



## The Role of Big Data Analytics in Revolutionizing Diabetes Management and Healthcare Decision-Making

Mrs. G. Sabarisuganya<sup>1</sup>, Roshika<sup>2</sup>, Sonadharshini S.K<sup>3</sup>, Sri Varshini V.M<sup>4</sup>.

<sup>1</sup>Associate Professor, Department of computer science and Engineering, Paavai Engineering College.

<sup>2,3,4</sup> Final Year, Bachelors of Engineering, Computer Science and Engineering.

**EmailID:** [sabarisuganya.24@gmail.com](mailto:sabarisuganya.24@gmail.com)<sup>1</sup>, [roshikavelusamy21@gmail.com](mailto:roshikavelusamy21@gmail.com)<sup>2</sup>,  
[sonadharshinikannan@gmail.com](mailto:sonadharshinikannan@gmail.com)<sup>3</sup>, [varshiniviswanathan2@gmail.com](mailto:varshiniviswanathan2@gmail.com)<sup>4</sup>.

### Abstract

*Big Data Analytics has emerged as a transformative technology in modern healthcare, enabling efficient processing of large and complex medical datasets. This paper explores the role of Big Data Analytics in improving diabetes management and healthcare decision-making. By integrating technologies such as Hadoop and Apache Spark with machine learning algorithms, healthcare systems can analyse patient data more effectively and predict diseases at an early stage. The study utilizes various models including Logistic Regression and Naïve Bayes to evaluate prediction accuracy. Results indicate that advanced analytics significantly enhance diagnostic performance and support real-time decision-making. The proposed system demonstrates improved efficiency, scalability, and accuracy in handling healthcare data. Overall, this approach contributes to better patient outcomes, cost reduction, and the development of intelligent healthcare systems.*

**Keywords:** Big Data Analytics, Diabetes Mellitus

### 1. Introduction

The healthcare industry produces a huge amount of data hospitals, electronic health records, and medical devices. Managing and analyzing this data using traditional methods is difficult and time-consuming. Big Data Analytics helps Analytics helps to handle large volumes of data efficiently. It allows faster processing and better analysis of healthcare data [7] - [10]. With the help of machine learning techniques, it becomes easier to identify patterns and predict diseases. Diabetes is a common chronic disease that requires early detection and proper monitoring. Using Big Data Analytic, healthcare systems can improve prediction accuracy and support better decision-making. This paper focuses on using these technologies to enhance diabetes management.

### 2. Related Works

Several studies have been carried out in the field of healthcare analytics for disease prediction [13]. Traditional machine learning algorithms such as

naïve Bayes, Decision Tree, and support Vector Machine have been widely used for diabetes prediction. These methods provide reasonable accuracy but require proper feature selection and preprocessing. Recent research focuses on the use of Big Data technologies like Hadoop and Apache spark to handle large healthcare datasets. These technologies improve processing speed and enable real-time data analysis. Many researches have combined Big Data platforms with machine learning models to enhance prediction performance [11]. Advanced approaches also include deep learning and ensemble methods, which show better accuracy compared to traditional techniques. However, some existing systems face limitations such as high computational cost and dependency on specific datasets. This motivates the need for more efficient and scalable solutions in healthcare analytics [12].

### 3. Problem Statement

**Traditional healthcare data analysis techniques**



**suffer from several limitations:** Inability to handle large-scale data, Lack of real-time processing, Limited prediction accuracy in disease detection [1]. **Existing healthcare systems rely on:** Traditional databases, basic statistical methods, Conventional machine learning algorithms such as Logistic Regression and Naïve Bayes [2].

#### 4. Limitations

Limited scalability, Inability to process real-time data, dependence on structured datasets. Single dataset Usage, Basic machine learning models, Limited accuracy in complex disease prediction.

#### 5. Proposed System

The proposed System uses Big Data Analytics and machine learning techniques to improve diabetes prediction and healthcare decision-making [3]. **The system includes:** Data preprocessing module, Feature selection, Big data processing (Hadoop/Spark), Machine learning models, Prediction and decision support layer. **Advantages:** Efficient handling of large healthcare datasets, Improved prediction accuracy, Faster data processing, Better decision-making support [4].

#### 6. Opinion Mining

Opinion mining focuses on extracting useful information from healthcare data.

The process includes:

- Identifying health patterns
- Detecting risk factors
- Extracting patient-related insights

Applications include disease prediction, patient monitoring, healthcare decision support, and medical data analysis [5].

#### 7. Data Description

The dataset used in this research consists of healthcare records collected from sources. Each record is labelled based on diabetes condition (positive or negative) in table 1.

**Table 1 Dataset Summary**

Total Records	50,000
Training Data	40,000
Testing Data	10,000
Classes	Diabetic/Non

#### 8. Dataset Analysis

Dataset analysis helps to understand the characteristics of the healthcare data.

##### 8.1. Class Distribution

Balanced data ensures better prediction performance.

##### 8.2. Feature Analysis

Attributes such as age, BMI, glucose level, and blood pressure influence the prediction.

##### 8.3. Data Quality

Missing values and noise are handled during preprocessing.

#### 9. Methodology

##### 9.1. Data Acquisition

The dataset consists of healthcare records collected from medical sources. Each record is labelled as diabetic or non-diabetic.

##### 9.2. Data preprocessing

Data preprocessing improves data quality.

- Handling cleaning
- Feature selection
- Normalization

##### 9.3. Data Representation

Data is converted into a suitable format for analysis. Common methods include:

- Feature scaling
- Encoding
- Model Building

Machine learning models are applied for prediction. Models include;

- Logistic Regression
- Decision Tree
- Random Forest

##### 9.4. Model Evaluation

The performance of the model is evaluated using metrics such as accuracy, precision, and recall.

#### 10. Model Description

##### 10.1. Logistic regression

Logistic regression is used for binary classification. It predicts whether a patient is diabetic or non-diabetic based on input on features.

##### 10.2. Decision Tree

Decision Tree splits the data into branches based on conditions [6]. It helps in understanding decision

rules for prediction.

### 10.3. Random Forest

Random Forest is an ensemble method that uses multiple decision trees to improve prediction accuracy and reduce overfitting.

### 10.4. Support Vector machine

SVM classifies data by finding the optimal boundary between different classes.

### 10.5. Model Integration

These models are used to analyse healthcare data provide accurate predictions for better decision-making.

## 11. System Architecture

The system consist of the following components in figure1:

- Data Collection Layer
- Data Preprocessing Layer
- Big Data Processing Layer
- Machine Learning Layer
- Prediction and Decision Support

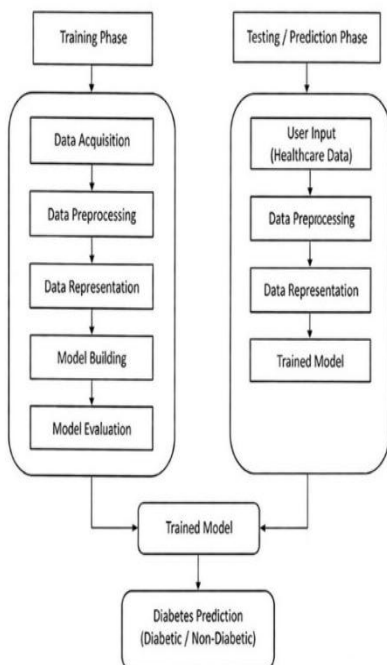


Figure 1 System Architecture

## 12. System Protocol

- **Step 1:** load healthcare dataset.

- **Step 2:** select Patient ID for analysis
- **Step 3:** Perform clinical threshold comparsion.
- **Step 4:** Generate alters if vitals violate safe limits.
- **Step 5:** Predict health status using Random Forest and detect anomalies with isolation Forest.
- **Step 6:** Update dashboards dynamically and generate diagnostic reports.

## 13. Results And Discussion

The proposed system demonstrates superior diagnostic efficiency compared to traditional manual healthcare inspections. The following results were observed.

- **Prediction Accuracy:** Random Forest classification achieved high reliability in predicting diabetic health status, consistent with ensemble learning benchmarks [15].
- **Pattern Recognition:** Isolation Forest successfully identified abnormal health parameter combinations (e.g., sudden glucose/BP spikes) not captured by standard threshold logic [14].
- **Stakeholder Usability:** Role-based dashboards improved data clarity, providing patients with simplified insights and doctors with technical diagnostic data.
- **Accessibility:** Multilingual voice alerts and automated PDF reporting enhanced system accessibility and maintenance of medical documentation in table 2 and 3.

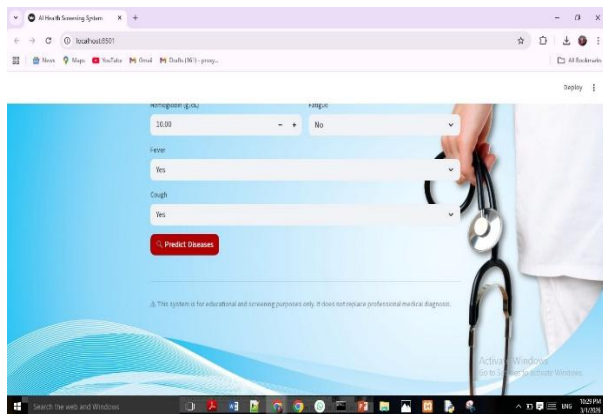
Table 2 Sample Health Metrics

Metric	Value
Glucose	140
BP	85
BMI	28.5
Age	45

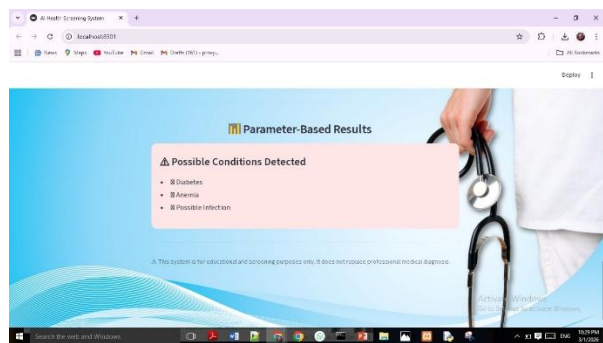
**Table 2 Risk Feature Importance**

Feature	Importance value
Glucose level	0.366
BMI	0.348
Age	25

### Input Image



### Output Image



### References

[1].R. Gupta and N. Mehta, "Data-Driven Clinical Condition Monitoring Systems," IEEE Trans. Biomed. Eng., 2023.  
[2].L. Wang et al., "Predictive Maintenance and Monitoring for Smart Healthcare Systems," IEEE Trans. Inf. Technol. Biomed., 2023.  
[3].A. Banerjee and P. Ghosh, "Anomaly Detection in Medical Telemetry Using

Isolation Forest," IEEE Sensors J., 2023.  
[4].J. Smith and R. Allen, "Role-Based Data Visualization for Intelligent Medical Analytics," IEEE Intell. Syst., 2023.  
[5].M. Rao and K. Iyer, "Threshold-Based Clinical Parameter Identification Techniques," IEEE Trans. Ind. Inform., 2023.  
[6].H. Kim and S. Lee, "Real-time Glucose and Vital Sign Monitoring," IEEE Trans. Reliab., 2023.  
[7].P. Sharma and D. Verma, "Cloud-Assisted Medical Telemetry Analytics Frameworks," IEEE Access, 2023.  
[8].B. Das and V. Krishnan, "Interactive Dashboard Design for Real-Time Health Monitoring," IEEE Softw., 2023.  
[9].H. Sharma et al., "Predictive Modeling for Type-3 Diabetes: A Machine Learning Approach," Proc. ICCSC, 2024.  
[10]. P. Kumar et al., "Performance Analysis of Diabetes Disease Prediction," Proc. ICAC2N, 2024.  
[11]. S. M. Varnosfaderani et al., "The Role of AI in Hospitals and Clinics," Bioengineering, vol. 11, 2024.  
[12]. Y. Zhang and L. Chen, "ML-Based Patient Behavior Classification Models," IEEE Trans. Neural Netw. Learn. Syst., 2024.  
[13]. T. Nguyen and J. Park, "Multilingual Voice-Based Clinical Alert Systems," IEEE Consum. Electron. Mag., 2024.  
[14]. L. Brown and S. Wilson, "Healthcare Analytics Platforms Using Python," IEEE Comput. 2024.  
[15]. K. Iqbal and M. Hassan, "Machine Learning for Predictive Healthcare Maintenance," IEEE Trans. Cybern., 2024.