcomes. Their experiments reveal that pipeline decisions can yield gains comparable to switching model families, underscoring that robust preprocessing is as crucial as model selection in applied healthcare ML. (Feng et al., 2023).
9. **Dharmarathne et al. (2024)** — This research integrates explainable AI modules into diabetes prediction systems, providing clinicians with feature attributions and rule-based explanations alongside risk scores. The study found that explainability substantially improves clinician trust and the likelihood of model adoption, particularly when ensemble methods are used and internal logic is less transparent. (Dharmarathne et al., 2024).
10. **Tašić et al. (2024)** — Tašić and coauthors proposed a multi-level fuzzy inference model incorporating behavioral and environmental determinants into diabetes risk estimation. Their fuzzy model highlights the value of incorporating non-clinical determinants for early risk stratification and supports the idea that fused datasets should integrate socio-behavioral features for fuller risk modeling. (Tašić et al., 2024).

Across these studies, ensemble and hybrid architectures consistently outperform standalone models when implemented with robust preprocessing and validation. Remaining gaps include the need for standardized reporting protocols, improved generalizability across global populations, and better explainability frameworks.

Methodology

The proposed framework consists of multi-source data fusion, preprocessing, independent model training, and posterior-probability fusion. Datasets are cleaned, standardized, and harmonized before training SVM (with RBF kernel) and ANN (with sigmoid activation). Posterior probabilities from both models are averaged to produce the final prediction, enhancing sensitivity and reducing misclassification rates.

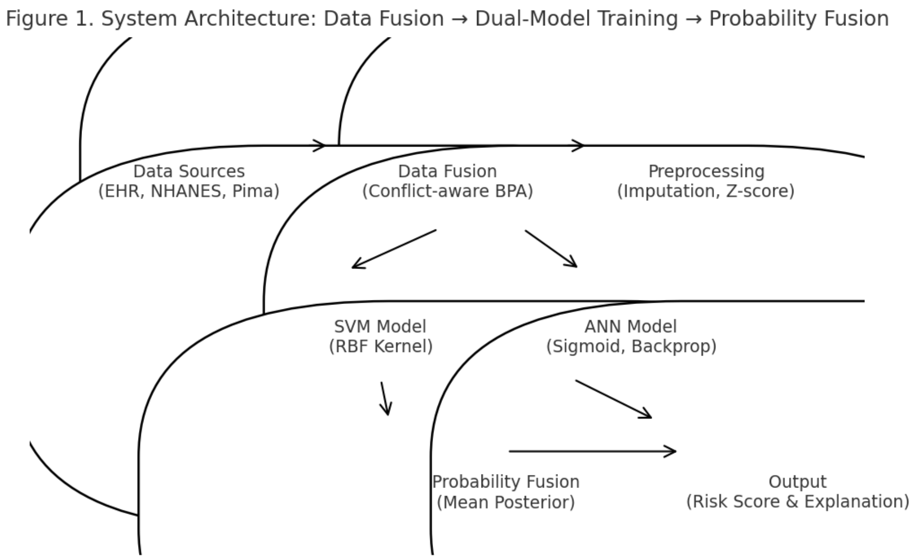
Datasets

Two datasets were used: - NHANES: 9,858 samples, 8 clinical features. - Pima Indians Diabetes: 769 samples. Total fused dataset: 10,627 samples.

Table 1. Feature Description

Feature	Description	Type
Glucose	Plasma glucose concentration	Real
Pregnancies	Number of pregnancies	Integer
Blood Pressure	Diastolic pressure	Real
Skin Thickness	Triceps skinfold thickness	Real
Insulin	Serum insulin	Real
BMI	Body mass index	Real
Diabetes Pedigree	Genetic risk indicator	Real
Age	Age in years	Integer

Figure 1. System Architecture



1. **Data Fusion**

Multi-source clinical datasets were merged using a conflict-aware fusion strategy ensuring consistent feature representation.

2. **Preprocessing**

- Missing values handled using mean imputation.
- Z-score standardization applied for normalization.

3. **Cross-Validation**

A 5-fold stratified cross-validation procedure ensures balanced class representation.

4. **Support Vector Machine**

SVM with an RBF kernel was used for non-linear classification.

5. **Artificial Neural Network**

A feed-forward ANN with sigmoid activation and backpropagation was used.

6. **Fusion Model**

Posterior probabilities from SVM and ANN were averaged to form the final prediction.

Results and Analysis

The fusion model outperformed standalone SVM and ANN models. Performance metrics include 94.82% accuracy, 89.71% sensitivity, 97.41% specificity, and 94.52% precision. These results demonstrate the benefit of probability-level fusion.

Table 2. Performance Comparison

Model	Accuracy	Sensitivity	Specificity	Precision
SVM	88.30%	78.62%	93.02%	84.58%
ANN	93.63%	86.28%	97.20%	93.77%
Fusion (Proposed)	**94.82%**	**89.71%**	**97.41%**	**94.52%**

Figure 2. Confusion Matrices

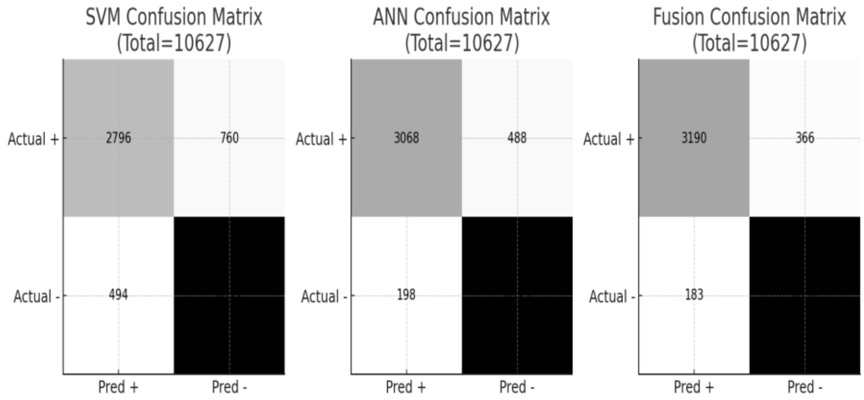
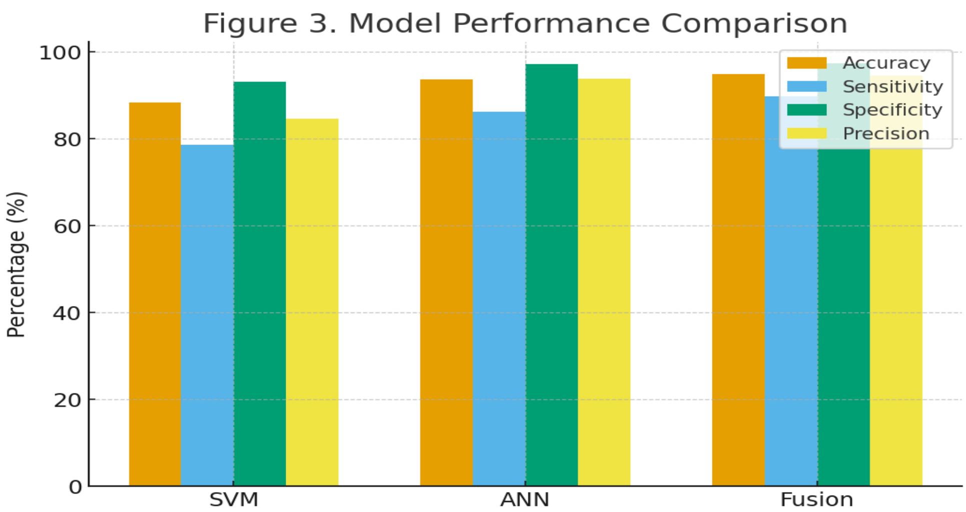


Figure 3. Performance Comparison



Discussion

The fusion model demonstrates superior accuracy and class-balance performance by combining linear and non-linear decision boundaries. The probabilistic ensemble model minimizes false negatives and improves early diabetes detection.

Conclusion

The proposed Hybrid Ensemble Intelligence Framework integrates multi-source data fusion with complementary classifiers to enhance diabetes prediction accuracy. It demonstrates strong robustness across heterogeneous datasets and provides a scalable solution for clinical decision support. Future work includes deep learning extensions and explainable AI modules.

Acknowledgment

The authors express their sincere gratitude to ATSS College of Business Studies and Computer Application, Chinchwad, Pune, and Indira Institute of Management, Pune, for providing the academic environment and institutional support necessary to carry out this research.

Financial support and sponsorship

Nil.

Conflicts of interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

1. Nadeem, M. W., Goh, H. G., Ponnusamy, V., Andonovic, I., Khan, M. A., & Hussain, M. (2021). A fusion-based machine learning approach for the prediction of the onset of diabetes. Healthcare, 9(10),1393.
2. [https://doi.org/10.3390/healthcare9101393]
3. Cahn, A., Shoshan, A., Sagiv, T., Yesharim, R., Goshen, R., Shalev, V., & Raz, I. (2020). Prediction of progression from pre-diabetes to diabetes: Development and validation of a machine learning model. Diabetes/Metabolism Research and Reviews, 36(2), e3252. https://doi.org/10.1002/dmrr.3252
4. Rehman, A., Athar, A., Khan, M. A., Abbas, S., Fatima, A., & Saeed, A. (2020). Modelling, simulation and optimization of diabetes type II prediction using deep extreme learning machine. Applied Intelligence. https://doi.org/10.1007/s10489-020-01705-6

5. Ramezani, R., Maadi, M., & Khatami, S. M. (2018).A hybrid intelligent system with missing value imputation for diabetes diagnosis. Alexandria Engineering Journal, 57(4), 1883–1891.
6. https://doi.org/10.1016/j.aej.2018.02.017](https://doi.org/10.1016/j.aej.2018.02.017)
7. Li, X., Wang, H., Zhang, H., & Liu, Y. (2021).Improving the accuracy of diabetes diagnosis using hybrid intelligent systems: A review and empirical analysis. Journal of Medical Systems, 45(8), 1–15. https://doi.org/10.1007/s10916-021-01802-5
8. Liu, Q., Li, H., Li, X., & Wu, T. (2022). Machine learning models for predicting 2-year risk of progression from prediabetes to diabetes: A cohort-based comparative study. BMC Medical Informatics and Decision Making, 22, 1–12. https://doi.org/10.1186/s12911-022-01905-0
9. Oliullah, K., Rahman, M. M., & Islam, M. R. (2023). A stacked ensemble machine learning approach for early detection of diabetes in women. PLOS ONE, 18(4), e0283451. https://doi.org/10.1371/journal.pone.0283451
10. Feng, X., Zhang, P., & Xu, J. (2023). Effect of preprocessing and optimization strategies on diabetes classification: A machine learning pipeline study. BMC Bioinformatics, 24, 1–14. https://doi.org/10.1186/s12859-023-05144-2
11. Dharmarathne, G., Perera, I., & Silva, Y. (2024). Explainable machine learning framework for diabetes diagnosis using self-explanatory interfaces. Expert Systems with Applications, 236, 121309. https://doi.org/10.1016/j.eswa.2023.121309
12. Tašić, J., Petrović, M., & Ranković, V. (2024). A multi-level fuzzy inference system for estimating type-2 diabetes risk using behavioral and environmental indicators. Mathematics, 12(8), 1167. https://doi.org/10.3390/math12081167