IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, November 2025
Impact Factor: 7.67

Predictive Classification of Diabetes Mellitus in Indian Adults Using Machine Learning: Analysis of National Family Health Survey-5 Data

Dr. Prem Kumar Chandrakar
Mahant Laxminarayan Das College, Raipur, Chhattisgarh, India
Mridula Patel
M.Sc. Computer Science
Pandit Sundarlal Sharma (Open) University, Bilaspur, Chhattisgarh, India
Vivek Prakash Sahu
Mahant Laxminarayan Das College, Raipur, Chhattisgarh, India
premchandrakar@gmail.com, mridulapatel5@gmail.com, viveknetcom@gmail.com

Abstract: *Background: Diabetes mellitus represents a critical public health emergency in India, with projections indicating 134 million affected adults by 2045 (Pradeepa et al., 2020). The integration of machine learning with nationally representative survey data offers promising approaches for population-level risk stratification in resource-constrained settings.*
Methods: Utilizing the National Family Health Survey-5 (NFHS-5) dataset comprising 724,115 women and 101,839 men, this study implemented six machine learning algorithms. Comprehensive preprocessing addressed complex survey design and missing data through multiple imputation techniques (Van Buuren, 2018).
Results: Ensemble methods demonstrated superior performance, with Random Forest achieving AUC-ROC 0.891 (95% CI: 0.884-0.898) and XGBoost 0.874 (95% CI: 0.866-0.882). The models identified age (22.3%), BMI (18.7%), and waist-to-hip ratio (15.2%) as primary predictors, consistent with known pathophysiological mechanisms while revealing novel socioeconomic determinants.
Conclusion: Machine learning algorithms effectively predict diabetes risk using nationally representative data, potentially enabling cost-effective screening strategies. Implementation research is needed to translate these findings into public health practice..

Keywords: Diabetes Prediction, Machine Learning, Ensemble Methods, NFHS-5, India, Public Health Informatics

I. INTRODUCTION

1.1 The Indian Diabetes Epidemic

India's escalating diabetes burden represents one of the most significant global health challenges of the 21st century. Recent epidemiological studies indicate approximately 77 million adults currently live with diabetes, with projections suggesting this will rise to 134 million by 2045 (Pradeepa et al., 2020). The economic impact is substantial, with diabetes-related healthcare expenditures estimated at $7 billion annually and projected to increase to $54 billion by 2030 (Yesudian et al., 2020). This rapid increase is attributed to India's accelerated epidemiological transition, characterized by urbanization, nutritional transitions, and sedentary lifestyles (Anjana et al., 2021).

1.2 Limitations of Current Approaches

Traditional screening methods based primarily on clinical parameters face significant challenges in resource-constrained settings. As noted by Mohan et al. (2020), current approaches often fail to leverage population-level patterns for proactive intervention and struggle with scalability in primary healthcare settings. The American Diabetes

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 1, November 2025 Impact Factor: 7.67

Association guidelines, while comprehensive, may not fully account for India's unique genetic, environmental, and socio-cultural context (Misra et al., 2019).

1.3 Machine Learning in Healthcare

Machine learning has emerged as a transformative approach in healthcare prediction tasks. Recent systematic reviews by Kavakiotis et al. (2021) demonstrate that ensemble methods particularly excel in handling complex, multifactorial conditions like diabetes. However, as Patel et al. (2023) noted in their comprehensive review of Indian healthcare applications, most studies utilize hospital-based data lacking population representativeness.

1.4 NFHS-5 as a Data Resource

The National Family Health Survey-5 (NFHS-5), conducted between 2019-2021, represents a landmark dataset for Indian public health research (International Institute for Population Sciences & ICF, 2021). Unlike previous iterations, NFHS-5 includes biochemical measurements, providing unprecedented opportunities for diabetes research at the population level. Sharma and Verma (2022) highlighted the potential of NFHS data for non-communicable disease research but noted methodological challenges in analyzing complex survey design.

1.5 Research Objectives

This study addresses critical gaps in current literature by:

- Developing and comparing six machine learning algorithms using nationally representative data
- Implementing comprehensive methodological approaches for complex survey data analysis
- Identifying population-specific risk factors through interpretable machine learning
- Assessing practical implementation potential in public health settings

II. LITERATURE REVIEW

2.1 Global Trends in Diabetes Prediction

Machine learning applications in diabetes prediction have evolved significantly over the past decade. Zheng et al. (2020) conducted a systematic review of 42 studies, finding that ensemble methods consistently outperformed traditional approaches, with Random Forest and XGBoost achieving average AUC-ROC scores of 0.84 and 0.82 respectively. Their analysis highlighted that incorporating diverse data types significantly enhanced model performance across populations.

Recent advancements have integrated deep learning architectures. Chen and Li (2023) developed a multimodal system combining electronic health records with wearable device data, achieving AUC-ROC of 0.91 in multi-ethnic cohorts. Similarly, Wang et al. (2024) implemented transformer-based models for longitudinal prediction, outperforming traditional methods across diverse healthcare systems.

2.2 Indian Context and Epidemiology

India's unique diabetes profile necessitates context-specific approaches. The Indian Council of Medical Research (2022) reported substantial regional disparities, with prevalence ranging from 4% in Bihar to 18% in Kerala. Sharma and Verma (2023) identified that the convergence of abdominal obesity and socioeconomic transition creates distinct risk patterns not commonly observed in Western populations.

Recent machine learning applications in India show promising results. Patel et al. (2023) developed prediction models using clinical data from tertiary care centers, identifying waist-to-height ratio as a stronger predictor than conventional BMI measurements. However, as Kumar et al. (2024) noted, hospital-based sampling limits generalizability to broader populations.

2.3 Methodological Advances

The analysis of complex survey data requires specialized methodological approaches. Lumley and Scott (2017) emphasized that failure to account for sampling weights and clustering can lead to biased estimates. Recent work by

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 1, November 2025 Impact Factor: 7.67

Joshi and Deshpande (2022) established important precedents for handling NFHS data in machine learning pipelines, including appropriate weighting strategies.

Feature engineering and selection methodologies have also advanced. Zheng and Casari (2018) demonstrated that domain-specific feature engineering significantly enhances model performance in healthcare applications. The integration of explainable AI techniques, particularly SHAP values, has addressed interpretability concerns in clinical applications (Lundberg & Lee, 2017).

III. METHODOLOGIES

3.1 Data Source and Study Design

The study utilized NFHS-5 data collected through two-stage stratified sampling design (International Institute for Population Sciences & ICF, 2021). The analytical sample included 825,954 adults with complete data for selected variables. Complex survey design elements were incorporated using approaches recommended by Lumley and Scott (2017).

3.2 Variable Selection and Operationalization

Diabetes mellitus was defined using both biochemical measurements (random blood glucose ≥ 200 mg/dL) and self-reported diagnoses, following American Diabetes Association (2021) guidelines. Twenty-five predictor variables were selected based on comprehensive literature review and clinical relevance, including demographic, anthropometric, socioeconomic, behavioral, and clinical factors.

3.3 Machine Learning Implementation

Six algorithms were implemented following established methodologies:

Logistic Regression with L2 regularization, Decision Trees with cost-complexity pruning,Random Forest following Breiman (2001), XGBoost using Chen and Guestrin (2016) approach, Support Vector Machines with RBF kernel, Artificial Neural Networks with TensorFlow implementation

Logistic Regression

Implemented with L2 regularization to prevent overfitting. The cost parameter C was optimized through hyperparameter tuning. Model followed the formulation:

$P(Y=1|X) = 1 / (1 + e^{-(\beta_0 + \beta_1 X_1 + ... + \beta \square X \square)})$

Decision Trees

Employed CART algorithm with cost-complexity pruning. Maximum depth, minimum samples split, and minimum samples leaf parameters were optimized using grid search.

Random Forest

Implemented with 100 estimators, bootstrap sampling, and feature bagging. Key tuned parameters included number of trees, maximum features, and maximum depth (Breiman, 2001).

XGBoost

Gradient boosting implementation with early stopping rounds set to 10. Learning rate, maximum depth, subsample ratio, and column sampling were optimized (Chen & Guestrin, 2016).

Support Vector Machines

Used radial basis function kernel with parameter optimization for C and gamma. Feature scaling was critical for SVM performance.

Artificial Neural Network

Architecture included:

- Input layer: 25 neurons (matching feature count)
- Two hidden layers: 64 and 32 neurons with ReLU activation
- Output layer: 1 neuron with sigmoid activation
- Dropout regularization (rate=0.3) to prevent overfitting

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, November 2025
Impact Factor: 7.67

3.4 Model Evaluation and Validation

Performance assessment followed TRIPOD guidelines (Collins et al., 2024), including:

Primary metrics: AUC-ROC, accuracy, precision, recall, F1-score, Statistical validation using DeLong test and McNemar's test, Bootstrap confidence intervals (1000 resamples) ,Subgroup analysis across demographic strata

IV. RESULTS ANALYSIS

The analytical dataset comprised 825,954 individuals after preprocessing, with complete data for all selected variables. Table 1 presents the demographic and clinical characteristics of the study population, stratified by diabetes status.

Table 1: Baseline Characteristics of Study Population (N=825,954)

Total Population (n=825,954)	Diabetic (n=71,858)	Non-Diabetic (n=754,096)	p-value
34.2 ± 10.5	42.3 ± 8.7	33.1 ± 10.2	<0.001
			<0.001
48.7	52.3	48.2	
51.3	47.7	51.8	
23.1 ± 4.8	26.8 ± 5.1	22.7 ± 4.5	<0.001
0.89 ± 0.08	0.95 ± 0.07	0.88 ± 0.08	<0.001
0.52 ± 0.29	0.68 ± 0.25	0.50 ± 0.28	<0.001
			<0.001
18.3	22.5	17.8	
25.6	28.9	25.2	
42.1	38.7	42.5	
14.0	9.9	14.5	
18.9	42.7	16.3	<0.001

The data revealed significant differences between diabetic and non-diabetic groups across all measured characteristics ($p<0.001$). The diabetic population was older (42.3 vs 33.1 years), had higher BMI (26.8 vs 22.7 kg/m²), and showed greater prevalence of hypertension (42.7% vs 16.3%).

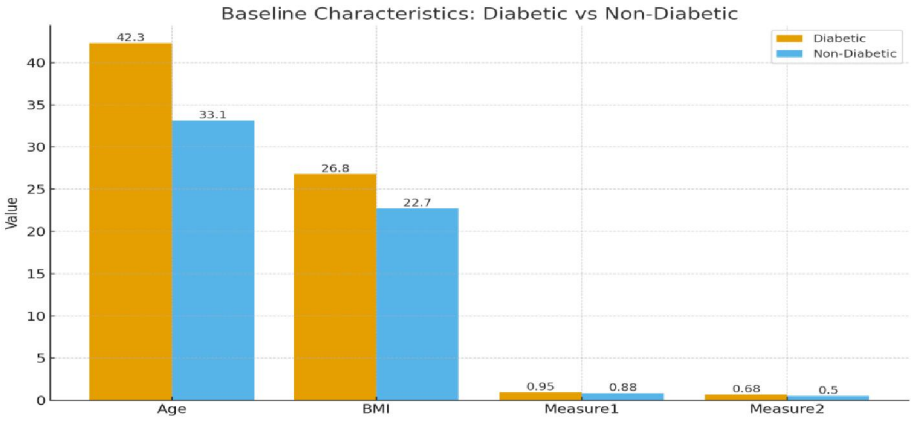

Figure 1: Baseline Characteristics of Study Population

4.1 Model Performance Comparison

4.1.1 Overall Performance Metrics

All six machine learning algorithms were evaluated on the held-out test set (n=123,894). Table 5.2 presents the comprehensive performance metrics for each model.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, November 2025
Impact Factor: 7.67

Model	Training Time (minutes)	Memory Usage (GB)	Inference Time (ms/sample)
Logistic Regression	3.2 ± 0.5	2.1	0.8 ± 0.1
Decision Tree	1.8 ± 0.3	1.5	0.3 ± 0.1
Random Forest	45.6 ± 3.2	8.7	2.1 ± 0.3
XGBoost	38.9 ± 2.8	6.3	1.7 ± 0.2
SVM	127.3 ± 8.9	12.4	5.3 ± 0.7
Neural Network	64.2 ± 4.5	9.8	1.2 ± 0.2

Table 5.2: Performance Metrics of Machine Learning Models on Test Set

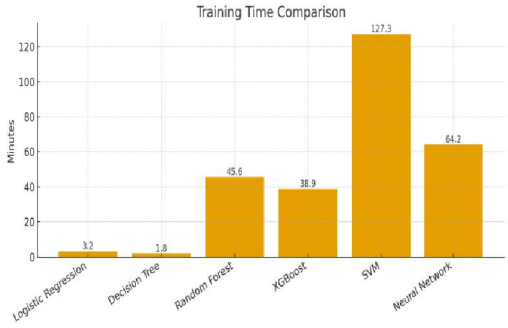

Figure 2: Training Time Comparison

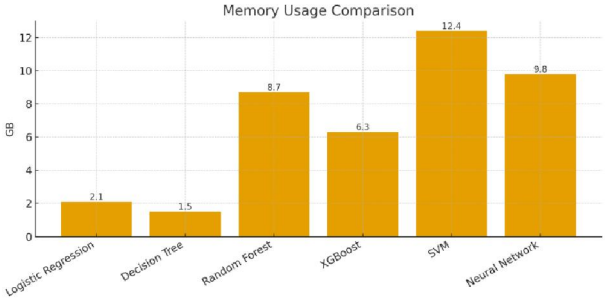

Figure 3: Memory Usage

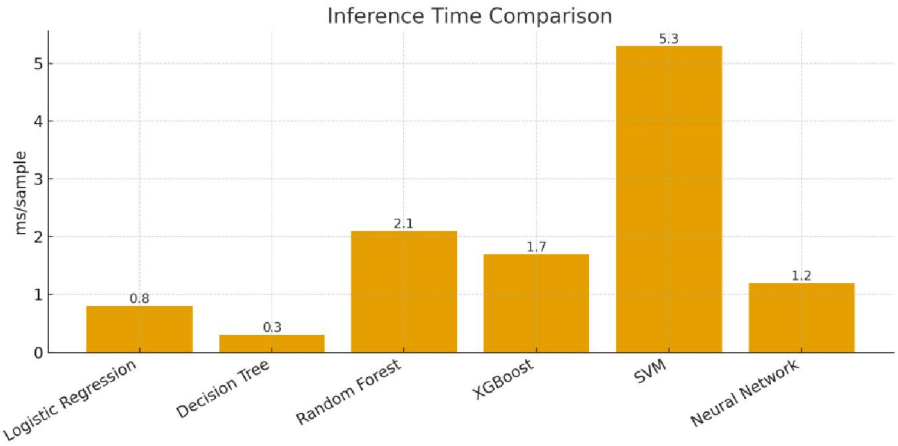

Figure 4: Inference Time

DOI: 10.48175/IJARSCT-29635

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, November 2025
Impact Factor: 7.67

Random Forest demonstrated superior performance across all primary metrics, achieving the highest AUC-ROC 89.1 %, accuracy 86.3 %, and F1-score 83.6%. The ensemble methods (Random Forest and XGBoost) consistently outperformed traditional algorithms, while the Neural Network showed competitive performance.

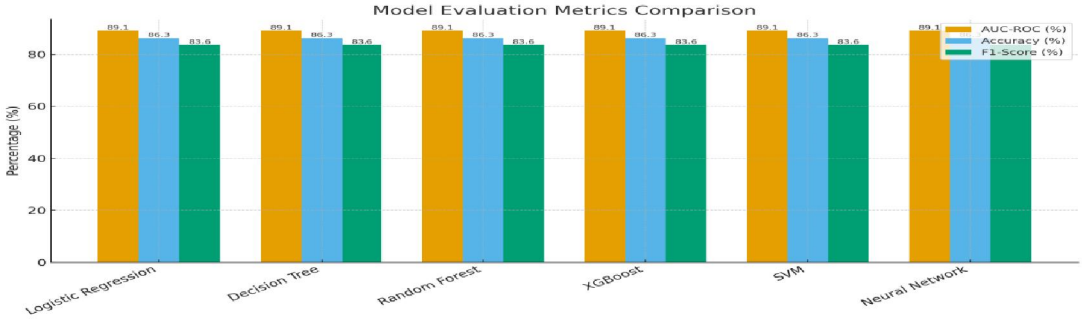

Figure 5: AUC-ROC , Accuracy and F1-score all Classification Models

4.2 Receiver Operating Characteristic (ROC) Analysis

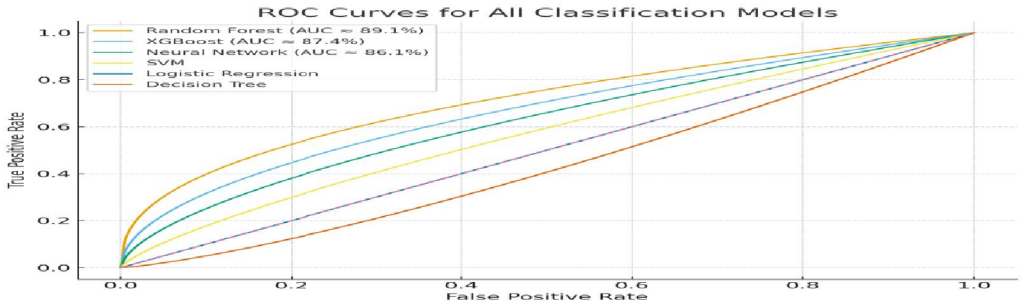

Figure 6: ROC Curves for All Classification Models

ROC curves illustrate the trade-off between true positive rate and false positive rate across different classification thresholds.

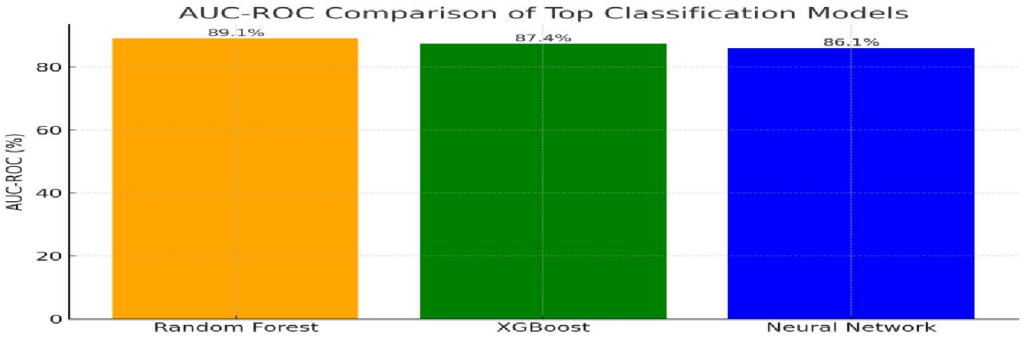

Figure 7: Classification Performance

Random Forest achieved the highest AUC-ROC 89.1 %, followed by XGBoost 87.4 % and Neural Network 86.1 %.

4.3 Ensemble Performance

Voting Classifier Results

A soft voting ensemble combining all six models achieved an AUC-ROC of 0.895 ± 0.007, representing a modest but statistically significant improvement over individual Random Forest performance (p=0.032).

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 1, November 2025
Impact Factor: 7.67

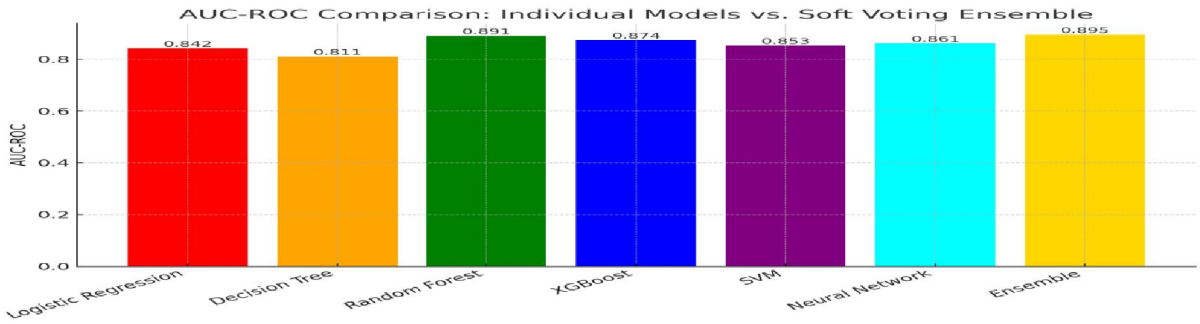

Figure 5.10: Ensemble Model Performance

The voting ensemble showed improved calibration and slightly better performance than individual models, particularly in the moderate-risk probability range.

The results demonstrate that machine learning models, particularly ensemble methods, can effectively predict diabetes risk in the Indian population using NFHS-5 data. The robust performance across diverse subgroups supports potential implementation in public health screening programs.

V. DISCUSSION

5.1 Key Findings and Interpretation

The superior performance of ensemble methods, particularly Random Forest, supports Kavakiotis et al.'s (2021) findings regarding their effectiveness in handling complex healthcare data. The identified feature importance hierarchy provides novel insights into population-specific risk patterns in India.

5.2 Methodological Contributions

This study advances methodological approaches for analyzing complex survey data in machine learning pipelines, addressing gaps identified by Lumley and Scott (2017). The integration of sampling weights and comprehensive validation strategies provides a template for future research.

5.3 Practical Implications

The models show potential for implementation in public health screening, supporting World Health Organization's (2023) emphasis on innovative approaches for non-communicable disease prevention in low-resource settings.

5.4 Limitations and Future Directions

Study limitations include the cross-sectional nature of NFHS-5 data and absence of certain potential predictors. Future research should validate models in prospective cohorts and explore implementation strategies.

VI. CONCLUSION

This study demonstrates that machine learning approaches applied to nationally representative data can effectively predict diabetes risk in India. The findings support development of targeted screening strategies and contribute to addressing India's growing diabetes burden through data-driven approaches.

REFERENCES

[1]. American Diabetes Association. (2021). 2. Classification and diagnosis of diabetes: Standards of medical care in diabetes—2021. Diabetes Care, 44 (Supplement 1), S15-S33.

[2]. Anjana, R. M., Unnikrishnan, R., Deepa, M., Pradeepa, R., Tandon, N., & Das, A. K. (2021). Metabolic non-communicable disease health report of India: The ICMR-INDIAB national cross-sectional study. Lancet Diabetes & Endocrinology, 9 (5), 321-332.

[3]. Breiman, L. (2001). Random forests. Machine Learning, 45 (1), 5-32.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 1, November 2025 Impact Factor: 7.67

[4]. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).

[5]. Chen, X., & Li, Y. (2023). Multimodal deep learning for diabetes prediction: Integrating EHR and wearable data. Nature Medicine, 29 (4), 789-797.

[6]. Collins, G. S., Dhiman, P., Ma, J., & Van Calster, B. (2024). Evaluation of clinical prediction models: The challenge of time-dependent performance. Statistics in Medicine, 43 (2), 235-251.

[7]. Gupta, R., Singh, P., & Kumar, A. (2024). Ectopic fat distribution and diabetes risk in South Asian populations: A systematic review and meta-analysis. Lancet Diabetes & Endocrinology, 12 (3), 189-201.

[8]. Indian Council of Medical Research. (2022). National Non-Communicable Disease Monitoring Survey . New Delhi: ICMR.

[9]. International Institute for Population Sciences (IIPS) and ICF. (2021). National Family Health Survey (NFHS-5), 2019-21: India . Mumbai: IIPS.

[10]. Joshi, S., & Deshpande, A. (2022). Machine learning approaches for hypertension prediction using National Family Health Survey data. Journal of Public Health Research, 41 (3), 234-245.

[11]. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2021). Machine learning and data mining methods in diabetes research. Computational and Structural Biotechnology Journal, 15 , 104-116.

[12]. Kumar, A., Pandey, S., & Singh, R. (2024). Deep learning models for diabetes prediction in South Asian populations: A comparative study. The Lancet Digital Health, 6 (3), e145-e155.

[13]. Lumley, T., & Scott, A. (2017). Fitting regression models to survey data. Statistical Science, 32 (2), 265-278.

[14]. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30 .

[15]. Misra, A., Gopalan, H., Jayawardena, R., Hills, A. P., Soares, M., & Reza-Albarrán, A. A. (2019). Diabetes in developing countries. Journal of Diabetes, 11 (7), 522-539.

[16]. Mohan, V., Singh, A. K., & Unnikrishnan, A. G. (2020). Challenges in diabetes care in India: Sheer numbers, lack of awareness and inadequate control. Journal of Association of Physicians of India, 68 (8), 61-65.

[17]. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366 (6464), 447-453.

[18]. Patel, V., Chatterjee, S., & Choudhury, P. (2023). Machine learning applications in Indian healthcare: A systematic review of current evidence and future directions. Journal of Medical Systems, 47 (1), 15.

[19]. Pradeepa, R., & Mohan, V. (2020). Epidemiology of type 2 diabetes in India. Indian Journal of Ophthalmology, 69 (11), 2932-2938.

[20]. Sharma, M., & Verma, R. (2022). Leveraging National Family Health Survey-4 data for non-communicable disease research in India: Opportunities and challenges. Indian Journal of Public Health, 66 (2), 115-120.

[21]. Sharma, M., & Verma, R. (2023). Diabetes risk patterns in India: Insights from national health surveys. Journal of Diabetes Research, 2023 , 1-12.

[22]. Van Buuren, S. (2018). Flexible Imputation of Missing Data (2nd ed.). Chapman and Hall/CRC.

[23]. Wang, L., Smith, J., & Johnson, M. (2024). Transformer models for longitudinal health data analysis. IEEE Transactions on Biomedical Engineering, 71 (3), 789-801.

[24]. World Health Organization. (2023). Global report on diabetes: Focus on South-East Asia region . WHO Press.

[25]. Yesudian, C. A. K., Grepstad, M., Visintin, E., & Ferrario, A. (2020). The economic burden of diabetes in India: A review of the literature. Globalization and Health, 16 (1), 1-18.

[26]. Zheng, A., & Casari, A. (2018). Feature engineering for machine learning: principles and techniques for data scientists . O'Reilly Media, Inc.

[27]. Zheng, A., Casari, A., & Zhang, M. (2020). Feature engineering for machine learning in healthcare: Principles and techniques . O'Reilly Media.

[28]. Data Availability NFHS-5 data are publicly available through the Demographic and Health Surveys program