EAI Endorsed Transactions
on Pervasive Health and Technology
Research Article EAI.EU

# Integrated Embedded system for detecting diabetes mellitus using various machine learning techniques

Rishita Konda[1*], Anuraag Ramineni[2], Jayashree J[3], Niharika Singavajhala[4] and Sai Akshaj Vanka[5]

[1,2,3]School of Computer Science and Engineering (SCOPE), VIT University, Katpadi, 632014, Tamil Nadu, India
[4]Mechanical Engineering, Vasavi College of Engineering, Hyderabad, 500031, Telangana, India
[5]Information Technology, Vasavi College of Engineering, Hyderabad, 500031, Telangana, India

## Abstract

INTRODUCTION: The goal of this study, titled "Integrated System for Detecting Diabetes Mellitus using Various Machine Learning and Deep Learning Algorithms," is to increase the precision and usability of diabetes diagnosis through the investigation and application of a wide range of machine learning and deep learning techniques.
OBJECTIVES: The objective of the study was to establish a comprehensive system for identifying diabetes mellitus by combining several machine learning and deep learning methods
METHODS: The methodology included every phase, from data gathering and preprocessing through advanced model development and performance assessment. The experiment demonstrated how combining several machine learning and deep learning techniques might completely transform diabetes detection. While praising accomplishments, the methodology also highlighted flaws in the data collection process. The goal of the roadmap for future improvements was to use technology to better detect and treat diabetes, which would ultimately help people of all ages and backgrounds.
RESULTS: The project's remarkable results demonstrate the legitimacy of the methodology chosen while also highlighting its potential to completely transform the diagnosis and treatment of diabetes
CONCLUSION: The conclusion of this project lays the ground for next developments, such as improved user interfaces and the expansion of dataset scope. Through these initiatives, the long-term objective of providing more precise and accessible diabetes diagnoses becomes a real possibility, providing significant advantages to people from a variety of age groups and demographics[6].

**Keywords:** Mellitus, Embedded Technique, Machine Learning, SGN Algorithm

Received on 22 December 2023, accepted on 15 March 2024, published on 21 March 2024



doi: 10.4108/eetpht.10.5497

*Corresponding author. Email: rishitakonda07@gmail.com

## 1. Introduction

An innovative initiative focused on developing an integrated system for identifying diabetes mellitus has been developed as a result of the pursuit of better healthcare diagnostics and management. This dataset is composed up of a wide range of features [1], including age, blood pressure readings, glucose levels, gender distinctions, and body mass index (BMI). "Outcome," the dataset's cardinal variable, divides people into two distinct classes.

The project is carried out through a succession of thoroughly planned steps: Data Acquisition and Preprocessing: The journey begins with the dataset's assimilation through the upload of a CSV file. The dataset then goes through an important preparation stage. The dataset's complexities become apparent at this point; missing data and superfluous columns demand attention. Utilizing cutting-edge preprocessing methods like imputation for missing values and the culling of unnecessary columns is

necessary to address these difficulties. Additionally, methods like transforming discrete variables into a format appropriate for analysis are used. Segmenting the Dataset and the Initial Analysis: The dataset segmentation choice becomes a crucial one. Training and testing subsets are successfully separated from the overall dataset. The foundation for the future analysis is laid by this partition, which is formed using a 70:30 ratio. An initial set of insights is obtained by using a 5-fold cross-validation approach [4]. The identification and application of appropriate procedures is at the heart of this project. Technique selection and model creation are key components. One of these, the SGN (Stochastic Gradient Descent + Neural Networks) model, stands out among the others. This innovative model is created by stacking the strengths of stochastic gradient descent with neural networks. A thorough review and synthesis of the available studies points the project in the direction of SGN as the ideal model. Hyperparameters are tuned for each method during model refinement.

This method acknowledges how complex data properties, feature engineering [5], and hyperparameter setups interact to effect model performance. This leads to customized settings for Stochastic Gradient Descent, Neural Networks, and Gradient Boosting. This study's significance in the constantly changing world of healthcare technology rests in its potential to transform diabetes diagnosis and management. The paper's ability to combine several machine learning and deep learning methodologies, enhancing the predictive potential of diabetes detection, is best demonstrated by the chosen model, SGN. This initiative acknowledges its successes but also emphasizes the need to solve problems like data inconsistencies, providing a roadmap for future advancements. The integrated system developed through this research, in conclusion, not only serves as an example of innovation, but also reveals the future direction of redefining diabetes diagnosis and care.

## 2. Related Work

The field of science has extensively studied the pursuit of utilizing deep learning and machine learning approaches to overcome issues in the identification of diabetes mellitus. This section provides an overview of some studies that are closely related to the goals and methodology of the current project. Several researchers have delved into this field, yielding insights that complement and fit with the goals of the current study. A systematic assessment of the efficiency of bedside examinations to identify peripheral arterial disease in individuals with diabetes mellitus was carried out by Chuter et al. in 2023[1]. They discovered that there isn't much proof to back up using a single in-person exam to diagnose peripheral artery disease in persons with diabetes mellitus. Chang et al. (2023) classified diabetes mellitus in Pima Indians using machine learning methods [2]. They discovered that the support vector machine (SVM) was the most effective approach. The management of gestational diabetes mellitus that was discovered early in pregnancy. [3] was evaluated by Simmons et al. in 2023. They discovered that a program of food and exercise is the most successful course of action. Joshua and co-authors created a mobile application for the management of type 2 diabetes mellitus in Joshua et al. (2023) [4]. The future scope of the project, which envisions improving the application's accuracy and user interface, is paralleled by this study. Vesa and Bungau [5] investigated new compounds in cardiovascular disease, dyslipidemia, and diabetes mellitus. Even though they are concentrating on molecular variables, the idea of feature engineering in machine learning is aligned with their investigation of novel elements. The current project's focus on feature selection and preprocessing methods is consistent with the goal of finding pertinent features for diabetes prediction. Kee et al. (2023) used machine learning to carry out a systematic evaluation of cardiovascular consequences in a diabetes prediction model [6].

They discovered that cardiovascular problems in diabetics can be predicted using machine learning models. Machine learning techniques were employed by Chou et al. in 2023 to forecast the onset of diabetes [7]. They discovered that the random forest method had the best performance. Hematological variables were employed by Mansoori et al. (2023) to predict type 2 diabetes mellitus using machine learning techniques. They discovered that logistic regression was the most effective algorithm [8]. Deep learning embedders and machine learning methods were utilized by Challagundla et al. (2023) to screen for illnesses [9]. They discovered that the neural network was the algorithm that performed the best. Deep learning methods were employed by Pal et al. (2022) to predict and identify diabetes mellitus [10]. They discovered that CNN was the algorithm that performed the best. Based on lifestyle factors, Ganie and Malik (2022) employed an ensemble machine learning strategy to predict type-II diabetes mellitus. They discovered that the random forest method [11] had the best performance.

A systematic evaluation of artificial intelligence and machine learning-based diabetic mellitus detection and self-management was carried out by Chaki et al. in 2022 [12]. They discovered that artificial intelligence and machine learning can be utilized to enhance diabetes mellitus detection, diagnosis, and management [14]. A variety of diabetes prediction models were created by Shin et al. in 2022 utilizing machine learning techniques [13]. They discovered that the SVM was the most effective algorithm. Based on the gut microbiome profile, Ge et al. (2022) employed machine learning algorithms to predict type 2 diabetes mellitus [15]. Alghamdi et al. (2017) predicted diabetes mellitus using SMOTE and an ensemble machine learning [16] technique. They discovered that the random forest method had the best performance. Along with more research, the ones mentioned over add to an increasing amount of literature

that reflects on the goals and procedures of this study. The study's scientific foundation is strengthened by the insights and results from this research, which also supports the importance of cutting-edge technology for diabetes detection and management.

## 3. Methodology

The initiative developed an effective and precise methodology for identifying diabetes using a methodical approach. The project team trained and tested various machine learning and deep learning algorithms using a dataset of more than 16,000 individuals. The objective was to develop an integrated strategy that combines the advantages of several approaches.

- Dataset collection and preparation:
  The dataset in CSV format was imported to start the project. Preprocessing was crucial in order to guarantee data relevancy and quality. Missing data and unnecessary columns were dealt with in this stage. The treatment of sparse attributes and the removal of unnecessary elements. Imputation techniques were used to deal with missing data by substituting the average or most frequent values. Discrete variables were also converted into a format that was appropriate for analysis.
- Initial analysis and dataset segmentation:
  The dataset was split into training and testing subsets in a 70:30 ratio. A 5-fold cross-validation strategy was used to ensure robust analysis and make it easier to assess the performance of the models.
- Algorithm Selection and Model Development:
  The methodology's foundation was the choice of the best techniques. Due to its ability to combine the strengths of stochastic gradient descent with neural networks, a unique model called SGN (Stochastic Gradient Descent + Neural Networks) became the front-runner. The SGN model was found to be the best one after a thorough analysis and review of the literature. Gradient Boosting, Naive Bayes, and Neural Networks were further methods.
- Performance Metrics for Model Performance Evaluation and Hyperparameter Tuning:
  An outstanding set of assessment metrics, including AUC, CA, F1, Precision, and Recall, were evident in the SGN model. For each strategy, hyperparameters were painstakingly tuned to increase model effectiveness.
- Model Refinement and Future Prospects:
  Hyperparameter tuning was undertaken to take data properties, feature engineering, and hyperparameter configurations into consideration, which have an impact on model performance. Gradient Boosting, Neural Networks, and Stochastic Gradient Descent all received the modified parameters.
- Model Comparison and Data Visualization:
  The SGN model was shown to be the most effective choice after thorough investigation, outperforming other approaches in terms of performance. To make it easier to understand the project's results graphically, data visualization techniques such as scatter plots were applied to the results.

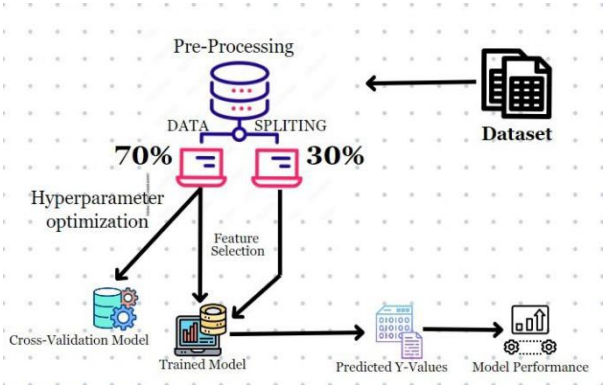


**Figure 1.** This figure illustrates sequential flow of the execution

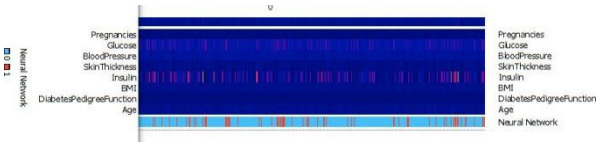


**Figure 2.** A heatmap visualization is employed in neural networks


**Figure 3.** Scattered plot illustrates relationship between glucose level and blood pressure

- Future Scope and Summary:
  The project's goals went beyond its immediate successes. Future efforts intended to improve the application's precision and usability, guaranteeing access to users of all ages. Initiatives for data collecting also aimed to broaden the dataset's breadth and improve the forecasting power of the program.

The methodology included every phase, from data gathering and preprocessing through advanced model development and performance assessment. The experiment demonstrated how combining several machine learning and deep learning techniques might completely transform diabetes detection. While praising accomplishments, the methodology also highlighted flaws in the data collection process. The goal of the roadmap for future improvements was to use technology to better detect and treat diabetes, which would ultimately help people of all ages and backgrounds.

## 4. Results

The study emphasizes the potential impact on healthcare management as well as the effectiveness of the suggested methodology in detecting diabetes.

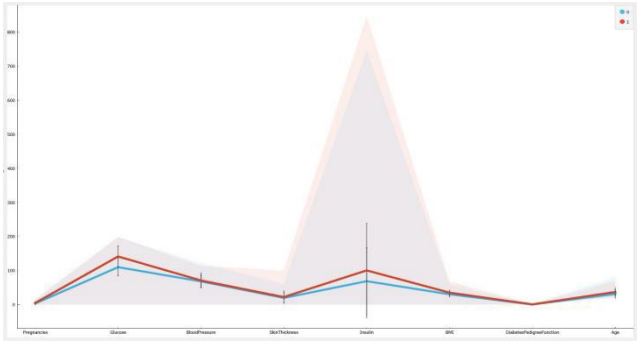

**Figure 4.** The line plot showcases a comprehensive comparison of attributes between Non-Diabetic (blue) and Diabetic (red) outcomes, effectively visualizing ranges, means, and error bars for each attribute.

Evaluation Metrics: After putting the various machine learning and deep learning algorithms into practice, a thorough review was carried out utilizing a number of metrics. These metrics, which offered numerical evaluations of the models' performance, comprised, it was crucial to evaluate the models' classification abilities using the metrics Area Under the Curve (AUC), Classification Accuracy (CA), F1-score, Precision, and Recall.

Performance Rankings: The integrated SGN model, which combines Stochastic Gradient Descent and Neural Networks, came up as the most promising performance among the many methodologies investigated. AUC, CA, F1, Precision, and Recall assessment scores were notable, reaching values of 0.978, 0.967, 0.963, 0.966, and 0.967 for each parameter, respectively. AUC, CA, F1, Precision, and Recall scores for neural networks were notably excellent, coming in at 0.982, 0.944, 0.943, 0.943, and 0.944respectively. Although it performed admirably, the Naive Bayes technique had some- what lower results, with AUC, CA, F1, Precision, and Recall values of 0.966, 0.938, 0.937, 0.938, and 0.938, respectively. Similar results were obtained using the Random Forest technique, which received scores of 0.951, 0.913, 0.910, 0.911, and 0.913 for AUC, CA, F1, and Precision, respectively.

Hyperparameter Optimization: To optimize the performance of each procedure, the project's methodology included hyperparameter tuning. Different configurations for various models were produced through the refinement process. With a learning rate of 0.299 and a tree count of 100, Gradient Boosting adopted the "xgboost" method. Utilizing Rectified Linear Unit (ReLU) activation, the Adam solver algorithm, and a regularization parameter (alpha) of 0.0001, 300 neurons were added to neural net- works in hidden layers. For classification and regression problems, stochastic gradient descent (SGD) was optimized with various loss functions. Lasso (L1) was used as the regularization model for SGD, with alpha set at 0.00005 and a constant learning rate of 0.01.

Future Implications and Conclusion: The project's remarkable results demonstrate the legitimacy of the methodology chosen while also highlighting its potential to completely transform the diagnosis and treatment of diabetes. The SGN model's crucial function, which is backed by Neural Networks and other methods, shows how many strategies were successfully combined to handle a significant healthcare concern. The study does, however, acknowledge some shortcomings, mostly attributable to irregularities in the database collection procedure. These flaws suggest chances for the accuracy and utility of the model to be improved and expanded. The initiative establishes a distinct course for the future when looking ahead. It seeks to improve the application's accuracy, increase its accessibility by catering to a range of age groups, and enlarge the dataset to boost its capacity for prediction.

The findings of this project's findings highlight the potential of combining different machine learning and deep learning approaches to develop an integrated system for diabetes detection. The SGN model's performance, together with those of other strategies, highlights the potential of this ground-breaking application to transform diabetes diagnosis and care.

Table 1. The table demonstrating the values obtained after executing various models

| MODEL | AUC | CA | F1 | PRECISION | RECALL |
|---|---|---|---|---|---|
| SGN (Stochastic gradient descent + Neural networks) | 0.978 | 0.967 | 0.963 | 0.966 | 0.967 |
| Neural Network | 0.982 | 0.944 | 0.943 | 0.943 | 0.944 |
| Random Forest | 0.951 | 0.913 | 0.910 | 0.911 | 0.913 |
| Gradient Boosting | 0.971 | 0.938 | 0.937 | 0.938 | 0.938 |
| Naïve bayes | 0.966 | 0.938 | 0.937 | 0.938 | 0.938 |

## 4. Conclusion

Finally, this research, dubbed thorough effort to utilize cutting-edge technologies for resolving a key healthcare challenge. The project's path took a number of careful phases, each of which helped to establish an integrated diabetes detection system in its entirety. As the voyage progressed, a number of crucial realizations arose, influencing both the results attained and the potential directions for future growth. The study prepared itself for the prediction task at hand by naming the target variable "Out- come" and associated it with binary values of 0 (Non-Diabetic) and 1 (Diabetic). The next step was preprocessing, which involved handling missing data, removing unnecessary columns, and using cutting-edge methods like imputation and feature selection. The dataset was strategically split into training and testing subsets in a 70:30 ratio, thus preserving its integrity and usefulness. To guarantee a thorough analysis phase, a 5-fold cross-validation process was also used. Following these preparations, a variety of deep learning and machine learning approaches were used, each carefully chosen and set up to maximize performance. Notably, SGN (Stochastic Gradient Descent + Neural Networks), a unique embedded machine learning technique, stood out as the top model, demonstrating exceptional performance across a range of evaluation metrics.

The success of the project was measured using a wide range of evaluation measures. AUC, Classification Accuracy (CA), F1-score, Precision, and Recall were among the metrics that were included. The SGN model was clearly superior in this thorough evaluation, with outstanding results for AUC (0.978), CA (0.967), F1 (0.963), Precision (0.966), and Recall (0.967). This was supported by Neural Networks' admirable performance, which strengthened the project's position on the viability of the approaches it had chosen. The study attempted the challenging task of optimization, guided by the knowledge that the performance of models depends on hyperparameter configuration. For best performance, each model underwent customized changes and had its neural network architecture specified. Additionally, the project showed its dedication to comprehensive comprehension by using appealing data visualization approaches that highlighted intricate patterns and linkages. As the initiative came to a close, its importance went beyond the short-term successes. It revealed a transparent future course, full of possibilities for improvement and growth. The application's accuracy and user interface need to be improved in order to ensure accessibility for a range of age groups. Additionally, the research hopes to have a greater impact by expanding the dataset and improving predicting abilities by acquiring more data. From preprocessing to model selection, each methodically carried out stage has been highlighted in the context of its results and ramifications. The success of the SGN model is evidence of the impact of technological innovation in healthcare.

The initiative has boldly started on a trajectory of growth, exemplifying the connection between technology and healthcare while accepting some constraints resulting from inconsistencies in data collection. This project delivers a resounding validation of the potential to change diabetes diagnosis and management through integrated machine learning solutions, with a solid basis, strategic insights, and a vision for the future.

## References


[1] Chuter, Vivienne, et al. "Effectiveness of bedside investigations to diagnose periph- eral artery disease among people with diabetes mellitus: a systematic review." Diabetes/metabolism research and reviews (2023): e3683.

[2] Chang, Victor, et al. "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms." Neural Computing and Applications 35.22 (2023): 16157-16173.

[3] Simmons, David, et al. "Treatment of gestational diabetes mellitus diagnosed earlyin pregnancy." New England Journal of Medicine 388.23 (2023): 2132-2144.

[4] Vesa, Cosmin Mihai, and Simona Gabriela Bungau. "Novel molecules in diabetesmellitus, dyslipidemia and cardiovascular disease." International Journal of MolecularSciences 24.4 (2023): 4029.

[5] Kee, Ooi Ting, et al. "Cardiovascular complications in a diabetes prediction model using machine learning: a systematic review." Cardiovascular Diabetology 22.1 (2023): 13.

[6] Chou, Chun-Yang, Ding-Yang Hsu, and Chun-Hung Chou. "Predicting the onsetof diabetes with machine learning methods." Journal of Personalized Medicine 13.3 (2023): 406.

[7] Mansoori, Amin, et al. "Prediction of type 2 diabetes mellitus using hematological factors based on machine learning approaches: a cohort study analysis." Scientific Reports 13.1 (2023): 663.

[8] Challagundla, Yagnesh, et al. "Screening of Citrus Diseases Using Deep Learning Embedders and Machine Learning Techniques." 2023 3rd International conference onArtificial Intelligence and Signal Processing (AISP). IEEE, 2023..

[9] Pal, Someswar, et al. "Deep learning techniques for prediction and diagnosis of diabetes mellitus." 2022 International mobile and embedded technology conference (MECON). IEEE, 2022.

[10] Ganie, Shahid Mohammad, and Majid Bashir Malik. "An ensemble machine learn- ing approach for predicting type-II diabetes mellitus based on lifestyle indicators." Healthcare Analytics 2 (2022): 100092.

[11] Chaki, Jyotismita, et al. "Machine learning and artificial intelligence based DiabetesMellitus detection and self-management: A systematic review." Journal of King Saud University-Computer and Information Sciences 34.6 (2022): 3204-3225.

[12] Shin, Juyoung, et al. "Development of various diabetes prediction models using machine learning techniques." Diabetes Metabolism Journal 46.4 (2022): 650-657.

[13] Malik, Majid Bashir, Shahid Mohammad Ganie, and Tasleem Arif. "Machine learn- ing techniques in healthcare informatics: Showcasing prediction of type 2

diabetes mellitus disease using lifestyle data." Predictive Modeling in Biomedical Data Mining and Analysis. Academic Press, 2022. 295-311.

[14] Ge, Xiaochun, et al. "Application of machine learning tools: Potential and usefulapproach for the prediction of type 2 diabetes mellitus based on the gut microbiome profile." Experimental and Therapeutic Medicine 23.4 (2022): 1-10.

[15] Ye, Jiancheng, et al. "Predicting mortality in critically ill patients with diabetes using machine learning and clinical notes." BMC medical informatics and decision making 20.11 (2020): 1-7

[16] Alghamdi, Manal, et al. "Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project." PloS one 12.7 (2017): e0179805.10.1109/AISP57993.2023.10134971.