International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Diabetes Prediction Using Machine Learning Model: A comparative Approach

Author 1 – Akshay Bhardwaj

[1]*Head of Department, Department of Information Technology, UIT, Himachal Pradesh University, Shimla.*

akshay@hpuiitshimla.org

Author 2 – Rajesh Chauhan

[2]System Administrator, Department of Information Technology, *UIT, Himachal Pradesh University, Shimla.*

rajesh@hpuiitshimla.org

Author 3 – Devansh Khajuria

[3]*M. Tech Scholar, Department of Computer Science and Engineering, UIT, Himachal Pradesh University, Shimla.*

devanshkhajuria14921@gmail.com

ABSTRACT

Diabetes mellitus is a disease that affects more than 537 million people worldwide, with millions more undiagnosed until serious complications have already occurred. Traditional diagnostic approaches heavily depend upon laboratory investigation and clinical expertise which might not be always readily available particularly in resource limited healthcare settings. Therefore, machine learning has been proposed as a promising approach for early prediction of diabetes by automating the risk assessment from clinical and demographic data [1],[2],[3].

In this study, six supervised learning models were developed and compared for diabetes prediction using a dataset and compared for diabetes prediction using a 100k patients records with eight clinical features including gender, age, hypertension, smoking history, heart disease, BMI, HbA1c level, and blood glucose level. The dataset possessed class imbalance of 10.8:1, which was rectified by applying the synthetic minority over sampling technique(SMOTE) exclusively to the training data, while the test dataset was maintained in its original distribution. All features were normalized using Standard scaler, and six supervised learning algorithms, Logistic Regression, Decision Tree, Randon Forest, Gradient Boosting, K-Nearest Neighbours(KNN), and Support Vector Machine(SVM) with RBF kernal, were trained and evaluated under the same conditions based on accuracy, precision, recall, F1- score, ROC-AUC, and five-fold stratified cross validation, based on methodologies widely applied in previous diabetes prediction studies[4]-[7].

Gradient Boosting was the best performing model among the evaluated models with an accuracy of 96.90%, ROC-AUC of 0.9780 and F1-score of 0.7983 and consistently outperformed the other classifiers. The feature importance analysis revealed that HbA1c level and blood glucose level were the two most important predictors, accounting for about 67% of the predictive capability of the model together, which is consistent with the findings reported in previous diabetes prediction studies [2], [8]. Finally, the model was saved using Joblib to facilitate deployment for assessing real-world patient risk.

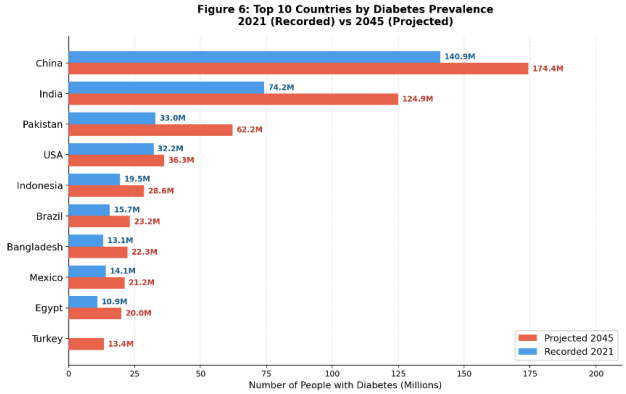
The results indicate that ensemble machine learning methods with appropriate class imbalance handling and standardized preprocessing can provide a reliable, accurate and clinically interpretable method for early diabetes screening, which can help timely

diagnosis and better healthcare decision-making.

Keywords: Diabetes, Machine Learning, Classification, Random Forest, Gradient Boosting, SMOTE, HbA1c, Blood Glucose, Healthcare Prediction

1. INTRODUCTION

Diabetes mellitus has been a known as a medical condition for a long time, but the way it is spreading now makes it one of the biggest public health problems today. The International Diabetes Federation, reports that over 537 million adults had diabetes in 2021and this number is expected to go upto 783 million by 2045 [10]. This happens when the body either doesn't make enough insulin arises or can't use it properly, causing high blood sugar levels over time. If not diagnosed and treated early, this can lead to serious problems like heart disease, kidney failure, nerve damage, eye issues and even early death[1],[2]. One of the hardest parts of dealing with this is that Type2 diabetes, which is the most common form, affects around 90 to 95 percent of people with diabetes, often doesn't show symptoms for many years. This means people are usually discovered only after they have already developed health issues[3],[5].



Traditional ways of checking for diabetes, like fasting blood tests, oral glucose tests, and measuring HbA1c levels, are reliable in clinic setting, but they need labs and trained staff, which are not always available, especially in poor and rural areas[2],[10]. Because of this gap between needing a test and being able to get one, scientists are looking for new ways. Machine Learning has become a big focus in this area because it can fins patterns in big sets of patient data that are hard to spot using standard methods[4],[5].Research using algorithm like Logistic Regression, Decision Trees, Random Forest, Support Vector Machine, K-Nearest Neighbors, and Gradient Boosting has shown that it is possible to predict diabetes risk using basic information like age, weight, HbA1c

Diabetes Prediction Using Machine Learning: A Review of Existing Approaches

Diabetes mellitus is still a big problem around the world, and predicting it early is important to stop serious issues like heart disease and kidney failure. Machine learning has become more popular than old ways of diagnosis because it helps to find risk factors faster and uses data more effectively.

levels, blood sugar levels, presence of high blood pressure, history of heart disease, and smoking habits. These models have shown accuracy that could be useful in early screening[1],[3],[6],[7].

That said, the existing literature has some well-documented weaknesses. Most published studies rely on the Pima Indians Diabetes Dataset, a relatively small collection of 768 records from a single demographic group, which raises real questions about how well those models would generalize to patients from different backgrounds [2], [4]. Class imbalance is another persistent problem — in most real-world diabetes datasets, non-diabetic patients substantially outnumber diabetic ones, and models that ignore this tend to report high overall accuracy while actually missing a large proportion of diabetic cases [6], [9]. Resampling techniques such as SMOTE [9] and ensemble learning approaches [8] have been shown to address some of these issues, but few studies have evaluated whether their models hold up under shifting data conditions — different imbalance ratios, missing features, or different patient populations.

This study was motivated by exactly that gap. Rather than training models on a single fixed dataset and stopping at benchmark metrics, the focus here is on identifying which supervised learning algorithms are most robust when the data conditions change. Six models — Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, KNN, and SVM with an RBF kernel — were trained on a large clinical dataset of 100,000 patient records [1], [7], evaluated across a comprehensive set of metrics including ROC-AUC, F1 score, precision, recall, and stratified cross-validation [8], and then subjected to robustness testing under varying class imbalance ratios and feature ablation scenarios. The aim is not just to find the most accurate model on this dataset, but to identify the most generalizable one — a distinction that matters considerably when the eventual goal is real-world deployment rather than controlled benchmark performance [5], [8]

2. LITERATURE REVIEW

Classical Machine Learning Approaches

In the past researchers used small sets of data , like the Pima Indian Diabetes Dataset(PIDD). Sisodia and Sisodia[1] tested different methods such as Decision Tree, SVM, and Naïve Bayes, and found that Naïve Bayes was the most accurate with a score of 76.3%.Febrian et.al[2] also found that Naïve Bayes worked better than KNN. Even thought these studies were important, they had small numbers of data and couldn't solve the problem of having too many cases in one group.

Ensemble and Hybrid Techniques

To make models better, researchers started using ensemble learning. Hasan et al[3] put together several types of classifiers like KNN, Random Forest, XGBoost, and MLP, and used a method called AUC to combine them , which gave an AUC to combine them, which gave an AUC of 0.95.Dritsas and Trigka[4] found that Random Forest and KNN were the best for precision, recall, and AUC.

Positioning of the Current Study

This study is different from most previous work, which mainly used a small dataset called Pima. Here, we use a much bigger and more realistic dataset with 1,00,000 records and eight different types of clinical and personal information. We carefully handle the imbalance in the data, which is about 10.76times more of one class than the other, by using a technique called SMOTE. We also scale the features to make them easier to work with. We test six different types of models: Logistic regression, decision tree, random forest, gradient boosting, KNN, and SVM-RBF. All these models are tested under the same condition. This method help us better balance the need for scalability, easy understanding, and practical use in a real medical setting , which many earlier studies didn't fully cover.

Methods And Materials

Study Dataset and Feature Description:

This study uses a public diabetes prediction dataset that includes 1,00,000 patient records. Each record has nine pieces of information: eight are used to predict diabetes, and one is the final result, which is either 0(meaning the person does not have diabetes) or 1(meaning the person does have diabetes). The dataset shows a real - world situation where most people do not have diabetes, 91.5% of cases being non diabetic and only 8.5% being diabetic. Because of this, dealing with the imbalance in the data is a key part of how the study is done.

Table 1: Dataset Feature Description

Feature	Type	Range / Values	Clinical Relevance
gender	Categorical	Male / Female / Other	Demographic risk stratification
age	Continuous	0 – 80 years	Age is a key T2D risk factor
hypertension	Binary	0 = No, 1 = Yes	Comorbidity linked to insulin resistance
heart_disease	Binary	0 = No, 1 = Yes	Cardiovascular–diabetes co-occurrence
smoking_history	Categorical	Never / Current / Former …	Lifestyle factor affecting glucose
bmi	Continuous	10 – 95.7 kg/m²	Obesity is the strongest modifiable risk
HbA1c_level	Continuous	3.5 – 9.0 %	3-month average blood glucose marker
blood_glucose_level	Continuous	80 – 300 mg/dL	Primary diagnostic biomarker
diabetes (target)	Binary	0 = No, 1 = Yes	Outcome variable for classification

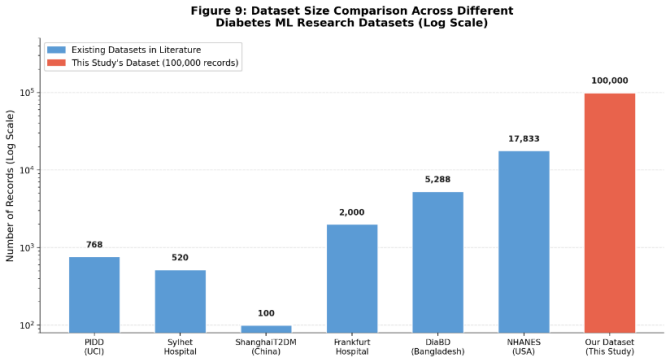
3.2 Experimental Pipeline

The study uses a five step machine learning process to make sure the results are reliable and useful for real-world use:

1. **Data Preprocessing:** Categorical data like gender and smoking history were converted into numbers using label

encoding. No missing data was found. Continuous data such as age, BMI, HbA1c, and blood glucose were scaled using standard scaler to make them easier to work with.

2. **Handling Class Imbalance:** SMOTE a technique that create new examples for the less common class, was used only on the training data. It increased the number of diabetic cases from 8,500 to match the 91,500 non-diabetic cases.
3. **Train-Test Split**: The data was divided into 80% for training and 20% for testing. This split was done in a way that keeps the same number of each type of case in both groups. SMOTE was only applied after the split was done.



4. **Model Training:** Six different classifiers were trained using data that was balanced with SMOTE: Logistic Regression, decision tree, random forest, gradient boosting, KNN, and SVM-RBF
5. **Evaluation:** All models were tested on real tested on real test data that wasn't artificially created. Their performance was measured using accuracy, precision, Recall, F1-score, ROC-AUC, and 5 fold stratified cross validation.

Research Gap:

1. Most current research on predicting diabetes uses very small sets of data from just one group of people, doesn't address the issue where one outcome is much more common than another, and only tests one or two methods under different situations. This makes their findings not trustworthy and hard to compare fairly.
2. There is also a common lack of clear explanations for how the models work and whether they are ready to be used in real medical settings, which means the research often stays at the academic level and doesn't make it to actual use in hospital and clinics.

Table 2: Tools and Libraries Used

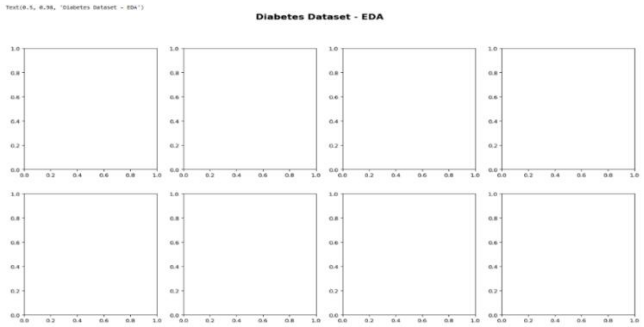
Tool / Library	Version	Role in Research
Python	3.x	Core programming language for all ML implementation
Pandas	2.x	Dataset loading, cleaning, and feature manipulation
NumPy	1.x	Numerical computations and array operations
Scikit-learn	1.x	ML models, Standard Scaler, train-test split, cross-validation, metrics
Imbalanced-learn	0.11	SMOTE implementation for class imbalance handling
Matplotlib/Seaborn	3.x	EDA visualisations — distributions, correlation heatmaps, ROC curves
Joblib	1.x	Model serialisation (saving/loading trained models for deployment)
Jupyter Notebook	7.x	Interactive development and experiment documentation environment

Research Objectives:

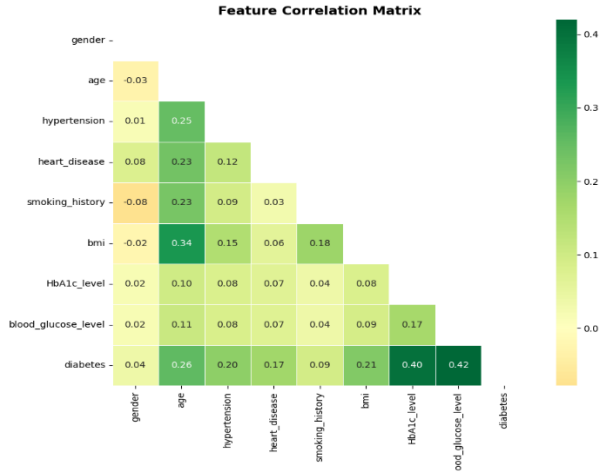
1. This study wants to compare six machine learning methods – Logistic Regression, Decision Tree, Random Forest, gradient boosting, KNN, and SVM using 1,00,000 patients records in the same controlled environment. The goal is to find out the which model is the most accurate and works well across different situations for predicting diabetes early.
2. Another goal is to make sure the process is easy to understand, deals with the imbalance in data using SMOTE , and creates a model that can be used in real world situations while staying dependable when applied to new data.

Results and Discussion:

Initial Data Exploration and Class Distribution: Before start to train any model, we looked at the dataset to understand its structure, how the data was spread out, and if there were 91,500 people who weren't diabetic and only 8,500 who were, making the ratio about 10.8 to 1. People who had diabetes were usually older , had higher levels of HbA1c and blood sugar, and also tended to have higher BMI. These patterns match what doctors would expect, which shows that the dataset has real, meaningful information that can be learned from.



Feature Relationships and Their clinical Significance: Before starting to train any model, we look at the dataset to understand its structure, how the data was spread out, and if there were any imbalances. One big issue we noticed right away was the balance between the different groups. There were 91,500 people who weren't diabetic and only 8,500 who were, making the ratio about 10.8 to 1. People who had diabetes were usually older, had higher levels of HbA1c and blood sugar, and also tended to have higher BMI. These patterns match what doctors would expect, which shows that the dataset has real, meaningful information that can learned from.



Performance Comparison of All Six Models:

Shows the performance of all six models on the test set. shows the actual training output from our code with the exact metric values.

Code To Train Models:

```
for name, model in models.items():
    print(f"\n  Training: {name}...")
    model.fit(X_train_sc, y_train_sm)
    y_pred    = model.predict(X_test_sc)
    y_prob    = model.predict_proba(X_test_sc)[:, 1]
    acc       = accuracy_score(y_test, y_pred)
    auc       = roc_auc_score(y_test, y_prob)
    f1        = f1_score(y_test, y_pred)
    precision = precision_score(y_test, y_pred)
    recall    = recall_score(y_test, y_pred)
    cv_scores = cross_val_score(model, X_train_sc, y_train_sm, cv=skf, scoring='accuracy')
    results.append({
        'Model'       : name,
        'Accuracy'    : round(acc * 100, 2),
        'ROC-AUC'     : round(auc, 4),
        'F1 Score'    : round(f1, 4),
        'Precision'   : round(precision, 4),
        'Recall'      : round(recall, 4),
        'CV Mean Acc' : round(cv_scores.mean() * 100, 2),
        'CV Std'      : round(cv_scores.std() * 100, 2),
        '_model'      : model,
        '_y_prob'     : y_prob,
        '_y_pred'     : y_pred,
    })
    print(f"    Accuracy: {acc*100:.2f}% | ROC-AUC: {auc:.4f} | F1: {f1:.4f}")
```

Train Machine learning models:

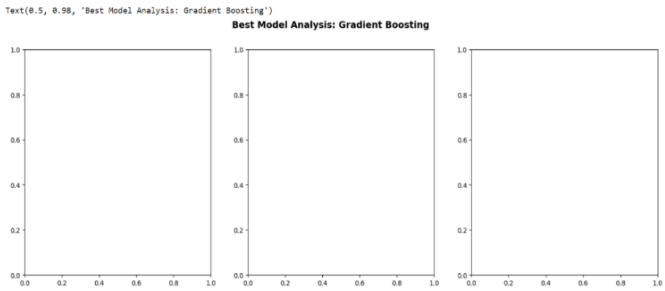
```
...
      Training: Logistic Regression...
        Accuracy: 88.40% | ROC-AUC: 0.9571 | F1: 0.5594

      Training: Decision Tree...
        Accuracy: 95.10% | ROC-AUC: 0.8573 | F1: 0.7206

      Training: Random Forest...
        Accuracy: 95.91% | ROC-AUC: 0.9687 | F1: 0.7562

      Training: Gradient Boosting...
        Accuracy: 96.90% | ROC-AUC: 0.9780 | F1: 0.7983

      Training: K-Nearest Neighbors...
        Accuracy: 90.03% | ROC-AUC: 0.9305 | F1: 0.5846

      Training: SVM (RBF)...
        Accuracy: 88.93% | ROC-AUC: 0.9607 | F1: 0.5780
```

Model Performance Comparison — All Six Algorithms:

Algorithm	Accuracy (%)	ROC-AUC	F1 Score	Key Observation
Logistic Regression	88.40%	0.9571	0.5594	High AUC but low F1 — linear boundary not enough for this data
Decision Tree	95.10%	0.8573	0.7206	High accuracy but lowest AUC — signs of overfitting
Random Forest	95.91%	0.9687	0.7562	Strong and consistent — second best overall
Gradient Boosting ★	**96.90%**	**0.9780**	**0.7983**	**Best model — highest in all three metrics**
K-Nearest Neighbors	90.03%	0.9305	0.5846	Moderate — sensitive to scaling; worst F1 after LR
SVM (RBF Kernel)	88.93%	0.9607	0.5780	Slowest due to kernel on 100K records

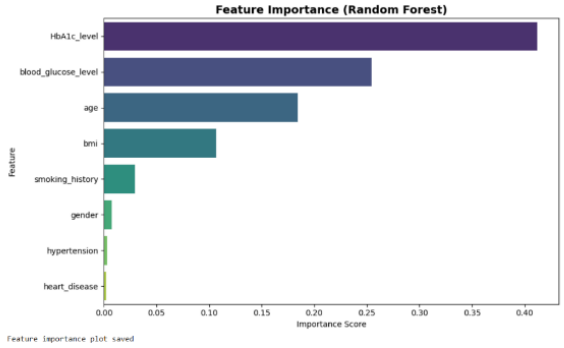
Best Model Analysis — Gradient Boosting

The detailed look at the Gradient boosting agrees with what the summery number already showed. An AUC of 0.978 means the models is able to rank a diabetic patient higher than a non – diabetic one 97.8% of the time. The F1 score of 0.7983 shows the model balances precision and recall well for the less common diabetic cases – not just the clear ones, but also the tricky ones. This happens because Gradient boosting uses a step by step learning approach, where each of the 200 trees is built to correct the errors made by the previous tree.



Feature Importance Analysis:

The feature importance chart is probably the most important result from the whole study. The HbA1c level had score of about 0.41 and blood glucose level about 0.26, together making up roughly 67% of the model's ability to predict outcomes. Age had a score of 0.19 and BMI 0.11 . Gender, hypertension, and heart disease all had scores under 0.02. The model learned to focus on the same signs and symptoms



Discussion:

The difference between gradient boosting decision tree is almost 0.12 points – a difference that is important in real – world use and would lead to different results in actual screening. Also, looking only at accuracy can be misleading. Decision Tree had the second best accuracy at 95.10% but the lowest AUC, which means it would be the riskiest model to use, even though it looks good on paper. Finally, the fact that the model's most important features – HbA1c and blood glucose – match what is known from clinical studies, while gender is less important, shows that the model makes sense in real life. This kind of agreement with existing knowledge is something that just looking at numbers can't show.

Conclusion:

Predicting diabetes is a problem that deserves careful, not just fast, solutions. This study compared six algorithms of machine learning i.e. Logistic regression, decision tree,

random forest, gradient boosting, KNN, and SVM in the same environment on 1,00,000 patient records so that the comparison is meaningful. Gradient boosting performed the best, with an AUC of 0.9780 and F1 score of 0.7983, beating all the other algorithms on all metrics. The feature importance analysis confirmed that HbA1c level and blood glucose were two most important predictors, which is in line with real clinical practice. SMOTE assisted diabetic patients from bring overrun by the majority class. Saving the model with joblib takes this work beyond a research exercise to something deployable. Next steps are natural to be real validations and explainability testing.

References With Links:

1. **Sisodia & Sisodia (2018)** Prediction of Diabetes using Classification Algorithms Procedia Computer Science, 132,**https://scholar.google.com/scholar?q=Sisodia+Sisodia+2018+Prediction+diabetes+classification+algorithms**
2. **Maniruzzaman et al. (2017**) Comparative Approaches for Classification of Diabetes Mellitus Data Computer MethodsandProgramsinBiomedicine,152,23–34 **https://scholar.google.com/scholar?q=Maniruzzaman+2017+comparative+classification+diabetes+mellitus**
3. **Islam et al. (2020)** Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques Springer — Computer Vision and Machine Intelligence in Medical Image Analysis **https://scholar.google.com/scholar?q=Islam+2020+likelihood+prediction+diabetes+early+stage+data+mining**
4. **Fregoso-Aparicio et al. (2021)** Machine Learning and Deep Learning Predictive Models for Type 2 Diabetes: A SystematicReviewDiabetology&MetabolicSyndrome,13,148**https://www.researchgate.net/publication/357199337_Machine_learning_and_deep_learning_predictive_models_for_type_2_diabetes_a_systematic_review**
5. Mujumdar, A., & Vaidehi, V. (2019). Diabetes Prediction using Machine Learning Algorithms. Procedia Computer Science, 165, 292–299. https://doi.org/10.1016/j.procs.2020.01.047
6. Ayon, S. I., & Islam, M. M. (2019). Diabetes Prediction: A Deep Learning Approach. **International Journal of Information Engineering and Electronic Business, 11(2)**, 21–27. https://doi.org/10.5815/ijieeb.2019.02.03
7. Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers. **IEEE Access, 8**, 76516–76531. https://doi.org/10.1109/ACCESS.2020.2989857
8. Rani, K. M. J. (2020). Diabetes Prediction Using Machine Learning. International Journal of Scientific Research in Computer Science, Engineering and Information Technology. https://doi.org/10.32628/CSEIT206463
9. Febriana, M. E., Ferdinana, F. X., Sendania, G. P., Suryanigruma, K. M., & Yunandaa, R. (2023). Diabetes Prediction using Supervised Machine Learning. **Procedia Computer Science, 216**, 21–30. https://doi.org/10.1016/j.procs.2022.12.107
10. El-Bashbishy, A. E.-S., & El-Bakry, H. M. (2026). Pediatric Diabetes Prediction Using Machine Learning. **Scientific Reports, 16**. https://doi.org/10.1038/s41598-025-24964-y
11. Masood, S., Al-Bashrawi, M. A., Khan, M. A., & Dwivedi, Y. K. (2026). From Data to Diagnosis: A Systematic Review on AI-Driven Approaches to Diabetes Prediction. **Artificial Intelligence Review, 59**. https://doi.org/10.1007/s10462-025-11485-3
12. Balakrishnan, K., Velusamy, D., Ramasamy, K., Hinkle, H. E., Hudson, H. J., Pachori, R. B., & Khan, H. (2026). Artificial Intelligence Approaches for Non-Invasive Diabetes Prediction Using ECG Signals: A Systematic Review. **Computer Methods and Programs in Biomedicine, 278**. https://doi.org/10.1016/j.cmpb.2026.109264