AI-Powered Healthcare Systems: A Comprehensive Study on Machine Learning-Based Diabetes Prediction and Online Medical Assistance

Ishika Prashad[1], Bhupinder Singh[2]

{ishikaishika1603@gmail.com[1], singhbhupinder651@gmail.com[2] }

Department of Computer Science and Engineering, Lovely Professional University, Phagwara[1,2]

Abstract. Diabetes is one of the most prevalent chronic diseases worldwide, and early detection is critical for preventing severe complications and improving patient outcomes. This study uses the publicly available Diabetes Prediction Dataset from Kaggle, which contains 100,000 patient records with demographic and clinical features such as age, gender, body mass index (BMI), glycated haemoglobin (HbA1c), blood glucose level, smoking history, hypertension and heart disease. Three machine learning models, a Support Vector Classifier, Gaussian Naïve Bayes and a Random Forest, were compared, with the Random Forest achieving the highest accuracy. After class balancing with resampling, the Random Forest was tuned with a cross-validated grid search and reached an accuracy of 95.20% and a ROC-AUC of 0.969 on a held-out test set of 19,230 patients. Feature-importance analysis identified HbA1c (38%), blood glucose level (29%), age (17%) and BMI (10%) as the most influential risk factors. To support real-world use, we developed a web interface in which users enter their health data and instantly receive a diabetes risk assessment. Overall, the results show that a tuned and interpretable ensemble model can be delivered as an accessible screening tool that supports proactive disease management.

Keywords: Diabetes prediction, Support Vector Classifier, Gaussian Naïve Bayes, Random Forest, Hyperparameter tuning, Feature importance analysis, Model validation, Web application

1 Introduction

Diabetes is a chronic disease that affects a large and growing number of people worldwide. Early detection is the key to reducing its impact and improving patient outcomes. Owing to the increasing volume of digital health data, machine learning (ML) and artificial intelligence (AI) techniques have become very useful for disease prediction, symptom analysis and the provision of online medical consultation [1], [2]. Several terms are central to AI-driven healthcare:

Machine learning (ML): systems that learn patterns from historical patient records.

Supervised learning: training on labelled data to identify health conditions [3].

Feature engineering: transforming raw clinical data (for example, age, BMI and blood glucose) into a representation that helps the model to perform better.

IICTDS 2025, December 19-20, Chandigarh, India
Copyright © 2026 EAI
DOI 10.4108/eai.19-12-2025.2362503

Classification models: algorithms such as support vector machines, Naïve Bayes and Random Forest that predict a health condition from patient features [4].

Online medical assistance systems: symptom checkers or patient portals that provide information about symptoms and risks, often through conversation [5].

AI-powered healthcare has grown rapidly in recent years owing to the availability of large clinical datasets, the increasing popularity of telemedicine, advances in computing power, and the rising demand for automated and remote diagnostic solutions. The research community is consequently developing predictive models, medical virtual assistants, NLP-based healthcare agents and IoT-enabled health-management devices [6]. Nevertheless, these systems still face problems of limited accuracy, low user engagement and a lack of model explainability. In particular, many prediction models are trained with unoptimised parameters, some do not analyse feature importance adequately, real-time prediction through user-friendly interfaces is rarely available [7], and existing conversational agents are not well integrated with predictive analytics tools. A robust, well-tuned, interpretable and easily accessible diabetes prediction system is therefore needed.

In this paper, we develop a diabetes risk model using a Support Vector Classifier, Gaussian Naïve Bayes and a Random Forest, combined with appropriate data preprocessing and feature engineering. We tune the hyperparameters, compare the models using standard performance metrics to identify the most effective classifier, and build a real-time web interface that provides online medical assistance in the form of an instant diabetes risk assessment.

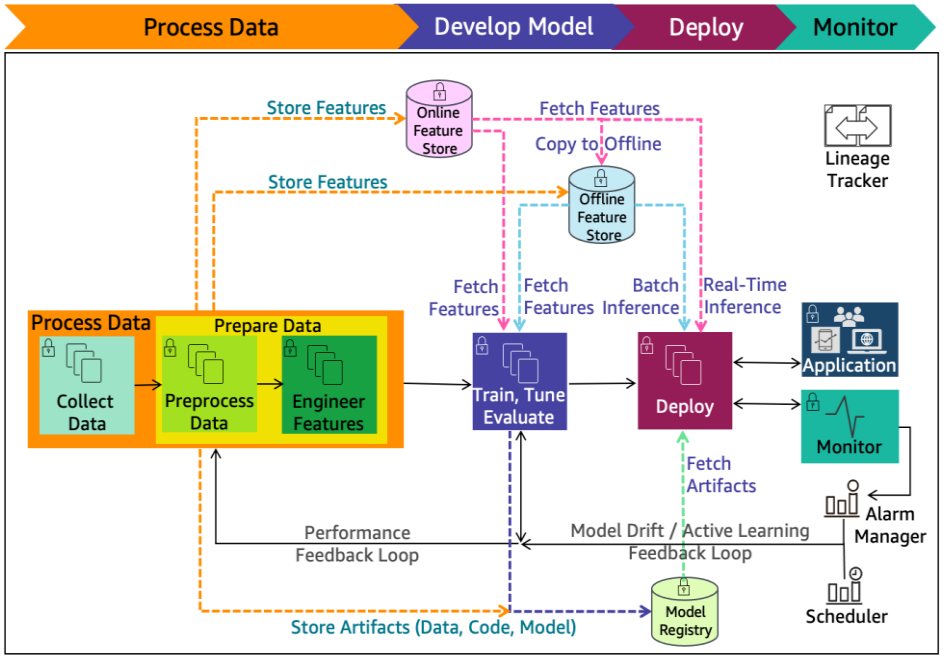


Fig 1. Basic machine learning architecture [9].

Figure 1 outlines the phases of a machine learning lifecycle: data collection, preprocessing, feature engineering, model development, deployment and monitoring. It also highlights the feedback loops, feature store and model registry components [9].

1.1 Related Work

Researchers have applied many machine learning methods, such as convolutional neural networks (CNN), Naïve Bayes, k-nearest neighbours (KNN), decision trees, support vector machines (SVM) and other neural networks, to disease prediction, risk assessment, personalised consultation and medical advice. Machine learning has also been used directly for diabetes risk prediction [8] and for cardiovascular risk prediction [24]. Table 1 summarises representative studies on healthcare chatbots and medical assistance systems, which are closest to the online assistance component of this work.

Table 1. Overview of existing studies

Author	**Objective**	**Techniques Used**	Result
L. Athota et al. (2020) [10]	Provide basic health information	N-gram, TF-IDF	Developed an application that extracts keywords from user queries.
P. Bansode et al. (2021) [11]	Act as a customer-service or personal assistant	Speech recognition, speech-to-text conversion, language translation modules	Used these modules to improve communication and to allow future integration with online doctor consultations.
F. Mehfooz et al. (2021) [12]	Raise awareness about COVID-19	Machine learning combined with real-world data	The system highlights COVID-19 prevention and severity using machine learning.
C. Balasubramaniam et al. (2020) [13]	Provide quick medical advice	AIML, natural language processing	Developed a clinical chatbot for quick consultation and initial precautions.
T. B. Tamizharasi et al. (2022) [14]	Collect medical details and predict diseases	Support vector machine algorithm	Introduced a chatbot that facilitates medical data collection and disease prediction.
N. A. I. Omoregbe et al. (2020) [15]	Provide personalised medical consultation	Natural language processing, fuzzy logic	Constructed a personalised consultation module based on users' self-reported responses.
M. Mittal et al. (2021) [16]	Assist with hospital FAQs	Machine learning algorithms	Linked a chatbot engine with hospital staff for smooth communication.
S. Ayanouz et al. (2020) [17]	Improve healthcare accessibility	Natural language processing, machine learning	Built an easy-to-use chatbot that provides medical help through simple interaction.

K. Sivaraj et al. (2021) [18]	Provide AI-driven medical services	Voice recognition, IoT devices	Developed a multi-featured AI chatbot to improve access to healthcare.
U. Bharti et al. (2020) [19]	Deliver healthcare advice after COVID-19	Natural language processing, AI	Designed a multilingual voice application for healthcare education.
M. Vijayarani et al. (2019) [20]	Provide mental healthcare support	AI, ML, NLP	Reviewed different medical chatbots for mental healthcare.
A. Munshi et al. (2020) [21]	Provide COVID-19 information and services	AI, ML, RASA framework	The proposed chatbot has potential for various uses but requires a large dataset.
A. K. Angappan et al. (2021) [22]	Offer herbal medicine information	AIML, Java	Created an AIML and Java system that provides information about herbal medicine.

The main contributions of this work are:

- Tuning the hyperparameters of the Random Forest to improve classification accuracy and generalisation.
- A comparison of machine learning methods on a diabetes classification task.
- A user-friendly web interface that allows anyone to assess their risk of diabetes.

2 Methodology

2.1 Dataset

The experiments use the Diabetes Prediction Dataset, a publicly available dataset hosted on Kaggle [25]. It contains 100,000 patient records (96,146 after the removal of 3,854 duplicates, of which 8,482, or 8.8%, are diabetic), each described by eight input features and one binary target variable (diabetes: 0 = non-diabetic, 1 = diabetic). The input features are gender, age, hypertension, heart disease, smoking history, BMI, HbA1c level and blood glucose level. The data are stored as a structured CSV file, and the classes are strongly imbalanced; this motivates the resampling described in Section 2.5. Table 2 describes each variable, and Section 3 reports the results obtained on this dataset.

Table 2. Description of the variables in the diabetes dataset

Dataset description	
Variable	**Description**
gender	Gender refers to the biological sex of the individual, which can influence susceptibility to diabetes. It has three categories: male, female and other.
age	Age is an important factor because diabetes is more common in older people. Age in the dataset ranges from 0.08 to 80 years.
hypertension	Hypertension is a condition in which the blood pressure in the arteries remains high for a long time. It takes the values 0 (no hypertension) and 1 (hypertension).
heart_disease	Heart disease, such as congestive heart failure, is a medical condition associated with an increased risk of diabetes. Values: 0 = no heart disease, 1 = heart disease.
smoking_history	Smoking history is a recognised risk factor for diabetes and its harmful effects. The categories are: not current, former, No Info, current, never and ever.
bmi	BMI is a measure of body fat based on weight and height. In the dataset it ranges from 10.01 to 95.69. Categories: <18.5 underweight, 18.5–24.9 normal, 25–29.9 overweight and ≥30 obese.
HbA1c_level	HbA1c is a blood test that shows the average blood sugar level over the previous 2–3 months. Higher levels indicate a higher risk of diabetes, and a value of 6.5% or more is a diagnostic indicator of diabetes.
blood_glucose_level	Blood glucose level is the amount of glucose present in the blood. High levels are a main sign of diabetes.
diabetes	Target variable: 0 = non-diabetic, 1 = diabetic.

2.2 Data Preprocessing

This stage ensures that the data are clean and suitable for model training.

- **Dataset loading:** the diabetes dataset was loaded from a structured CSV file.
- **Handling missing values:** the dataset contains no missing values, which was verified before training.
- **Duplicate removal:** 3,854 duplicate records were removed, leaving 96,146 records.
- **Categorical encoding:** one-hot encoding was applied to the gender and smoking history variables.
- **Value ranges:** age ranges from 0.08 to 80 years, BMI from 10.01 to 95.69, HbA1c from 3.5% to 9.0% and blood glucose from 80 to 300 mg/dL; no records were removed as outliers.

2.3 Feature Engineering

Good features improve model performance. The smoking history variable was regrouped into four meaningful categories: non-smoker (never), current, past (former, not current and ever) and no information. The following steps were also applied:

- The numerical variables were brought to the same scale by standardisation (zero mean and unit variance).
- The categorical variables (gender and smoking history) were one-hot encoded.

2.4 Model Training and Hyperparameter Tuning

Three models were trained:

- Support Vector Classifier (SVC) with a radial basis function kernel and C = 1
- Gaussian Naïve Bayes, with the variance-smoothing parameter selected from three candidate values
- Random Forest Classifier [23]

2.4.1 Train–Test Split

The dataset was divided into a stratified training set (80%, 76,916 records) and a test set (20%, 19,230 records), using a fixed random seed. The test set keeps the natural class distribution of the data and was not used during tuning or resampling.

2.4.2 Hyperparameter Tuning

For the Random Forest, the following hyperparameters were tuned with a grid search (GridSearchCV) using five-fold cross-validation on the training set, with accuracy as the scoring metric:

- Number of estimators (50, 100, 200)
- Minimum samples per leaf (1, 2, 4)
- Minimum samples per split (2, 5)
- Maximum depth (5, 10, unlimited)

2.5 Resampling Techniques

To balance the classes, two resampling techniques were combined: the Synthetic Minority Oversampling Technique (SMOTE), which generates synthetic samples of the minority (diabetic) class, and random undersampling (RandomUnderSampler), which removes samples from the majority (non-diabetic) class. SMOTE raised the minority class to half the size of the majority class, and undersampling then reduced the majority class until the minority class reached 80% of its size. Resampling was performed inside the cross-validation pipeline, so that it affected only the training folds and never the validation or test data.

2.6 Model Evaluation Metrics

The models were evaluated using the following metrics:

1. Accuracy
2. Precision, recall and F1-score
3. Confusion matrix (Table 4)
4. Area under the ROC curve (ROC-AUC)

2.7 Feature Importance Analysis

The Random Forest provides an inherent measure of how much each feature contributes to its predictions. Extracting and visualising these feature importances can show which clinical variables drive the model's decisions and therefore makes the model more interpretable.

2.8 System Workflow

The workflow of the proposed system consists of six stages, which are described below. The web interface that exposes the final stage to users is described in subsection 2.9.

1. Data loading and preprocessing

The diabetes data are loaded from the CSV file.

- The data are checked for duplicates, and missing values are handled appropriately.
- Any categorical variables are converted into numerical variables.

2. Feature engineering

- The category labels of the smoking history variable are regrouped.
- The gender and smoking history variables are converted into one-hot encoded variables.

3. Model training and hyperparameter tuning

- The input features (X) and the target variable (y) are identified.
- A preprocessing pipeline is defined.
- The parameter grid for the Random Forest classifier is specified.
- The best parameters are found by hyperparameter tuning with GridSearchCV.
- The model is fitted to the training data.

4. Model evaluation

- The model is tested on unseen data, and the performance metrics are collected.
- The classification report and confusion matrix are produced.

5. Feature importance analysis

- The feature importances are extracted from the trained model.

6. Prediction

- The user's input data are transformed so that the model can process them.
- The prediction is returned by the trained model.

2.9 Web Interface for Online Medical Assistance

To make the trained model usable without any programming knowledge, a real-time web interface was developed in Python using the Streamlit framework. The application loads the trained Random Forest together with its preprocessing pipeline, so that user input is transformed in exactly the same way as the training data.

The interface works as follows. The user enters eight values in an input form built from Streamlit widgets: gender, age, hypertension (yes or no), heart disease (yes or no), smoking history, BMI, HbA1c level and blood glucose level. When the form is submitted, the entries are passed through the same preprocessing pipeline as the training data, and the trained model returns the predicted class. The result is displayed immediately as a message stating whether the user is predicted to be diabetic or non-diabetic as shown in Figure 2..

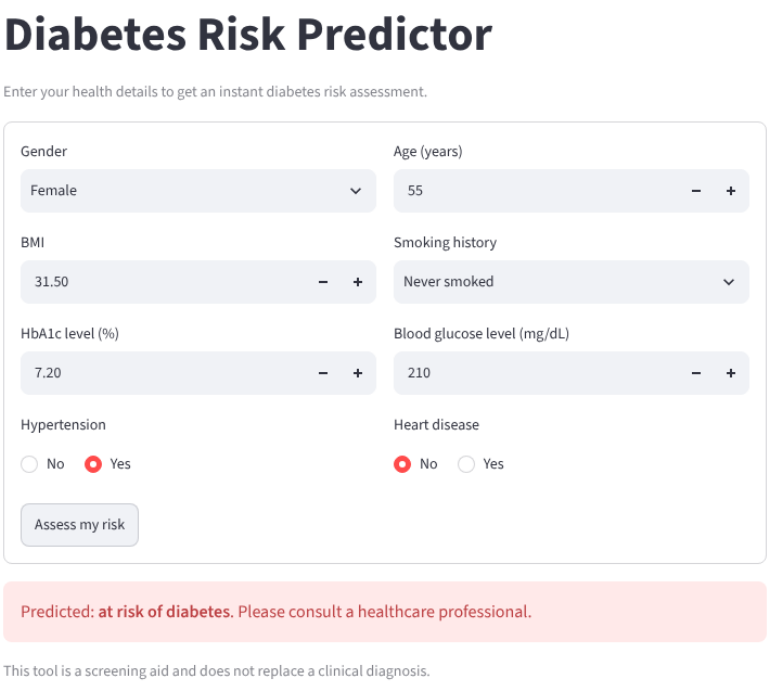


Fig 2. Web interface of the diabetes risk predictor

The interface is intended as a screening aid that supports early awareness and online medical assistance. It does not replace a clinical diagnosis, and users with a positive result should consult a healthcare professional.

3 Results and Discussion

3.1 Model Tuning and Test Accuracy

The Random Forest classifier was trained on data that had been preprocessed with one-hot encoding and resampling (SMOTE and RandomUnderSampler), and its hyperparameters were tuned with GridSearchCV. The optimal parameter values were:

max_depth = unlimited, min_samples_leaf = 1, min_samples_split = 5, n_estimators = 200.

With these parameters, the Random Forest achieved an accuracy of 95.20% and a ROC-AUC of 0.969 on the test set. The grid search selected the largest number of trees (200) and unrestricted depth, which suggests that the model benefits from capacity. Even so, a forest of this size produces a prediction in a fraction of a second, which is fast enough for the real-time interface described in Section 2.9.

3.2 Feature Importance

Figure 3 shows the importance of each feature in the tuned Random Forest. HbA1c level was the most influential feature (37.9%), followed by blood glucose level (28.8%), age (16.8%), BMI (10.3%), hypertension (2.3%) and heart disease (1.6%).

Together, HbA1c level and blood glucose level account for 66.7% of the total importance, so the model relies mainly on the two direct markers of glycaemic control. This agrees with clinical practice, in which an HbA1c value of 6.5% or more is a diagnostic indicator of diabetes (Table 2). Age and BMI together contribute a further 27.1%, which is consistent with the higher risk of diabetes in older and overweight people, whereas hypertension and heart disease contribute little. The six features shown in Figure 3 account for 97.7% of the total importance; the encoded gender and smoking history categories share the remaining 2.3%.

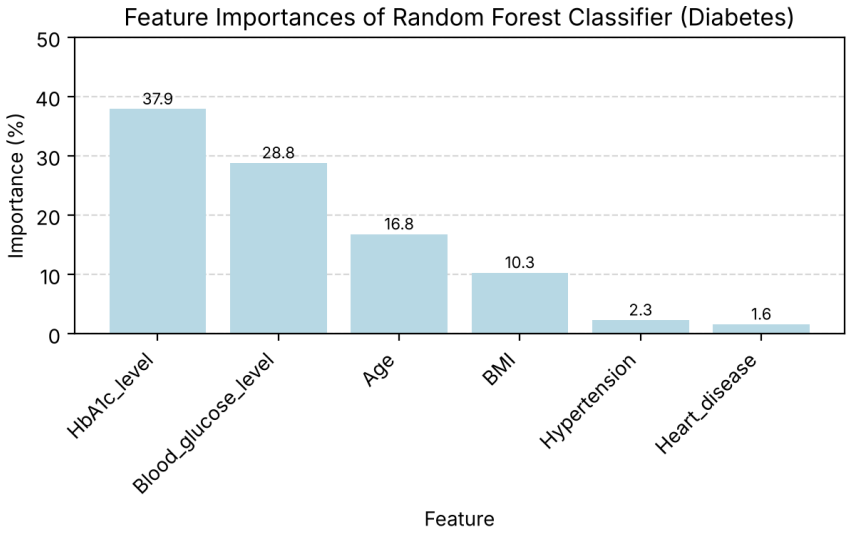


Fig 3. Feature importance of the Random Forest classifier for diabetes prediction

3.3 Classification Performance

Table 3 compares the three models on the test set, and Table 4 shows the confusion matrix of the tuned Random Forest.

Table 3. Performance of the models on the test set (19,230 records)

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Random Forest	95.20%	70.4%	78.6%	0.743	0.969
Support Vector Classifier	90.29%	47.3%	88.4%	0.616	0.964
Gaussian Naïve Bayes	68.46%	21.2%	94.8%	0.346	0.913

3.4 Confusion Matrix Analysis

Table 4. Confusion matrix of the Random Forest on the test set

	Predicted non-diabetic	Predicted Diabetic
Actual non-diabetic	16,973	561
Actual Diabetic	363	1,333

Table 4 shows that, of the 17,534 non-diabetic patients in the test set, 16,973 were classified correctly and 561 were wrongly flagged as diabetic (false positives). Of the 1,696 diabetic patients, 1,333 were identified correctly and 363 were missed (false negatives). From these counts, the precision for the diabetic class is 70.4%, the recall (sensitivity) is 78.6%, the specificity is 96.8% and the F1-score is 0.743. Accuracy is therefore high mainly because the non-diabetic majority is classified very well, whereas the diabetic class is harder: about one in five diabetic patients is missed, and about three in ten positive predictions are false alarms.

Table 3 shows that the Random Forest achieved the highest accuracy (95.20%), precision (70.4%), F1-score (0.743) and ROC-AUC (0.969) of the three models. The Support Vector Classifier was second best (accuracy 90.29%, ROC-AUC 0.964), and Gaussian Naïve Bayes was clearly the weakest (accuracy 68.46%). Both of these models reached a higher recall (88.4% and 94.8%) than the Random Forest, but only at the cost of many more false positives (1,671 and 5,976, respectively), which gives them a precision of only 47.3% and 21.2%. In a screening setting, a missed case is more costly than a false alarm, so the decision threshold of the Random Forest could be lowered to raise recall at the expense of precision; this is a worthwhile direction for future work.

3.5 Discussion

The results show that a tuned Random Forest separates diabetic from non-diabetic cases better than the Support Vector Classifier and Gaussian Naïve Bayes on this dataset. The most important features coincide with established clinical risk markers, which supports the interpretability of the model and shows that its decisions are clinically plausible. The web interface turns the model into an accessible tool, so that individuals can obtain an instant risk assessment without specialist knowledge.

Several limitations should be noted. The study uses a single public dataset, so the reported performance may not transfer to other populations or clinical settings. The dataset also lacks other known risk factors, such as family history, diet and physical activity. In addition, the diabetic class is a minority, so precision and recall for this class are clearly lower than the overall accuracy. Finally, the model has not been validated on independent clinical data, and the interface has not yet been evaluated with real users.

4 Conclusion and Future Scope

In this work, we developed and evaluated a machine learning approach for diabetes prediction using a public dataset of patient records. We trained a Support Vector Classifier, Gaussian Naïve Bayes and a Random Forest, balanced the classes with SMOTE and random undersampling, and tuned the Random Forest with a cross-validated grid search. The tuned Random Forest outperformed the other two models, with a test accuracy of 95.20% and a ROC-AUC of 0.969, and the feature importance analysis showed that HbA1c level and blood glucose level are the dominant predictors, followed by age and BMI. We also implemented a real-time web interface that gives users an instant diabetes risk assessment from their health data, which makes the model accessible to non-specialists.

In future work, we plan to validate the model on external clinical datasets, to include additional risk factors such as family history, diet and physical activity, and to improve recall for the diabetic class, to provide calibrated risk probabilities together with model explanations such as SHAP values. We also plan to integrate the predictor with a conversational agent, so that users can receive guidance in natural language, and to evaluate the interface with real users.

References

[1] Kajal, K., Saini, K., Akhtar, N., Agarwal, D., Rabbani, S., Pervej, Y.: Machine learning for the interpretation and prognosis of chronic illnesses. International Journal of Scientific Research in Science, Engineering and Technology. p. 112 (2024)

[2] An, Q., Rahman, S., Zhou, J., Kang, J. J.: A comprehensive review on machine learning in healthcare industry: classification, restrictions, opportunities, and challenges. Sensors. p. 4178 (2023)

[3] Nandini, S., Likhitha, S.: A multi-condition health prognosis standard based on machine learning algorithms. IEEE GCCIT Conference Proceedings. (2024)

[4] Rahmawati, E., Wibowo, A., Warsito, B.: Evaluation of machine learning algorithms for classifying user perceptions of a child health monitoring application. Jurnal Informatika. p. 90 (2025)

[5] Verma, D., Kushwah, K., Jain, M., Jain, P., Pal, R.: Remote interpretation based on symptoms. International Journal of Research in Applied Science and Engineering Technology. pp. 3739–3747 (2022)

[6] Li, J., Carayon, P.: Health care 4.0: A vision for smart and connected health care. IISE Transactions on Healthcare Systems Engineering. pp. 1–10 (2021)

[7] Collins, G. S., et al.: Clinical prognosis standards using machine learning in oncology: challenges and recommendations. BMJ Oncology. (2025)

[8] Sefid, F., et al.: Predicting diabetes risk using machine learning: a comparative study on the Yazd Health Study (YaHS). Iranian Journal of Diabetes and Obesity. (2025)

[9] AWS: Machine learning lifecycle architecture diagram. AWS Well-Architected Machine Learning Lens. (2024)
[10] Athota, L., Shukla, V. K., Pandey, N., Rana, A.: Chatbot for healthcare using AI. (2020)
[11] Bansode, P., Korekar, G., Badhale, K., Patil, M.: Medical chatbot. International Journal of Advanced Scientific Research and Engineering Trends. (2021)
[12] Mehfooz, F., Jha, S., Singh, S., Soni, S., Sharma, N.: Medical chatbot for novel COVID-19. Springer Proceedings. (2021)
[13] Balasubramaniam, C., Velmurugan, S., Saravanan, M.: Design and development of smart healthcare chatbot application using AI/ML. (2020)
[14] Tamizharasi, T. B., Rajkumar, S.: Building a medical chatbot using support vector machine learning algorithm. NCSET Conference Proceedings. (2020)
[15] Omoregbe, N. A. I., Ndaman, I. O., Misra, S., Abayomi-Alli, O. O., Damasevicius, R.: Text messaging-based medical interpretation using NLP and fuzzy logic. (2020)
[16] Mittal, M., Battineni, G., Singh, D., Nagarwal, T., Yadav, P.: Web-based chatbot for frequently asked queries in hospitals. (2021)
[17] Ayanouz, S., Abdelhakim, B. A., Benhmed, M.: A smart chatbot architecture based on NLP and machine learning for healthcare assistance. (2020)
[18] Sivaraj, K., Jeyabalasuntharam, K., Ganeshan, H., Nagendran, K., Alosious, J., Tharmaseelan, J.: Medibot: end-to-end voice-based AI medical chatbot with a smartwatch. (2021)
[19] Bharti, U., et al.: Medbot: conversational AI-powered chatbot for delivering tele-health after COVID-19. (2021)
[20] Vijayarani, M., Balamurugan, G.: Chatbot in mental health care. Indian Journal of Psychiatric Nursing. (2020)
[21] Munshi, A., Bhardwaj, S., Ramalingam, V. V.: Smart chatbot for COVID-19 using RASA. (2020)
[22] Angappan, A. K., Khan, N. J., Sharan, M., Ahamed, S. F.: AI-based healthcare chatbot system. EasyChair Proceedings. (2021)
[23] Tumpanjawat, T.: Diabetes EDA and random forest hyperparameter tuning. Kaggle Notebook. (2024)
[24] Sitharamulu, V., Maturi, S. M., Murugesan, M., Dudekula, M. R., Battu, H. R.: Efficient machine learning algorithms for cardiovascular risk prognosis. Engineering, Technology & Applied Science Research. (2025)
[25] Mustafa, M.: Diabetes prediction dataset. Kaggle (2023). https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset