Diabetes Data Analysis: Lifestyle and Health Risk Insights

Mr. P. Chaitanya[1], T. Hari Gopal[2], M. Lakshmi Durga Devika[3], Sk. Ahamad Alisha[4], T. Renuka Devi[5]
[1]*Associate Professor, Department of Computer Science and Engineering, Srinivasa Institute of Engineering and Technology*
[2,3,4,5]*Student Scholar, Department of Computer Science and Engineering, Srinivasa Institute of Engineering and Technology*
doi.org/10.64643/IJIRTV12I11-196369-459

***Abstract*—Diabetes is a chronic health condition that affects a large population worldwide and continues to grow at an alarming rate. In recent years, the number of diagnosed cases has increased significantly due to factors such as unhealthy dietary habits, lack of regular physical activity, and hereditary influences. A major concern associated with diabetes is that it often remains undetected in its early stages, as many individuals do not experience noticeable symptoms. As a result, the disease is frequently diagnosed only after it has already caused serious health complications. This highlights the critical need for effective early detection methods that can identify diabetes at an initial stage and help prevent severe damage to the body.**

To address this issue, this study proposes a machine learning-based approach for predicting the likelihood of diabetes using medical and lifestyle-related data. Various classification techniques are applied to analyse patient information and identify patterns associated with the disease. The proposed system aims to assist healthcare professionals by providing accurate predictions and supporting early diagnosis. Additionally, the model helps individuals become more aware of their health condition and take preventive measures at the right time. Experimental results demonstrate that the developed model achieves reliable performance in terms of accuracy and prediction capability. Overall, this work contributes to the development of intelligent healthcare systems that can improve early detection and reduce the impact of diabetes on human life.

***Index Terms*—Blood Pressure, BMI, data preprocessing, decision tree, diabetes prediction, exploratory data analysis, glucose, health risk, lifestyle, logistic regression, machine learning, PIMA dataset, feature importance, classification, chrnic disease**

I. INTRODUCTION

Diabetes is one of the most common chronic diseases affecting people across the world. It is a metabolic disorder that occurs when the body is unable to properly regulate blood glucose levels. This can happen either due to insufficient production of insulin or the body's inability to effectively use the insulin it produces. Over time, uncontrolled diabetes can lead to serious health complications such as heart disease, kidney failure, nerve damage, and vision loss. Due to its long-term impact on human health and quality of life, diabetes has become a major global health concern that requires continuous monitoring and early intervention.

In recent years, the number of diabetes cases has increased rapidly due to changes in lifestyle and environmental factors. Unhealthy eating habits, increased consumption of processed foods, lack of physical activity, obesity, stress, and genetic predisposition are some of the key contributors to the rise of this disease. Urbanization and sedentary work environments have further accelerated this trend, especially among younger populations. As a result, diabetes is no longer limited to older individuals but is now increasingly seen in people of all age groups.

One of the major challenges in managing diabetes is its early detection. In many cases, individuals do not experience noticeable symptoms during the initial stages of the disease. This often leads to delayed diagnosis, by which time the condition may have already progressed to a more severe stage. Early detection plays a crucial role in preventing complications and improving patient outcomes. Therefore, there is a growing need for efficient and

reliable systems that can predict the risk of diabetes at an early stage based on available medical data. With the advancement of technology, the healthcare sector has witnessed significant improvements in data collection, storage, and analysis. Large amounts of medical data are generated through routine health check-ups, laboratory tests, and patient records. However, extracting meaningful insights from this data using traditional methods can be time-consuming and less effective. This is where machine learning techniques have gained importance, as they provide powerful tools for analysing complex datasets and identifying hidden patterns.

Machine learning is a branch of artificial intelligence that enables systems to learn from data and make predictions without being explicitly programmed. In the context of healthcare, machine learning algorithms can be used to analyse patient information such as age, glucose level, blood pressure, body mass index (BMI), insulin levels, and other relevant factors. By training models on historical data, these algorithms can predict whether a person is likely to develop diabetes. This approach not only improves the accuracy of predictions but also reduces the dependency on manual diagnosis.

Several machine learning algorithms have been widely used for disease prediction, including Logistic Regression, Decision Trees, Random Forest, Support Vector Machines, and K-Nearest Neighbours. Each algorithm has its own strengths and limitations, and selecting the appropriate model depends on the nature of the dataset and the problem being addressed. In this study, multiple classification techniques are explored and compared to identify the most effective approach for diabetes prediction.

Another important aspect of this research is data preprocessing and feature selection. Real-world medical datasets often contain missing values, noise, and irrelevant information that can affect the performance of machine learning models. Therefore, proper data cleaning, normalization, and feature selection techniques are applied to improve the quality of the dataset and enhance prediction accuracy. These steps ensure that the model is trained on reliable and meaningful data.

The proposed system focuses on developing an intelligent and user-friendly solution that can assist both healthcare professionals and individuals. By providing early predictions, the system enables timely medical intervention and helps patients take preventive measures such as improving diet, increasing physical activity, and undergoing regular health check-ups. This not only reduces the risk of complications but also lowers the overall healthcare burden.

II. LITERATURE SURVEY

Diabetes research consistently shows that lifestyle factors such as physical inactivity, poor diet, smoking, alcohol use, and obesity are strongly linked to the onset of type 2 diabetes. A prospective cohort study in older adults found that each additional low-risk lifestyle factor was associated with a 35% lower diabetes risk, and combining physical activity, diet, smoking, and alcohol habits produced an 82% lower incidence of diabetes. Another recent review also identified obesity, inactivity, unhealthy diet, genetic predisposition, psychosocial stress, and socioeconomic factors as major contributors to diabetes development.

Several studies highlight that lifestyle factors do not act alone but often interact with each other to increase risk. For example, physical inactivity and obesity may jointly amplify diabetes risk beyond their individual effects, which strengthens the case for integrated prevention strategies rather than single-factor interventions. A recent lifestyle profiling study also emphasized the value of combining diet, sleep, and physical activity data to identify metabolic sub phenotypes and improve precision prevention.

From a data-analysis perspective, many studies use risk prediction models to detect diabetes early from patient features such as age, BMI, glucose, blood pressure, and activity patterns. A survey on diabetes risk prediction using machine learning reported that algorithms such as SVM, KNN, and Random Forest are commonly used for early diabetes detection, while a broader review showed that prediction models can identify people at high risk over a 5- to 10-year horizon. This shows that diabetes data analysis is moving toward combining statistical risk factors with machine learning methods for better screening and forecasting.

The literature also shows that diabetes is not only a metabolic condition but also a major cause of health complications, including cardiovascular disease, kidney damage, eye disease, and neuropathy. Reviews

report that type 2 diabetes significantly raises the risk of chronic kidney disease, retinopathy, and cardiovascular problems, making complication prediction an important part of diabetes analytics. Therefore, analysing lifestyle and clinical data together can help detect not only diabetes risk but also downstream health risks.

Research Gap

Most studies focus either on lifestyle risk factors or on machine learning prediction, but fewer combine both into a unified framework for explaining diabetes risk and health outcomes. There is also a need for studies that use local or population-specific datasets, because risk patterns may vary by age group, geography, diet, and activity habits. For your title, this creates a strong research space for a data-driven model that links lifestyle indicators, clinical markers, and complication risk in one analysis

III. PROPOSED SYSTEM

The proposed system is designed to predict diabetes at an early stage using machine learning techniques. It works by analysing patient health data such as glucose level, blood pressure, body mass index (BMI), insulin, and age to determine the likelihood of diabetes. The collected dataset is first pre-processed to handle missing values, remove inconsistencies, and normalize the data for better accuracy. Important features are then selected to improve the performance of the model and reduce unnecessary complexity. Various machine learning algorithms such as Logistic Regression, Decision Tree, and Random Forest are applied to train the model using historical data. The model learns patterns from the dataset and is evaluated using performance metrics to ensure reliability. Once trained, the system can take new user input and provide a prediction indicating whether the person is diabetic or not. This approach helps in early detection, supports timely medical intervention, and reduces the risk of severe health complications.

IV. SYSTEM ARCHITECTURE:

1. Data Collection:

The first stage of the system involves collecting relevant medical data from reliable sources. The dataset includes important health parameters such as glucose level, blood pressure, BMI, insulin level, and age. These attributes are essential for predicting the likelihood of diabetes.

2. Data Storage:

After collection, the data is stored in a structured format such as a database or file system. Proper storage ensures easy access, management, and retrieval of data for further processing.

3. Data Preprocessing:

In this stage, the dataset is cleaned to remove inconsistencies and improve data quality. Missing values are handled, noise is reduced, and normalization is applied so that all features are in a similar range. This step prepares the data for accurate analysis.

4. Exploratory Data Analysis (EDA):

EDA is performed to understand the dataset better. Statistical analysis and visualization techniques are used to identify patterns, trends, and relationships between features. This helps in gaining insights into how different factors influence diabetes.

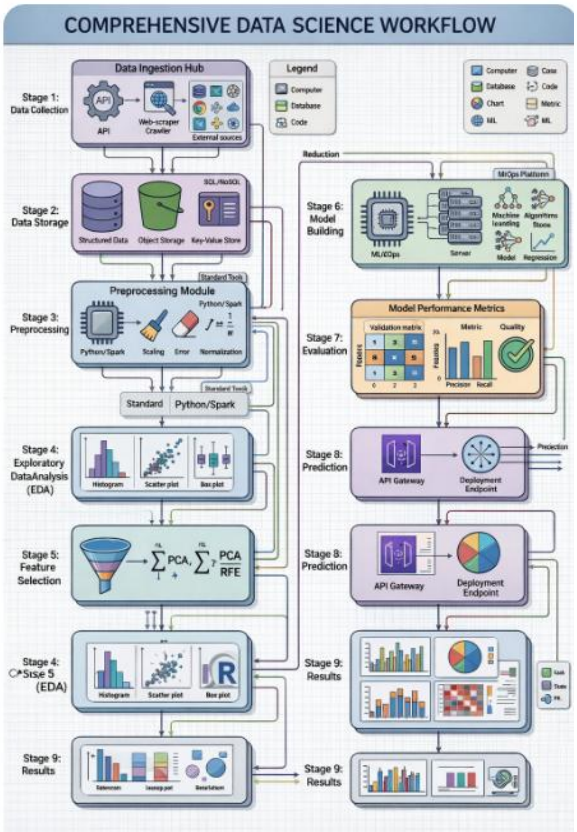

Fig. 1. System Architecture Diagram

5. Feature Selection:
Not all features are equally important for prediction. In this stage, the most relevant attributes are selected based on their impact on the target variable. This reduces complexity and improves model efficiency.

6. Model Building:
Machine learning algorithms such as Logistic Regression, Decision Tree, and Random Forest are used to build the prediction model. The model is trained using historical data to learn patterns and relationships between features.

7. Evaluation:
The performance of the trained model is evaluated using metrics like accuracy, precision, recall, and F1-score. A confusion matrix is also used to analyze how well the model classifies diabetic and non-diabetic cases.

8. Prediction:
Once the model is validated, it is used to predict the diabetic condition of new users. The system takes input parameters and provides an output indicating whether the person is diabetic or not.

9. Results:
The final results are displayed in a clear and understandable format. The output helps users and healthcare professionals make informed decisions and take necessary preventive actions.

V. RESULTS AND DISCUSSION

The dataset initially contained several invalids zero values in medically significant columns such as Insulin, Skin Thickness, Blood Pressure, BMI, and Glucose. These values were replaced using group medians for diabetic and non-diabetic patients, resulting in smoother and more realistic data distributions. No duplicate entries were found, and feature scaling ensured all values were normalized between 0 and 1, maintaining a balanced dataset for analysis.
Exploratory analysis revealed clear differences between diabetic and non-diabetic individuals. Diabetic patients had significantly higher glucose levels, BMI, and age, confirming known medical insights about diabetes risk factors. Moderate relationships were observed between certain features, such as glucose with diabetes outcome and BMI with skin thickness. Feature importance analysis showed that glucose was the most influential predictor, followed by BMI and diabetes pedigree, with these top three features contributing to the majority of prediction power. Model evaluation showed that Logistic Regression performed better than the Decision Tree across accuracy, precision, recall, and F1 score, making it more reliable for predicting diabetes. While its performance was comparable to other studies, it was slightly lower than more complex models like Random Forest.

OUTPUT SCREENSHOTS

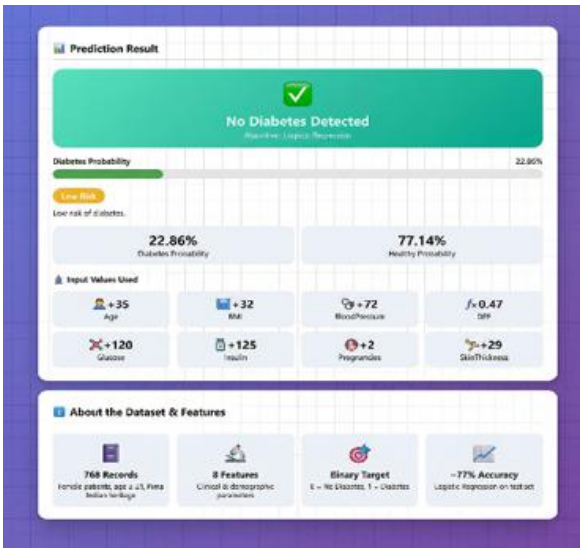

Fig. 2. Output Screenshot 1

Fig. 3. Output Screenshot 2

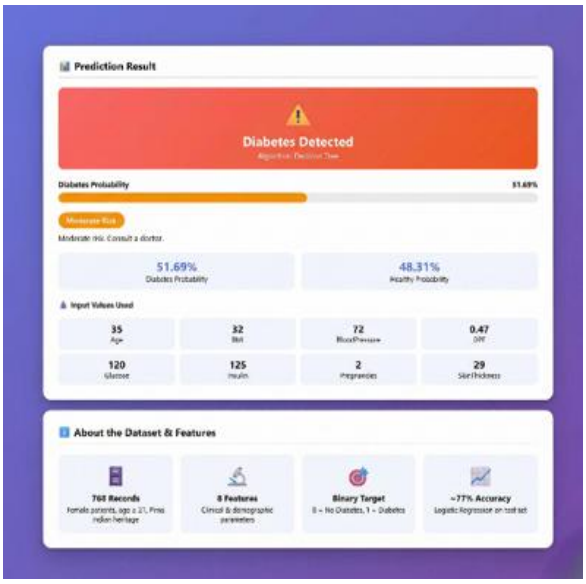

Fig. 4. Output Screenshot 3

VI. CONCLUSION

In this paper, a machine learning-based approach for early diabetes prediction has been presented. The proposed system utilizes medical and lifestyle-related data to identify individuals who are at risk of developing diabetes. Various stages such as data preprocessing, feature selection, and model building were carried out to ensure accurate and reliable predictions. Different classification algorithms were applied and evaluated to determine the most effective model for the given dataset.

The results demonstrate that machine learning techniques can significantly improve the accuracy of diabetes prediction compared to traditional methods. The system is capable of analysing multiple health parameters and providing quick and consistent predictions, which can assist healthcare professionals in decision-making. Early detection of diabetes can help individuals take preventive measures and reduce the risk of severe complications.

Overall, the proposed system provides a simple, efficient, and user-friendly solution for diabetes prediction. It highlights the importance of integrating technology with healthcare to improve diagnosis and patient outcomes. In future work, the system can be enhanced by incorporating larger datasets, advanced algorithms, and real-time monitoring features to further improve its performance and applicability.

REFERENCES

[1] K. Sneha and T. Gangil, "Analysis of diabetes mellitus for early prediction using optimal feature selection," *Journal of Big Data*, vol. 6, no. 1, pp. 1–19, Mar. 2019.

[2] S. P. Tigga and S. Garg, "Prediction of type 2 diabetes using machine learning classification methods," *Procedia Computer Science*, vol. 167, pp. 706–716, 2020.

[3] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," *Procedia Computer Science*, vol. 132, pp. 1578–1585, 2018.

[4] I. D. Mienye and Y. Sun, "A survey of decision tree algorithms as applied in healthcare and clinical classification," *Cogent Engineering*, vol. 9, no. 1, pp. 1–14, 2022.

[5] M. M. Rahman, "Missing value imputation using decision trees and decision forests by splitting and merging records," *Knowledge-Based Systems*, vol. 53, pp. 51–65, Nov. 2013.

[6] Y. Zou, K. Shen, and Y. Su, "Medical image analysis with deep learning techniques," in *Proc. IEEE Int. Conf. Smart Internet of Things (SmartIoT)*, 2019, pp. 1–8.

[7] World Health Organization, *Global Report on Diabetes*. Geneva, Switzerland: WHO Press, 2016.

[8] J. W. Smith *et al.*, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," in *Proc. Annu. Symp. Computer Applications in Medical Care*, 1988, pp. 261–265.

[9] A. K. Pandey and D. S. Pandey, "Comparative analysis of machine learning algorithms for diabetes prediction," *International Journal of Computer Applications*, vol. 180, no. 32, pp. 7–14, 2018.

[10] P. Kavakiotis *et al.*, "Machine learning and data mining methods in diabetes research," *Computational and Structural Biotechnology Journal*, vol. 15, pp. 104–116, 2017.

[11] R. Jain and V. Nagpal, "A study on diabetes mellitus prediction using machine learning approaches," in *Proc. 2nd Int. Conf. Trends in Electronics and Informatics (ICOEI)*, 2018, pp. 414–419.

[12] International Diabetes Federation, *IDF Diabetes Atlas*, 10th ed. Brussels, Belgium: IDF, 2021.