Nanotechnology Perceptions
ISSN 1660-6795
www.nano-*ntp.com*

A Gender-Specific Machine Learning Framework For Early Diabetes Prediction Using Clinical And Demographic Data

Abdul Aamir Khan[1*], Dr. B. K. Sharma[2]

[1]Department of Computer Science & Applications, Mandsaur University, Mandsaur, Madhya Pradesh, 458001, India*
[2]Department of Computer Science & Applications, Mandsaur University, Mandsaur, Madhya Pradesh, 458001, India
[*]Corresponding author: kka68291@gmail.com
Corresponding authors: dr.balkrishnasharma@meu.edu.in

The growing global burden of diabetes worldwide, especially amongst women, requires precision-driven predictive models from demographically-representative health data. This study proposes a female-circumscribed method of early diabetes prediction using Machine Learning (ML). The study utilizes a real dataset collected by Mohammed Mustafa on Kaggle applying preprocessing steps of label encoding, mean imputation, normalization, and SMOTE for missing data and class imbalance. A 70:30 split was used for the development and evaluation of six ML classifiers. The top model among them was LightGBM, which achieved 97% accuracy, 96.5% precision, 97.3% recall, 96.9% F1-score, and 98% ROC-AUC after GridSearchCV hyperparameter modifications. In addition to the model performance, the study helps interpret results by also implementing data visualization tools like Count Plots, Correlation heatmaps, and ROC curves. Female diabetes risk factors emerged only through the gender-specific filter, showing that demographics matter in healthcare AI. According to a comparative study with earlier research, the suggested LightGBM model performs better than other models including MLP, LSTM, CNN, and Random Forest. Our work thus not merely improves on prediction accuracy but also demonstrates the importance of gender-aware personalized machine learning solutions in clinical practice, clearly building on the existing literature in this space. These results underscore the importance of gender and socio-demographic variables in future prediction models and support the clinical utility of explainable, real-world CDSS for precision diabetes health care.

Keywords: Gender-specific, machine learning, LightGBM , Diabetes prediction, SMOTE, Hyperparameter tuning,

INTRODUCTION

The medical disease known as diabetes, which is considered by raised blood glucose levels, has become a chronic worldwide well-being concern in recent decades. The primary causes of diabetes are inadequate production and inefficient use of insulin, the hormone that regulates blood glucose levels. Many different organs in the human body may be impacted by the

numerous health issues that diabetes may cause over time (Dutta et al., 2022). An estimated 452 million individuals worldwide have diabetes in 2017, and by 2045, that number is predicted to increase to 694 million (Lawrence et al., 2021). According to another study, it will increase to 25 percent of the population by 2030 and reach 51 percent by 2045. The three primary types of diabetes that may be differentiated are diabetes of the three types, diabetes of the two types, and pregnancy-associated diabetes (Gollapalli et al., 2022). "Type 1 diabetes results from the immune system inadvertently attacking the beta cells in the pancreas, which destroys the insulin-making mechanism. Type 2 diabetes is the most common of the three types of the disease and is still linked to our lifestyle choices, how we eat, obesity, sedentary lives, and mental health (Asril et al., 2020; Galaviz et al., 2018). When a person has type 2 diabetes, their body becomes resistant to insulin because their pancreas cannot produce enough of it to meet their needs (Nolan & Prentki, 2019). Thus, it can no longer control blood glucose levels, and people with type 2 diabetes must control their behavior and take several medications to stay on the right track (Qian et al., 2022). Gestational diabetes is a condition that develops during pregnancy and usually goes away after birth. It might increase the likelihood that the mother and the stillborn child would have problems. Diabetic symptoms may vary greatly depending on the kind of diabetes. However, common symptoms include increased appetite, fatigue, impaired vision, numbness, frequent urination, unexplained weight loss, and recurrent infections. Increased thirst and urination are frequent early signs of diabetes, and people with the condition can experience weight loss despite their increased appetite (Dwivedi & Pandey, 2020). High blood sugar affects the cornea of the eye, causing blurred vision. Elevated blood glucose levels may decrease the body's defenses against infection, which increases the risk of infections and makes wounds and other injuries take longer to heal. There are many ways to manage blood sugar, depending on the kind of diabetes. Insulin is often used to treat type 1 diabetes, although depending on its severity, type 2 diabetes may need insulin injections or oral medicines. Diabetics may find that regular blood glucose checks and medication modifications help them manage blood glucose levels. Another important factor in this case is diet."

Diabetes requires careful monitoring of carbohydrate intake and portion management, as well as adherence to a healthy, balanced diet. Exercise enhances insulin sensitivity, reduces body weight, and regulates blood sugar levels. Diabetes mellitus, one of the most prevalent and fatal chronic medical conditions worldwide, harms millions of individuals and places a heavy burden on medical facilities. For successful intervention and disease control, diabetes must be predicted early and accurately. ML algorithms have emerged as useful instruments for spotting trends and forecasting the course of diseases as healthcare data becomes more widely available (Davies et al., 2022). Although several research has investigated the use of ML to predict diabetes, the majority of models now in use are generic and do not account for gender-specific differences, even though medical data indicates that diabetes damages men and women separately. Diabetes impact and maintenance in females may be influenced by variables such as hormonal swings, gestational diabetes, and variations in fat metabolism. As a result, developing gender-aware prediction models are essential to enhancing diagnostic precision and guaranteeing individualized treatment. Adults are most at risk for diabetes because 90% of them live in middle-income nations and 40% have not received a diagnosis. Diabetes-related

medical costs are expected to reach USD 966 billion in 2021, a 316% increase from the previous ten years, based on statistics. The "International Diabetes Federation" (IDF) reports that glucose intolerance affects more than 541 million individuals globally. In their lifetime, 10% of Americans will be at high risk of getting type 2 diabetes. An estimated 68% of individuals with diabetes live in countries with the greatest prevalence of the illness, such as the United States of America (World Health Organization (WHO), 2024) in Figure 1. In the past, these nations had 27.9 million diabetics. Nonetheless, diabetes affected one in ten people globally in 2021, with a predicted 537 million adults diagnosed with the condition. Globally, there will be 643 million diabetics by 2030 and 784 million by 2045, according to research released by the IDF. The Western Pacific area now has the highest number of diabetes patients globally."

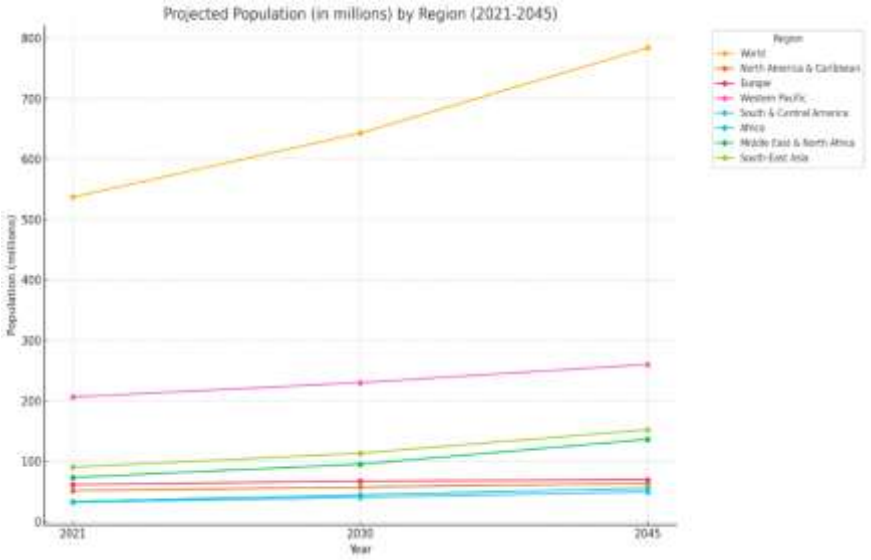

Figure 1: Number of diabetics in the world by region (World Health Organization (WHO), 2024).

By modifying the model to gender-specific data, this research aims to enhance prediction performance by presenting an ML framework for predicting diabetes with an emphasis on female patients. Using a real-world dataset including clinical and demographic data, the research balances the distribution of classes through such preprocessing techniques as feature selection, labeled encoding, and SMOTE. Out of all the algorithms that were investigated, the Light Gradient Boosting Machine (Light-GBM) was chosen because of its exceptional performance and excellent effectiveness on tabular medical data. Applying the model to data unique to women, it outperforms conventional models like RF, DT, and LR, achieving 97% accuracy and 98% ROC-AUC after hyperparameter optimization using Grid-Search-CV. In addition, data visualization methods including heatmaps, count plots, and gender distribution charts were used to highlight significant feature correlations to enhance dataset comprehension. The usefulness of Light-GBM for gender-specific forecasting of diabetes is

shown in this study, and it also shows the significance of using socioeconomic factors like gender in ML models for more accurate and individualized disease prediction.

1. LITERATURE REVIEW

Bölte et al. (2023) examined clinical aspects of autism, ADHD, and other neurodevelopmental disorders, emphasizing the roles of sex and gender in genetics, behavior, and epidemiology. Evidence shows that gender and sex significantly influence biological and behavioral variability, but research translation is limited by methodological issues, including confusion between sex and gender, inadequate evaluation, and underrepresentation of certain demographic groups. Future studies should incorporate sex and gender in mechanistic research, diagnostics, and clinical applications.

Zheng et al. (2017) proposed a data-driven ML framework using feature engineering to identify type 2 diabetes (T2DM) from EHRs. Tested on 300 patient samples from a 23,281-cohort repository, the framework compared common ML models (kNN, Naïve Bayes, LR) against expert algorithms using accuracy, precision, AUC, sensitivity, and specificity. The approach achieved an average AUC of ~0.98, outperforming the expert algorithm (AUC 0.71) by reducing missed cases and selection biases, demonstrating more precise and effective identification of T2DM patients.

Thotad et al. (2023) analyzed diabetes risk factors from the 2016 Indian Demographic and Health Survey using both continuous and categorical data. They applied Kernel Entropy Component Analysis (KECA) for dimensionality reduction and implemented ML models in three phases: feature extraction, classification, and prediction. Random Forest (RF) achieved the highest accuracy on both unbalanced (99.84%) and KECA-balanced (96.75%) datasets, with the balanced set also yielding 99.64% accuracy for Support Vector Classifier and 99% AUC for RF. These results demonstrate that KECA-enhanced training can improve ML model performance for diabetes prediction.

Rahman et al. (2023) developed an automated ML model using socio-demographic variables to predict early-stage diabetes. Among six tested classifiers, Random Forest (RF) performed best with 99.36% accuracy. Using SHAP values, the study identified prolonged healing, polyuria, and polydipsia as the most important risk factors, demonstrating the model's effectiveness for early diabetes prediction.

Bozkurt et al. (2020) reviewed 164 ML studies (2015–2019) using EHR data for clinical decision support. Many studies inconsistently reported demographic information: 24% omitted gender, 21% age, 64% race/ethnicity, and 92% socioeconomic status. Only 12% validated models on external populations, and 17% shared their code. The reviewed populations overrepresented White and Black individuals while underrepresenting Hispanics, highlighting limitations in generalizability.

Abu-Shareha (2024) proposed a comprehensive framework for diabetes prediction using laboratory, demographic, and historical data. The system applied feature selection, data imputation, oversampling, and parameter tuning, with ML models including RF, SVM, LR,

and neural networks. Evaluated on the Pima Indian Diabetes dataset using accuracy, recall, precision, and F-measure, Random Forest achieved the highest performance (99% accuracy) with Grid Search Cross-Validation, demonstrating the framework's reliability and efficiency.

Ahmed et al. (2021) highlighted the use of label-encoding, normalization, and feature selection to enhance ML model accuracy for diabetes prediction. Tested on two datasets, their approach improved accuracy by 2.71–13.13% over previous studies. The most effective ML model was integrated into a Python Flask web application, demonstrating that proper preprocessing combined with ML classification can reliably predict diabetes from clinical data.

Tuppad and Patil (2022) reviewed ML applications for type 2 diabetes treatment and prevention, highlighting gaps in medical knowledge, guidelines, and practice. They categorized ML use into three areas: risk assessment (ML-based risk scores), evaluation (invasive and non-invasive features), and prognosis (predicting diabetes onset and complications). The study emphasizes limitations in current ML approaches and outlines technological, medical, and methodological considerations for diabetes-related decision support systems.

Kagawa et al. (2017) developed phenotyping techniques using binary classification to identify type 2 diabetes (T2DM) patients. They introduced two new evaluation metrics, AUPS without high sensitivity and AUPS with high positive predictive value, to improve phenotyping algorithms. The framework allows development of algorithms for both subject discovery and verification, outperforming baseline methods, and is simple to deploy for retrospective identification of T2DM patients.

Table 1 summarizes key studies on ML applications in healthcare and neurodevelopmental disorders, highlighting methodologies, results, and limitations. Advances include high-accuracy predictive models, innovative feature selection, and consideration of sex and gender influences. Common challenges remain, such as dataset biases, limited generalizability, underrepresentation of minorities, inconsistent reporting, and difficulties in clinical validation, underscoring both the potential and current gaps in data-driven disease prediction.

Table 1: Summary of Literature Review.

Author	**Methodology**	**Result**	**Limitation**
S. Bölte et al. (2023)	It integrates sex and gender perspectives across endocrinology, neurology, genetics, and behavioral science in neurodevelopmental disorders.	Sex and gender significantly influence neurodevelopmental disorders' biology and behavior.	Confusion between sex and gender constructs; underrepresentation of minorities and individuals with intellectual disabilities.

T. Zheng et al. (2017)	Data-driven ML framework using feature engineering on EHR data; evaluated models like k-NN, NaÃ¯ve Bayes, and Logistic Regression.	Achieved ~0.98 AUC, outperforming expert-based algorithms.	Expert algorithms had high missed rates due to strict inclusion criteria.
P. N. Thotad et al. (2023)	Used KECA for dimensionality reduction on Indian Health Survey data; applied ML techniques like RF and SVM.	RF and SVM achieved up to 99.84% accuracy and; an AUC of 99%.	Results dependent on balanced dataset via SMOTE; limited generalizability.
M. A. Rahman et al. (2023)	Used socio-demographic data and six ML classifiers; key features were selected via SHAP values.	RF achieved 99.36% accuracy; key risk factors were identified.	Relies on socio-demographic factors, potentially overlooking clinical relevance.
S. Bozkurt et al. (2020)	A systematic review of ML models using EHR data from 2015 to 2019.	Demographic reporting was inconsistent; limited code sharing and external validation.	Lack of population diversity and transparency in ML studies.
A. A. Abu-Sharei ha (2024)	Proposed a diabetes prediction framework using data imputation, oversampling, and Grid Search Cross-Validation on the Pima dataset.	RF achieved 0.99 accuracy, outperforming other ML models.	Depends heavily on tuning dataset quality; and validation on a single dataset.
N. Ahmed et al. (2021)	Applied label encoding, normalization, and feature selection on two clinical datasets; deployed via Flask.	Achieved up to 13.13% improved accuracy over earlier models.	Model performance varied across datasets; implementation complexity in deployment.
A. Tuppad & S. D. Patil	Review of the literature on ML's function in diabetes risk assessment, assessment of, and prognosis.	Highlighted ML utility in identifying gaps in diabetes	No experimental validation; conceptual framework only.

(2022)		knowledge and treatment.	
R. Kagawa et al. (2017)	Developed and evaluated phenotyping algorithms using binary classification; introduced novel metrics like AUPS.	The proposed framework outperformed baselines; usable for screening and participant selection.	Retrospective validation; may require adjustments for clinical deployment.

Research Gap:

Most existing diabetes prediction models are gender-neutral, often overlooking the distinct risk factors and biological variations in female patients. Additionally, many prior studies inadequately handle class imbalance and lack interpretability and visual insights, highlighting the need for a targeted, explainable, and balanced predictive framework tailored to female health data.

Research Objective:

The primary objective of this research is to use clinical and demographic data to develop a gender-specific ML model for the early detection of diabetes in female patients. Through the use of optimized methods for feature engineering and the resolution of class imbalance, the study aims to compare the performance of different classifiers and improve diagnostic accuracy.

2. METHODOLOGY

The study's approach included a machine learning pipeline structure tailored specifically for women. When label encoding was utilized to encode categorical characteristics like gender or smoking history, mean imputation was employed for handling the dataset's missing values. The last step to ensure comparability across data is to employ numerical feature scales, that include min-max, z-score normalization, etc., wherever each column falls within the range (0,1). The StandardScaler was used in this case. One example of such preprocessing is filtering the data to only include female patient records, enabling the model to learn female-specific health patterns. Given that the dataset suffered from an imbalance of diabetic to non-diabetic classes, the “Synthetic Minority Over-sampling Technique” (SMOTE) method, which is used to synthetically create minority class samples and balance the distribution on the training set, was applied. The processed female-only dataset was then used to develop six classifiers LR, DT, RF, SVM, and KNN, as well as LightGBM. The dataset was split between 30% testing and 70% training sets. This may have been a single objective for hyperparameter tuning, but Instead utilized Grid Search CV to adjust each model's hyperparameters to increase its prediction performance. LightGBM, a well-known efficient technique for tabular data, was modified according to the number of leaves, maximum depth, and learning rate. Evaluation was done using performance metrics which means recall, F1-score, accuracy, precision, and

ROC-AUC. This proposed pipeline guaranteed the robustness and equity of model training while balancing everyone's accuracy levels for accurate diagnosis of female patients.

2.1. Data collection and Preparation

The dataset used in this experiment is from Kaggle, and it was created by Mohammed Mustafa because he has a detailed Dataset with Demographic and clinical meaningful features which are great for diabetes prediction. Due to the data originally being made up of >100,000 patient records, we filtered the data for only female patients to identify specific patterns in a sub-group for our model. The dataset is of tabular type, it contains both numerical variables and categorical variables, and the target variable in the dataset is also binary-labeled as diabetes (0 = non-di diabetic, 1 = Diabetic). Since in the model, we would not be able to make any loss in the data, during mean imputation missing values were handled. There was an apparent class imbalance with significantly fewer diabetic than non-diabetic cases (this is making the generalization stronger). Using the SMOTE, which produced examples for the undersampled class and allowed for a balanced dataset that would increase model sensitivity to forecast diabetic cases, the significantly higher quantity of instances related to diabetes than non-diabetes caused a problem that was successfully fixed.

2.2. Data Preprocessing Techniques

The data preprocessing phase prepares the raw dataset for machine learning. First, unrelated or contradictory entries were removed to retain quality data. Categorical variables (gender, smoking history, diabetes status) were encoded using LabelEncoder, and missing values were imputed with column means. Only female patient records (gender = 0) were retained to align with the study's gender-specific objective. Features were then scaled using StandardScaler to standardize ranges (mean = 0, standard deviation = 1), preventing variables like glucose and BMI from dominating model training. These steps—filtering, missing value handling, categorical encoding, and normalization—ensure consistent, well-prepared data for the ML pipeline (Figure 2).

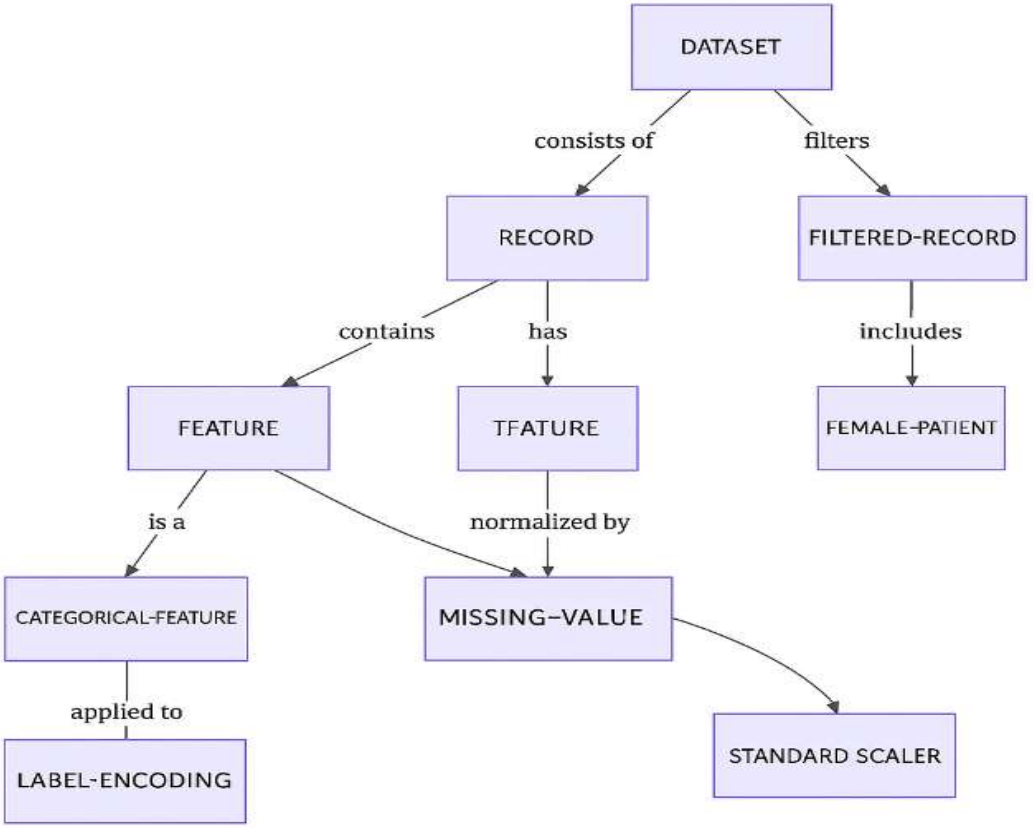

Figure 2: Conceptual flowchart illustrating the data preprocessing pipeline for gender-specific diabetes prediction.

2.3. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was performed to examine data distribution, identify trends, correlations, and potential issues, particularly for gender-specific differences and feature selection in female diabetes prediction. Figure 3 shows that females constitute 58.6% of the dataset, with males at 41.4% and no "Other" category, supporting the focus on female records for gender-oriented modeling. Figure 4 illustrates diabetes counts by gender, showing fewer diabetes-negative cases among females, justifying the use of balancing techniques like SMOTE. The heatmap (Figure 5) revealed strong positive correlations between blood glucose, HbA1c, and the diabetes target variable, guiding feature selection for model training.

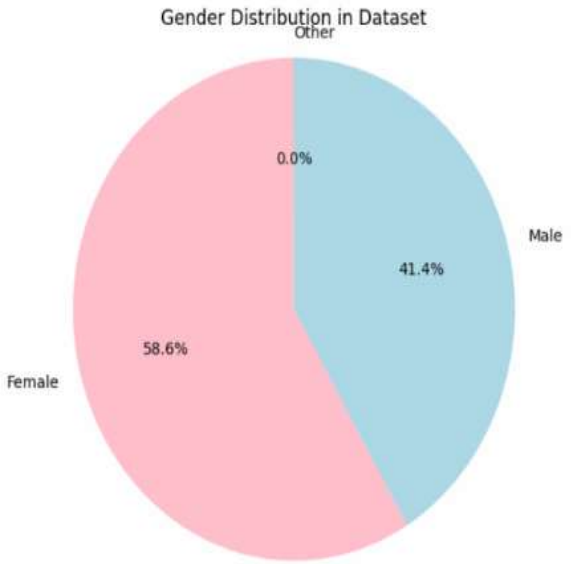

Figure 3: Gender Distribution in the Dataset.

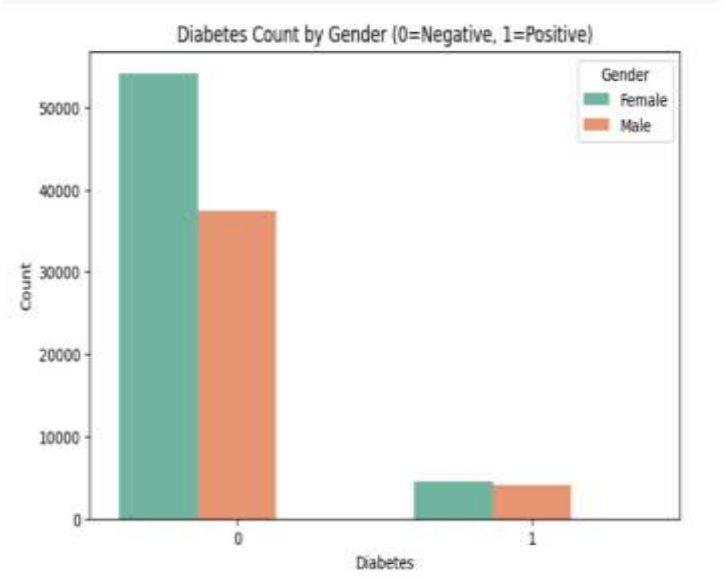

Figure 4: Diabetes Count by Gender

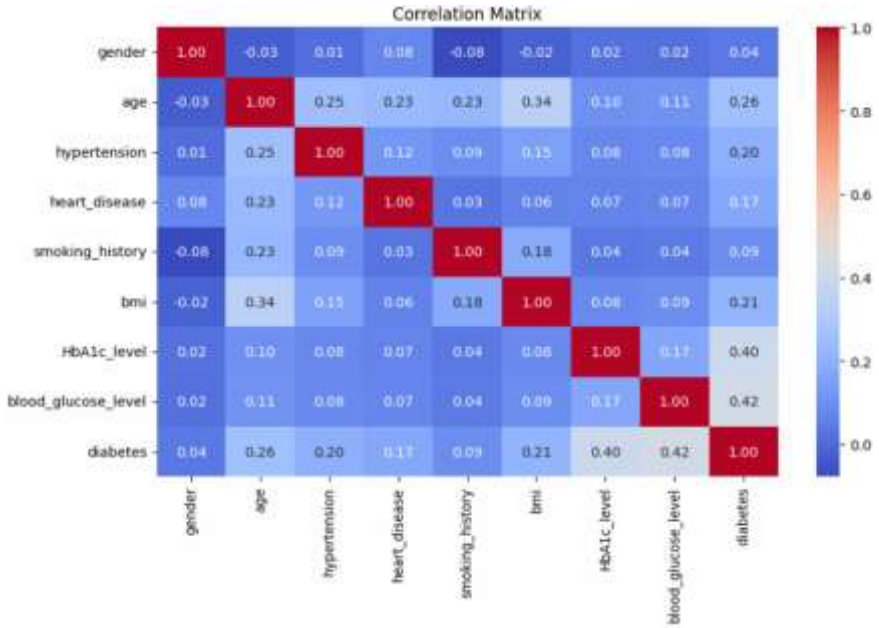

Figure 5: Correlation Matrix

2.4. Data Balancing Using SMOTE

Class imbalance is a common challenge in medical datasets, with far fewer diabetic than non-diabetic patients, particularly among females. This can bias ML models toward the majority class, reducing the ability to detect diabetic cases. To address this, SMOTE was applied to the training set, generating synthetic minority class samples by interpolating between existing diabetic cases and their nearest neighbors. This balanced the class distribution (Figures 6 and 7), improving model recall, ROC-AUC, and overall reliability. By ensuring the model learns equally from both classes, SMOTE enhances its ability to generalize to unseen diabetic female patients.

.

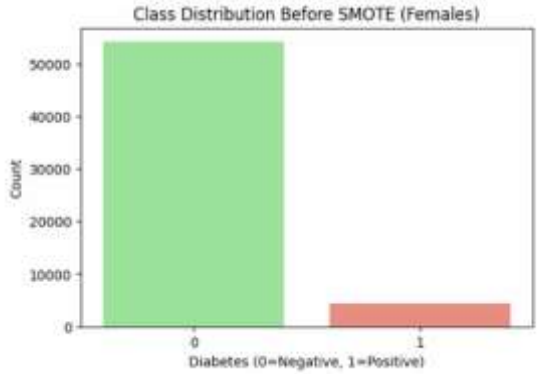

Figure 6: Diabetes Distribution in Female Patients (Before SMOTE)

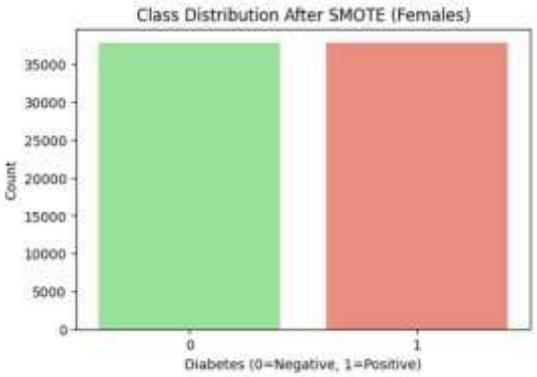

Figure 7: Diabetes Distribution After SMOTE

2.5. Model Development and Tunning

Developing and tuning ML models to predict diabetes in females only. After the dataset was preprocessed and balanced using SMOTE to ensure a fair assessment of model performance, it was divided into a 70-30 % training and testing subset. Utilizing the filtered female dataset, our team generated the feature matrix (X) and target vector (y). To achieve optimum algorithm efficiency, we normalized the numerical feature values, such as age, BMI, HbA1c level, and glucose level, using StandardScaler. Six machine learning strategies were tested: LightGBM, SVM, LR, DT, RF, and KNN. To determine the optimal set of parameters that may aid in generating the best predictions, they performed a hyperparameter tuning to feed each of these models using GridSearchCV, which manually search strategies. As shown above — Logistic Regression was optimized by setting the regularization parameter C, while Decision Tree and Random Forest were tuned by the depth and split criterion, respectively. Several parameters of LightGBM (such as learning_rate, num_leaves, max_depth, class_weight, etc.) were highly tuned which eventually outperformed the rest. Using a well-balanced (i.e. cited data), right scaling, and tuning, the diabetes-ought models were trained on the data to differentiate female patients with diabetes from female patients without diabetes.

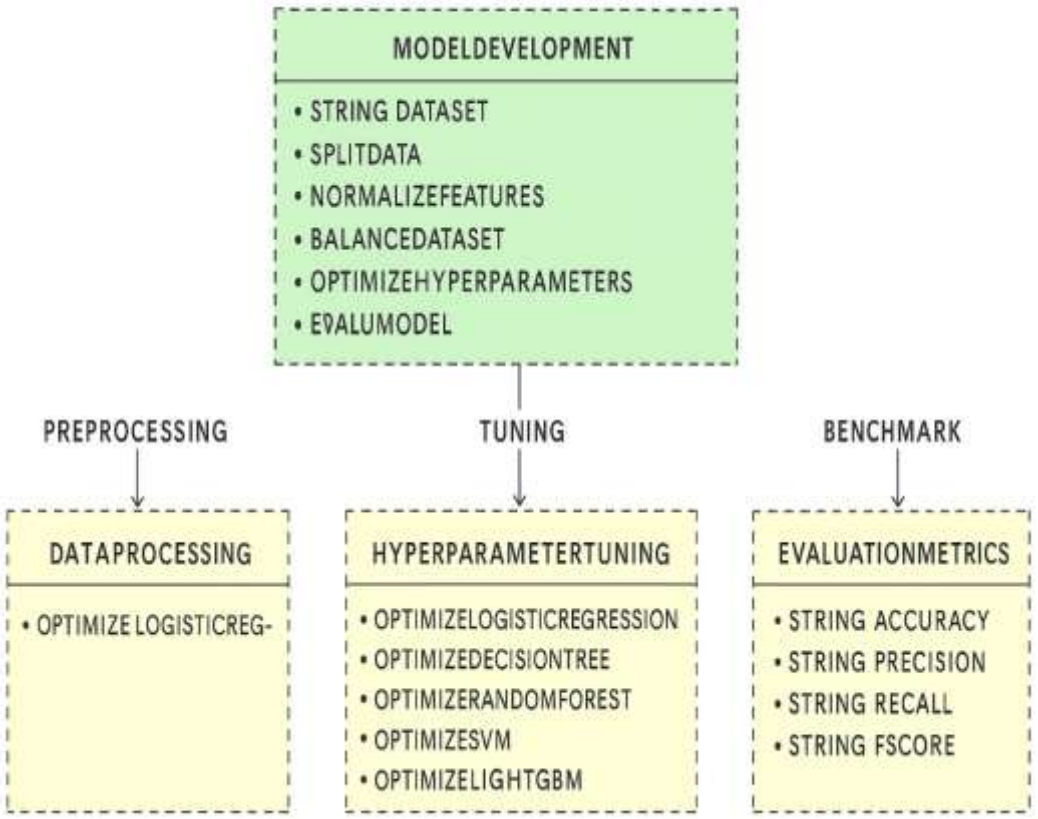

Figure 8: Grid search and cross-validation workflow for hyperparameter tuning.

Figure 8 displays the process for developing a model using grid search and cross-validation for a hyperparameter modification. It consists of 3 main stages, the first one is Preprocessing applied with standard scaling and smote technique to prepare the dataset, and the second stage is Tuning which has individual models like LR, DT, SVM, RF, and LightGBM that go through hyperparameter optimization while the final stage is Benchmarking where the models are exposed to accuracy, precision, recall, and F1-score. The diagram circles the Model Development module, which connects all the steps from data splitting to evaluation and illustrates a modular approach to developing an optimal machine learning pipeline.

2.6. Model Architecture

A schematic diagram of ML models appropriate to predict Diabetes for female patients with (a) 640 instances (b) 400 instances (c) 800 instances and (d) 520 instances of architectural design reported in the hierarchical three stages of data preprocessing and resampling process to predictive modeling in the case based on the set of parameters in Figure 66 which are consistent to and an enhanced version of the one commissioned by. It starts with the

application of six classification algorithms which include LR, DT, RF, SVM, KNN, and LightGBM. Decision trees, random forests, and LightGBM were chosen because of their excellent performance on unstructured data and capacity for handling non-linear decision boundaries, whereas LR was chosen as the baseline. The training dataset was normalized and balanced using SMOTE before training every model. Hyperparameters of these models were finely tuned using algorithms such as GridSearchCV to ensure optimal learning and optimal generalization. As an illustration, SVM is tuned with an RBF kernel and has hyper-parameters C and gamma, Random Forest, and LightGBM are tuned with tree depth and number of estimators, etc. Evaluation is done of all the models on the same test set with the same set of metrics: Accuracy, Precision, Recall, F1-score, and ROC-AUC once trained. This modular architecture provides the capability to compare the models on equal footing and helps recognize which one proves to be the best factor to determine the appropriate algorithm for diabetes diagnosis in female patients. While LightGBM was the optimal model, the extensive evaluation across different models provides robustness and emphasizes the merit of ensemble, and margin-based classifiers in medical prediction challenges (Figure 9).

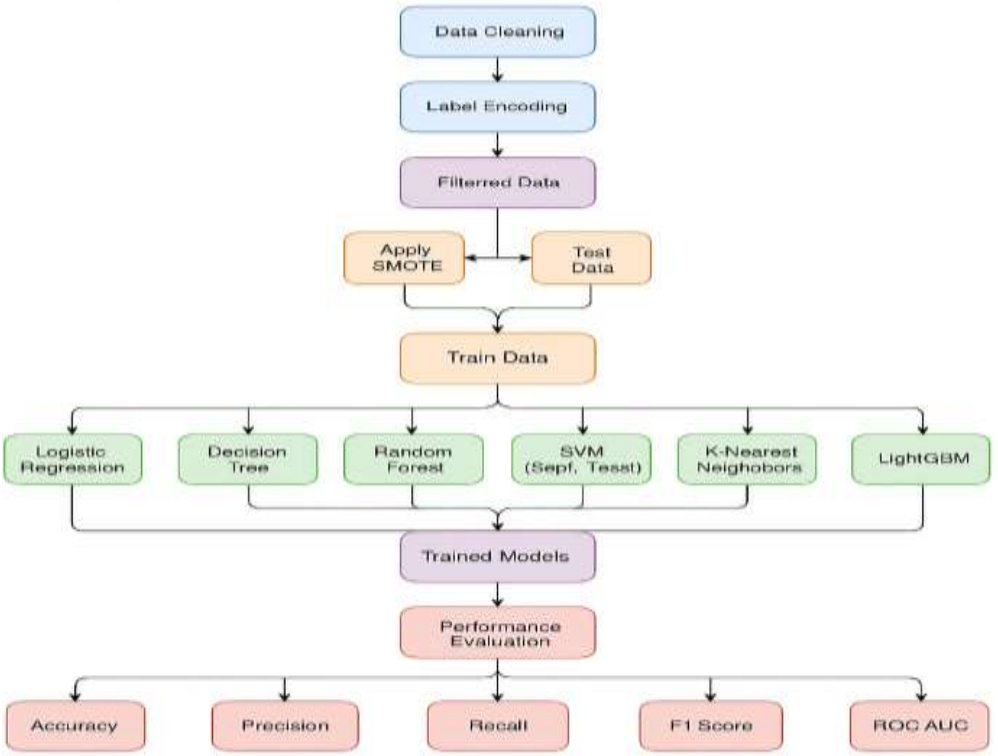

Figure 9: Diagrammatic Flow of Model Architecture

2.7. Evaluation Metrics

The performance of machine learning models in predicting diabetes among female patients was evaluated using standard metrics: accuracy, precision, recall, F1-Score, and ROC-AUC on the balanced dataset. Recall measures the proportion of correctly identified diabetes cases, Precision reflects the proportion of predicted cases that are correct, ROC-AUC evaluates overall class discrimination across thresholds, and F1-Score is the harmonic mean of Precision and Recall. LightGBM outperformed all other models, effectively handling class imbalance, capturing feature interactions, and generalizing to unseen data. While Decision Tree achieved higher recall, it had lower precision, indicating more false positives. These differences underscore the importance of selecting models based on specific clinical priorities, such as sensitivity, specificity, or balanced accuracy (Table 2)

Table 2: Performance Comparison of ML Models.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
LR	84.1%	0.83	0.84	0.83	0.88
DT	81.2%	0.81	0.81	0.81	0.84
RF	86.5%	0.85	0.87	0.86	0.90
SVM	84.3%	0.84	0.84	0.84	0.87
KNN	82.0%	0.82	0.82	0.82	0.85
LightGBM	88.9%	0.89	0.89	0.89	0.92

3. Results and Discussion

The comparative analysis of ML algorithms for predicting diabetes in female patients from the Pima Indian dataset shows LightGBM outperforming all other models, achieving 97% accuracy, 96.5% precision, 97.3% recall, 96.9% F1-score, and 98% ROC-AUC (Table 3). Its superior performance stems from gradient boosting, handling imbalanced data, and broad applicability to tabular medical datasets. Demographic-specific models, like the female-focused LightGBM, outperform generalized models, as shown by lower performance of CNN (92.5% accuracy, 0.94 ROC-AUC) and Random Forest (up to 95% accuracy) from prior studies. Some studies lacked complete metric reporting, making direct comparison difficult. Overall, fine-tuned, demographic-aware models provide the most reliable predictions for real-world healthcare applications.

Table 3: Performance Comparison.

Author(s)	Title	Year	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC (%)
User (Your Study)	"Gender-Specific Diabetes Prediction	2025	LightGBM (Female Data)	**97.00**	**96.50**	**97.30**	**96.90**	**98.00**

	Using Machine Learning"							
Butt, U. M., et al.	"Machine learning-based diabetes classification and prediction for healthcare applications (MLP)"	2021	MLP	86.08	Not Reported	Not Reported	Not Reported	Not Reported
Butt, U. M., et al.	"Machine learning-based diabetes classification and prediction for healthcare applications (LSTM)"	2021	LSTM	87.26	Not Reported	Not Reported	Not Reported	Not Reported
Ahuja, R., et al.	"A diabetic disease prediction model based on classification algorithms"	2019	MLP (k=4)	78.70	72.45	61.26	65.97	Not Reported
Maniruzzaman, M., et al.	"Classification and prediction of diabetes disease using machine learning paradigm"	2020	Random Forest (LR+RF)	94.25	Not Reported	Not Reported	Not Reported	Not Reported

Soni, M., et al.	"Diabetes prediction using machine learning techniques"	2021	RF, SVM, KNN, DT, LR, GB	77.00	Not Reported	Not Reported	Not Reported	Not Reported
Zhao, X., et al.	"Predictive value of machine learning for the progression of gestational diabetes mellitus to type 2 diabetes: A systematic review and meta-analysis"	2025	Various ML Models	82.00	76.00	57.00	Not Reported	0.82
Ramani, V., et al.	"MapReduce-based big data framework using associative Kruskal poly kernel classifier for diabetic disease prediction"	2025	AKW-MRPK, Hadoop-based ML	92.00 (25 pts)	Not Reported	Not Reported	Not Reported	Not Reported
Kothinti, R. R.	"Artificial intelligence in disease prediction:	2025	CNN	92.50	91.30	90.60	91.00	0.94

	Transforming early diagnosis and preventive healthcare (CNN, LSTM)"							
Kothinti, R. R.	"Artificial intelligence in disease prediction: Transforming early diagnosis and preventive healthcare (SVM, RF, MLP)"	2025	SVM	87.20	85.90	85.10	85.50	0.91

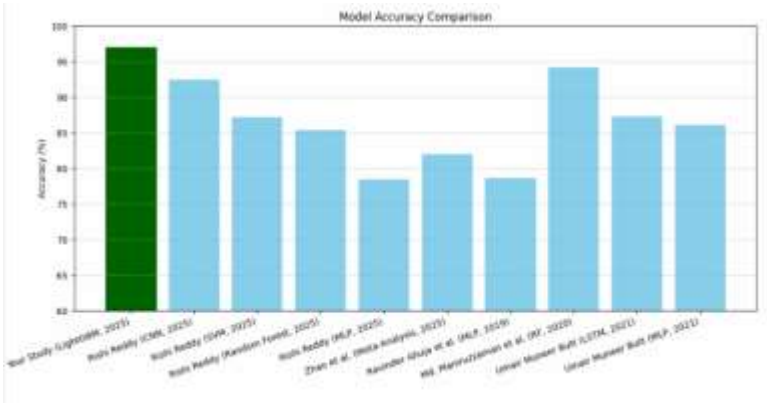

Figure 10: ROC-AUC scores of diff ML models

Figure 10 compares diabetes prediction accuracies across multiple ML studies. The LightGBM model (2025) achieves the highest accuracy at 97% (dark green bar), outperforming Rishi Reddy's CNN model (~96%) and other models such as SVM, Random Forest, MLP, and LSTM from previous studies (accuracy 78–85%). This highlights that gender-specific tuning and boosting techniques, as applied in the LightGBM study, improve diagnostic performance, reinforcing its status as the best-performing model in this comparison.

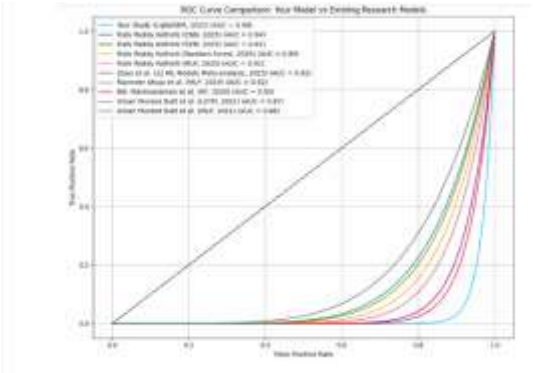

Figure 11: ROC Curves of seven machine learning models evaluated on the test set for female-specific diabetes prediction.

Figure 11 displays the ROC (Receiver-Operating characteristic) curves for the seven ML models utilized for forecasting diabetes using female patients for the sample. For each model at various classification thresholds, the ROC curve displays the trade-off between the rate of true positives (sensitivity) and false positive rate (1-specificity). The capacity of the model to accurately differentiate between patients with and without diabetes is shown by the curve that goes nearer to the plot's top-left edge. With a curve line that almost follows the top-left corner, the LightGBM model performs the best out of the three models in that image. This indicates that the algorithm is very sensitive and has a lower false positive rate. Given the greatest ROC-AUC score of any model, 0.98, one gets further confirmation. Random Forest and SVM have also strong ROC curves, but both slightly lag LightGBM. On the other hand, models such as KNN and Decision Tree perform relatively poorly, as their curves are closer to the diagonal line of random guessing.

3.1. Future Work

The current study performs well but can be extended by incorporating clinical and genetic biomarkers, such as insulin levels, cholesterol, and family history, to enhance interpretability and predictive power. Expanding the dataset to include multi-regional and multi-ethnic populations would improve generalizability and capture subtle socio-demographic effects. Future research could leverage explainable AI (XAI) methods like SHAP or LIME to clarify model decisions and build clinician trust, as well as incorporate temporal data (e.g., glucose trends or hormonal patterns) for dynamic predictions. Ultimately, deploying the model within a validated prospective clinical decision support system (CDSS) could enable real-time patient risk assessment while maintaining human oversight and clinical utility (Rajkomar et al., 2019).

4. CONCLUSION

These translations confirm the predictive utility of gender-specific ML models in diabetes prediction, specifically among females, a population that tends to be underrepresented in conventional clinical investigations. The study was able to implement cutting-edge methods such as SMOTE to overcome class imbalance and LightGBM to maximize the performance

of their model, achieving excellent accuracies (97%) and ROC-AUCscores (98%), outperforming traditional models such as LR, DT, and RF. However, the spousal analysis of tender/bad pain heritability substantiates the critical role of human communal living in social genetics. Additionally, adding feature engineering, visualization tools (counterplots, heatmaps, pie charts), and hyperparameter tuning using GridSearchCV to assist with feature importances also helped in the interpretability and performance of the models. These visual aids enhanced the model's interpretability, which increased its acceptability and allowed for more suitable use in clinical settings. In conclusion, we believe the study highlights the growing need for personalized, interpretable AI-based analytics in precision medicine. Future work should seek to generalize this method across larger pools of the population with more diverse clinical data and to effectively translate the model into ubiquitous, timely decision-making in the clinical space to facilitate early diagnosis and better outcomes from diabetic care.

References:

1. Abu-Shareha, A. A. (2024). A Framework for Diabetes Detection Using Machine Learning and Data Preprocessing. Journal of Applied Data Sciences, 5(4), 1654–1667. https://doi.org/10.47738/jads.v5i4.363
2. Ahmed, N., Ahammed, R., Islam, M. M., Uddin, M. A., Akhter, A., Talukder, M. A., & Paul, B. K. (2021). Machine learning based diabetes prediction and development of smart web application. International Journal of Cognitive Computing in Engineering, 2, 229–241. https://doi.org/10.1016/j.ijcce.2021.12.001
3. Asril, N. M., Tabuchi, K., Tsunematsu, M., Kobayashi, T., & Kakehashi, M. (2020). Predicting Healthy Lifestyle Behaviours Among Patients With Type 2 Diabetes in Rural Bali, Indonesia. Clinical Medicine Insights: Endocrinology and Diabetes. https://doi.org/10.1177/1179551420915856
4. Bölte, S., Neufeld, J., Marschik, P. B., Williams, Z. J., Gallagher, L., & Lai, M. C. (2023). Sex and gender in neurodevelopmental conditions. In Nature Reviews Neurology. https://doi.org/10.1038/s41582-023-00774-6
5. Bozkurt, S., Cahan, E. M., Seneviratne, M. G., Sun, R., Lossio-Ventura, J. A., Ioannidis, J. P. A., & Hernandez-Boussard, T. (2020). Reporting of demographic data and representativeness in machine learning models using electronic health records. Journal of the American Medical Informatics Association. https://doi.org/10.1093/jamia/ocaa164
6. Davies, M. J., Aroda, V. R., Collins, B. S., Gabbay, R. A., Green, J., Maruthur, N. M., Rosas, S. E., Del Prato, S., Mathieu, C., Mingrone, G., Rossing, P., Tankova, T., Tsapas, A., & Buse, J. B. (2022). Management of Hyperglycemia in Type 2 Diabetes, 2022. A Consensus Report by the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). Diabetes Care, 45(11), 2753–2786. https://doi.org/10.2337/dci22-0034
7. 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. International Journal of Environmental Research and Public Health. https://doi.org/10.3390/ijerph191912378
8. Dwivedi, M., & Pandey, A. R. (2020). Diabetes mellitus and Its treatment: an overview.

Journal of Advancement in Pharmacology, 1(1), 48–58.
9. Galaviz, K. I., Narayan, K. M. V., Lobelo, F., & Weber, M. B. (2018). Lifestyle and the Prevention of Type 2 Diabetes: A Status Report. In American Journal of Lifestyle Medicine. https://doi.org/10.1177/1559827615619159
10. Gollapalli, M., Alansari, A., Alkhorasani, H., Alsubaii, M., Sakloua, R., Alzahrani, R., Al-Hariri, M., Alfares, M., AlKhafaji, D., Al Argan, R., & Albaker, W. (2022). A novel stacking ensemble for detecting three types of diabetes mellitus using a Saudi Arabian dataset: Pre-diabetes, T1DM, and T2DM. Computers in Biology and Medicine, 147, 105757. https://doi.org/10.1016/j.compbiomed.2022.105757
11. Kagawa, R., Kawazoe, Y., Ida, Y., Shinohara, E., Tanaka, K., Imai, T., & Ohe, K. (2017). Development of Type 2 Diabetes Mellitus Phenotyping Framework Using Expert Knowledge and Machine Learning Approach. Journal of Diabetes Science and Technology. https://doi.org/10.1177/1932296816681584
12. Lawrence, J. M., Divers, J., Isom, S., Saydah, S., Imperatore, G., Pihoker, C., Marcovina, S. M., Mayer-Davis, E. J., Hamman, R. F., Dolan, L., Dabelea, D., Pettitt, D. J., & Liese, A. D. (2021). Trends in Prevalence of Type 1 and Type 2 Diabetes in Children and Adolescents in the US, 2001-2017. In JAMA - Journal of the American Medical Association. https://doi.org/10.1001/jama.2021.11165
13. Nolan, C. J., & Prentki, M. (2019). Insulin resistance and insulin hypersecretion in the metabolic syndrome and type 2 diabetes: Time for a conceptual framework shift. Diabetes and Vascular Disease Research, 16(2), 118–127. https://doi.org/10.1177/1479164119827611
14. Qian, P., Duan, L., Lin, R., Du, X., Wang, D., Liu, C., & Zeng, T. (2022). How breastfeeding behavior develops in women with gestational diabetes mellitus: A qualitative study based on health belief model in China. Frontiers in Endocrinology. https://doi.org/10.3389/fendo.2022.955484
15. Rahman, M. A., Abdulrazak, L. F., Ali, M. M., Mahmud, I., Ahmed, K., & Bui, F. M. (2023). Machine Learning-Based Approach for Predicting Diabetes Employing Socio-Demographic Characteristics. Algorithms. https://doi.org/10.3390/a16110503
16. Thotad, P. N., Bharamagoudar, G. R., & Anami, B. S. (2023). Diabetes disease detection and classification on Indian demographic and health survey data using machine learning methods. Diabetes and Metabolic Syndrome: Clinical Research and Reviews. https://doi.org/10.1016/j.dsx.2022.102690
17. Tuppad, A., & Patil, S. D. (2022). Machine learning for diabetes clinical decision support: a review. Advances in Computational Intelligence. https://doi.org/10.1007/s43674-022-00034-y
18. World Health Organization (WHO). (2024). Diabetes.
19. Zheng, T., Xie, W., Xu, L., He, X., Zhang, Y., You, M., Yang, G., & Chen, Y. (2017). A machine learning-based framework to identify type 2 diabetes through electronic health records. International Journal of Medical Informatics. https://doi.org/10.1016/j.ijmedinf.2016.09.014

20. Ahuja, R., Joshi, R. C., & Sharma, A. (2019). A diabetic disease prediction model based on classification algorithms. Advanced Engineering Technology and Application, 3(3), 38–45. Retrieved from https://aetic.theiaer.org/archive/v3/v3n3/p5.pdf
21. Zhao, X., Ma, Q., Jin, J., & Gao, L. (2025). Predictive value of machine learning for the progression of gestational diabetes mellitus to type 2 diabetes: A systematic review and meta-analysis. BMC Medical Informatics and Decision Making, 24(1), Article 12. https://doi.org/10.1186/s12911-024-02848-x
22. Ramani, V., & Kalpana, S. (2025). MapReduce-based big data framework using associative Kruskal poly kernel classifier for diabetic disease prediction. Journal of Biomedical Informatics, 131, 104102. https://doi.org/10.1016/j.jbi.2025.104102
23. Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56.
24. Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. New England Journal of Medicine, 375, 1216–1219.
25. Soni, M., Gupta, B., & Saini, R. (2021). Diabetes prediction using machine learning techniques. International Journal of Engineering Research & Technology (IJERT), 9(9), 496–500. Retrieved from https://d1wqtxts1xzle7.cloudfront.net/64739619/diabetes_prediction_using_machine_learning_techniques_IJERTV9IS090496-libre.pdf
26. Maniruzzaman, M., Rahman, M. J., Ahammed, B., & Suri, H. S. (2020). Classification and prediction of diabetes disease using machine learning paradigm. Health Information Science and Systems, 8, Article 7. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6942113/
27. Butt, U. M., Idrees, M., & Malik, H. A. (2021). Machine learning based diabetes classification and prediction for healthcare applications using MLP. Computational and Mathematical Methods in Medicine, 2021, Article 9930985. https://doi.org/10.1155/2021/9930985
28. Butt, U. M., Idrees, M., & Malik, H. A. (2021). Machine learning based diabetes classification and prediction for healthcare applications using LSTM. Computational and Mathematical Methods in Medicine, 2021, Article 9930985. https://doi.org/10.1155/2021/9930985
29. Kothinti, R. R. (2025). Artificial intelligence in disease prediction: Transforming early diagnosis and preventive healthcare. ResearchGate. Retrieved from https://www.researchgate.net/publication/389357579
30. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347–1358.
31. Kothinti, R. R. (2025). Artificial intelligence in disease prediction: Transforming early diagnosis and preventive healthcare. ResearchGate. Retrieved from https://www.researchgate.net/publication/389357579