

EARLY STAGE DIABETES RISK PREDICTION AND PREVENTION SYSTEM USING ML

¹ V.Latha sri, ² Bommidi Hasini, ³ Erukula Akshaya, ⁴ Ananthagiri Madhuri, ⁵ M.Sridhar

¹Assistant Professor, ²³⁴⁵Students

Department of ECE

Siddhartha Institute of Technology & Sciences, Narapally

lathasri@siddhartha.org.in, 24tq1a0446@siddhartha.co.in, 24tq1a0414@siddhartha.co.in,

24tq1a0444@siddhartha.co.in, 25tq5a0403@siddhartha.org.in

ABSTRACT

Diabetes mellitus is a chronic metabolic disorder that affects a very large and growing number of people, and a considerable share of them remain undiagnosed until complications such as nerve damage, kidney disease, or vision loss have already begun. Early detection allows lifestyle changes and medical treatment to slow or even prevent the progression of the disease. This paper presents an early stage diabetes risk prediction and prevention system that uses machine learning to estimate an individual's risk from simple symptoms, clinical measurements, and lifestyle information, and then provides practical prevention guidance based on that risk.

The system uses two kinds of input. The first is a symptom questionnaire covering signs that commonly appear in the early stage of diabetes, such as excessive urination, excessive thirst, sudden weight loss, weakness, blurred vision, delayed healing, itching, and irritability, together with age and gender. The second is a set of clinical and lifestyle measurements, including fasting glucose, body mass index, blood pressure, waist circumference, family history, physical activity, and dietary habits. Where available, readings from a glucometer, a digital blood pressure monitor, and a weighing scale can be entered directly, which reduces transcription errors.

The collected data is cleaned, encoded, and scaled, and missing clinical values are handled carefully because zero readings in such datasets often represent missing measurements rather than true values. Several classifiers are trained and compared, including logistic regression, k-nearest neighbours, a support vector machine, a decision tree, random forest, and gradient boosting. Because missing a diabetic case is more harmful than a false alarm, the models are evaluated with particular attention to recall and F1 score in addition to accuracy, precision, and the area under the ROC curve.

The selected model assigns each user to a low, moderate, or high risk category along with a probability value. The prevention module then generates advice that matches the risk level and the specific factors contributing to it, such as increasing daily physical activity, reducing refined carbohydrate intake, managing weight, or consulting a physician for a confirmatory HbA1c test. A web dashboard allows users to record repeated measurements over time and view their risk trend, while health workers can review the risk profile of a screened community.

The system is intended as a screening and awareness tool rather than a diagnostic instrument, and every high-risk result is accompanied by a recommendation to seek laboratory confirmation. It can be used in college health centres, primary health

campus, and workplaces where large numbers of people can be screened quickly at low cost. Future enhancements include integration with Bluetooth-enabled glucometers and wearable activity trackers, support for regional languages, and periodic retraining with locally collected data.

I. INTRODUCTION

Diabetes has become one of the most widespread non-communicable diseases in the world, and India carries one of the largest numbers of affected adults. Type 2 diabetes, which accounts for the great majority of cases, develops gradually over several years. During this period many people experience mild symptoms that they do not associate with a serious condition, and because routine screening is not common outside hospitals, the disease is often diagnosed only when complications appear.

Clinical diagnosis relies on laboratory tests such as fasting plasma glucose, oral glucose tolerance, and glycated haemoglobin. These tests are reliable, but people must first decide to take them, and that decision usually depends on awareness of risk. In rural areas and among young working adults, access to a physician is limited and preventive check-ups are rare. A simple tool that can estimate risk from symptoms and basic measurements, and that can explain in plain terms why the risk is high, would encourage people to seek proper testing much earlier.

Machine learning is well suited to this screening problem. Publicly available and hospital-collected datasets contain records of patients with and without diabetes together with their symptoms, measurements, and history. Classifiers trained on these records can learn the combinations of factors associated with the disease and can apply them to new individuals in a fraction of a second. From an electronics perspective, low-cost sensing devices such as glucometers, blood pressure monitors, and smart scales already provide the numerical inputs needed, which makes it practical to combine measurement and prediction in one workflow.

This paper describes a system that brings these elements together. It collects symptom and measurement data, preprocesses it, compares several classification models, selects the most suitable one with emphasis on recall, and presents the result as an understandable risk level with personalised prevention advice. A dashboard supports repeated use so that changes in risk can be tracked over time. The aim is to shift attention from late diagnosis towards early awareness and prevention, while leaving diagnosis and treatment firmly in the hands of qualified doctors.

II. LITERATURE SURVEY

Research on diabetes prediction using computational methods began with statistical risk scores derived from large population studies. These scores assign points to factors such as age, body mass index, waist circumference, family history, physical inactivity, and history of high blood pressure, and the total is used to classify a person as low or high risk. Such scores have been validated in several countries and are widely used in public health programmes because they are simple and require no laboratory tests. However, studies comparing them with measured glucose values have reported that

their sensitivity varies considerably between populations, particularly in South Asian groups who tend to develop diabetes at lower body mass index values. Researchers have therefore suggested that data-driven models trained on local populations could improve screening accuracy, while retaining the practical advantage of relying on easily obtained information that a health worker can collect in a few minutes.

The widely used Pima Indians diabetes dataset has served as a benchmark for a large number of machine learning studies. It contains measurements such as glucose concentration, blood pressure, skin fold thickness, insulin, body mass index, a pedigree function, and age for several hundred women. Studies using this dataset have applied logistic regression, naive Bayes, decision trees, k-nearest neighbours, support vector machines, and neural networks, typically reporting accuracy between seventy and eighty percent. A recurring observation is that many records contain zero values for measurements such as insulin and skin thickness that cannot physically be zero, so the treatment of these missing values strongly affects the reported results. Authors who used median imputation or model-based imputation generally obtained more reliable outcomes than those who used the raw values, and several recommended that preprocessing choices be reported as carefully as model choices.

A second line of work has focused on symptom-based datasets collected through questionnaires administered in hospitals. One commonly used dataset records age, gender, and the presence or absence of symptoms such as polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, and obesity. Because these features are mostly binary and directly related to the clinical picture of early diabetes, tree-based and ensemble classifiers have achieved high accuracy on such data, in some studies exceeding ninety five percent. Researchers have noted that polyuria and polydipsia consistently emerge as the strongest predictors, which agrees with clinical understanding and increases confidence in the models. They have also cautioned that questionnaire data depends on honest self-reporting and that results from a single hospital may not transfer directly to the general population.

Ensemble learning and careful evaluation have received increasing attention in recent studies. Random forest, gradient boosting, and extreme gradient boosting have been reported to outperform single classifiers on both measurement-based and symptom-based datasets, largely because they capture interactions between features and are robust to noise. At the same time, researchers have emphasised that accuracy alone is an inadequate measure for medical screening, since a model that misses diabetic patients may still appear accurate when the negative class dominates. Recall, F1 score, and the area under the ROC curve are therefore preferred, and techniques such as class weighting and synthetic minority oversampling are applied to improve detection of positive cases. Interpretability methods such as feature importance and Shapley values have been used to make predictions understandable to clinicians and patients, and to confirm that the model relies on medically reasonable factors.

Recent work has also explored the integration of prediction models with sensing devices and mobile applications. Studies have described systems in which glucometer readings, blood pressure measurements, and activity data from wearable devices are

transmitted to a smartphone or web server where a trained model estimates risk. Other studies have focused on the prevention aspect, showing that structured lifestyle advice on diet, exercise, and weight management can significantly reduce the progression from prediabetes to diabetes. Authors have pointed out that the value of a prediction system increases greatly when it is linked to clear guidance and to follow-up measurements, because awareness alone rarely changes behaviour. The system proposed in this paper follows this direction by combining symptom and measurement based prediction with risk-specific prevention advice and a history of repeated readings for each user.

III. SYSTEM ANALYSIS

The early stage diabetes risk prediction system accepts information about a person through a short symptom questionnaire and a set of clinical and lifestyle fields. Symptoms include frequent urination, excessive thirst, sudden weight loss, weakness, increased hunger, blurred vision, itching, delayed healing, and irritability. Measurements include fasting glucose, body mass index, blood pressure, and waist circumference, which can be typed in or taken from a glucometer, a digital blood pressure monitor, and a weighing scale. Lifestyle fields cover physical activity, diet, sleep, and family history. The input is validated for physically impossible readings, missing values are imputed, categorical answers are encoded, and numerical values are scaled. A trained classifier produces a probability of diabetes, which is mapped to low, moderate, or high risk using thresholds chosen to keep recall high. The factors that contribute most to the result are identified, and the prevention module selects advice suited to those factors. Each assessment is stored so that the user and health workers can follow changes over time.

Existing System

At present, diabetes is usually detected in one of two ways. Some people are diagnosed during a hospital visit for an unrelated complaint, when a routine blood test shows high glucose. Others are identified during occasional health camps, where glucose is checked with a glucometer and a simple cut-off value is used to decide whether the person should see a doctor. Paper-based risk scores are sometimes used in public health programmes, but they are filled manually, are not adapted to the local population, and are rarely recorded in a form that allows follow-up. Symptoms such as frequent urination, fatigue, or slow healing are commonly ignored or attributed to other causes, because people are not aware of their connection with diabetes. Online symptom checkers exist, but most give generic warnings without explaining which factors matter or what the person should do next. Records from camps and clinics are kept separately and are seldom used to monitor how a person's risk changes over months, so opportunities for early lifestyle intervention are frequently missed.

Disadvantages of Existing System

- Many people are diagnosed only after complications of diabetes have already developed.
- Screening depends on occasional health camps and hospital visits.

- A single glucose cut-off ignores symptoms, lifestyle, and family history.
- Paper-based risk scores are not adapted to the local population.
- Early symptoms are often ignored because their significance is not understood.
- Online symptom checkers give generic warnings without specific guidance.
- Records from camps and clinics are not linked, so follow-up is difficult.
- Changes in an individual's risk over time are not monitored.
- Prevention advice, when given, is general rather than based on personal risk factors.
- Health workers lack a summary view of risk across a screened community.

Proposed System

The proposed system is a machine learning based screening and prevention platform for early stage diabetes. Users or health workers enter symptoms, basic measurements, and lifestyle details through a web form, and readings from a glucometer, blood pressure monitor, and weighing scale can be recorded directly. The data is validated, zero or impossible readings are treated as missing and imputed, categorical answers are encoded, and numerical fields are scaled. Logistic regression, k-nearest neighbours, a support vector machine, a decision tree, random forest, and gradient boosting are trained on combined symptom and measurement data, with class weighting to handle imbalance. The models are compared using accuracy, precision, recall, F1 score, and ROC-AUC under stratified cross-validation, and the model with the best recall and F1 balance is selected. The output is shown as low, moderate, or high risk with a probability and the main contributing factors, and the prevention module gives matching advice on diet, activity, weight, and confirmatory testing. A dashboard stores repeated assessments and summarises community risk.

Advantages of Proposed System

- Estimates diabetes risk from simple symptoms and low-cost measurements.
- Combines symptoms, clinical readings, lifestyle, and family history in one model.
- Model selection emphasises recall so that fewer diabetic cases are missed.
- Explains which factors contribute most to each person's risk.
- Provides prevention advice matched to the individual's specific risk factors.
- Accepts readings from common devices such as glucometers and BP monitors.
- Stores repeated assessments so that risk trends can be followed over time.
- Gives health workers a summary of risk across a screened group.
- Encourages high-risk users to seek laboratory confirmation early.
- Runs on ordinary computers and phones without specialised hardware.

IV. METHODOLOGY

The methodology starts with assembling a training dataset from a public symptom-based dataset and a measurement-based dataset, supplemented with records gathered during screening camps with consent. Each dataset is examined for duplicates, impossible values, and missing entries, and zero values in glucose, blood pressure, and body mass index are replaced by medians computed within each outcome class of

the training portion. Binary symptom answers are encoded as zero or one, gender and activity level are one-hot encoded, and continuous measurements are standardised. Exploratory analysis with correlation matrices and distribution plots identifies strong predictors and redundant features. The data is split into training and test sets with stratification, and imbalance is addressed with class weights or synthetic oversampling applied only within training folds. Six classifiers are trained and tuned using grid search with five-fold cross-validation. The chosen model is evaluated on the test set with accuracy, precision, recall, F1 score, ROC-AUC, and a confusion matrix, risk thresholds are fixed, and the model is deployed behind a web service.

SYSTEM ARCHITECTURE

The architecture of the diabetes risk prediction and prevention system is organised into layers that lead from raw user input to personalised guidance. The Data Acquisition Layer collects questionnaire answers and readings from glucometers, blood pressure monitors, and weighing scales through a web or mobile form. The Validation Layer checks that readings fall within physically possible ranges and flags incomplete entries. The Preprocessing Layer imputes missing values, encodes categorical answers, and scales numerical measurements. The Feature Engineering Layer derives body mass index categories, age groups, and a symptom count, and selects the most informative features. The Model Layer holds the trained classifiers and the selected production model. The Risk Scoring Layer converts predicted probabilities into low, moderate, or high risk using recall-oriented thresholds. The Explanation and Prevention Layer identifies the main contributing factors and selects matching advice. The Storage Layer keeps assessment history for each user, and the Dashboard Layer presents personal risk trends and community summaries to users and health workers.

Early stage diabetes risk prediction and prevention system using ML

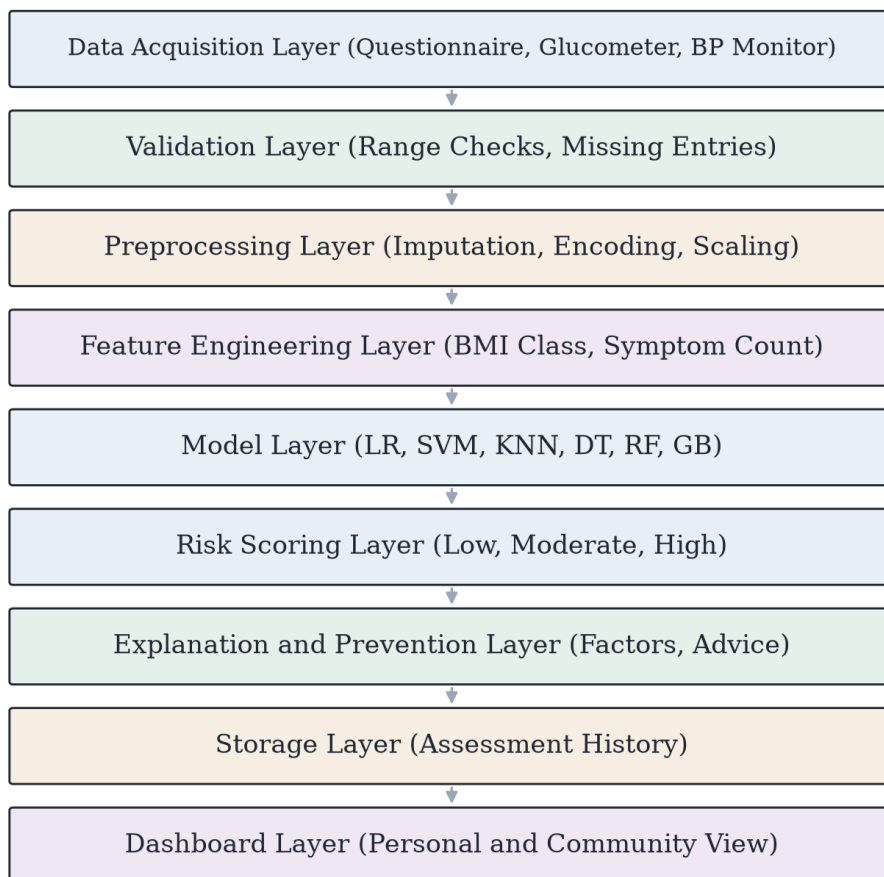


Fig. 1 System Architecture

V. RESULT AND OUTPUT



VI. CONCLUSION

This paper has presented an early stage diabetes risk prediction and prevention system that uses machine learning to estimate risk from symptoms, low-cost measurements, and lifestyle information. By combining these inputs in a single model, the system provides a more complete picture of risk than a glucose cut-off or a paper-based score, and it does so quickly enough to be used in screening camps, college health centres, and workplaces.

Several classifiers were compared, and ensemble methods such as random forest and gradient boosting gave the strongest performance, particularly in terms of recall and F1 score. Careful handling of missing clinical readings and class imbalance proved important, because these steps directly affect how many diabetic cases the model is able to detect. Emphasising recall in model selection reflects the fact that a missed case is far more costly than a follow-up test that turns out negative.

The explanation and prevention components turn the prediction into practical guidance. Users learn which of their own factors, such as weight, inactivity, or specific symptoms, are raising their risk, and receive advice that addresses those factors directly. The ability to record repeated measurements from common devices allows users and health workers to see whether lifestyle changes are having an effect over time. The system does not replace clinical diagnosis, and every high-risk result directs the user to confirmatory laboratory testing.

Future work can extend the system with direct Bluetooth connectivity to glucometers and wearable activity trackers, so that readings are captured automatically without manual entry. Support for regional languages and voice input would make the

questionnaire accessible to more people, and retraining with data collected from local screening camps would improve accuracy for the population served. Integration with primary health centre records could further support continuity of care for high-risk individuals.

REFERENCES

- [1] Kumar, R. D., Prudhviraaj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 through intensive investigation with supervised machine learning algorithm. In *Handbook of Artificial Intelligence and Wearables* (pp. 145-158). CRC Press.
- [2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In *The International Conference on Artificial Intelligence and Smart Environment* (pp. 557-564). Cham: Springer Nature Switzerland.
- [3] Sv satyakrishna, shirisha rangu, bhargavi nalacheruve. (2024) Prospective investigation on colorectal cancer with SMOTE on machine learning Algorithm
- [4] Dr.G.Vishnu Murthy, BhargaviNalacheruve 1Professor, Department of computer Science & engineering, Anurag University, TS, India. 2Student, Department of computer Science & engineering, Anurag University, TS, India.
- [5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, "Real-Time Object Detection in Drone Surveillance Using YOLOv5," in *Proc. 2025 3rd Int. Conf. IoT, Communication and Automation Technology (ICICAT)*, Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.
- [6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, "Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks," in *Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment*, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.
- [7] R. D. Kumar, V. N. S. Manaswini, "Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology," in *Blockchain for Smart Cities*, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.
- [8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, "An advanced movie recommender using collaborative filtering and sentiment analysis," *International Research Journal of Modernization in Engineering Technology and Science*, vol. 7, no. 7, July 2025, doi: 10.56726/IRJMETS81618.
- [9] Ravi Kumar Banoth, Ramana Murthy B V, "Automatic crop recommendation system using LightGBM and decision tree machine learning models," *Journal of Machine and Computing*, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.
- [10] Ravi Kumar Banoth, Dr. B.V. Ramana Murthy, "Smart agriculture through IoT and machine learning for analyzing carbon footprints," in *Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE)*, Apr. 2025.
- [11] Ravi Kumar Banoth, B. V. Ramana Murthy, "Soil image classification using transfer learning approach: MobileNetV2 with CNN," *SN Computer Science*, vol. 5, art. no. 199, 2024, doi: 10.1007/s42979-023-02500-x.