Leveraging Machine Learning Models for Diabetes Prediction

Ashwin N Hebbar[1][0009-0000-2949-6992], Shivam Sai Kiran[2][0009-0007-4649-1254],

Mahalakshmi B R[3][0009-0006-2546-3513], Dr. Bharath Setturu[4][0000-0001-6800-9579]

[1,2,3,4]School Of Mathematics and Natural Sciences, Chanakya University, India

ashwinhebbar2k02@gmail.com, shivamsaikiran111@gmail.com, mahalakshmibr02@gmail.com, bharath.s@chanakyauniversity.edu.in,

Abstract —Diabetes is one of the deadliest health conditions mankind faces today, with millions being affected by it worldwide, this is where early detection can significantly reduce complications and save lives. This study evaluates the predictive capabilities of several popular machine learning algorithms *K-Nearest Neighbors (KNN)*, *Decision Trees*, *Random Forests*, and *Logistic Regression* for diabetes diagnosis. A thorough literature review establishes the groundwork, followed by model implementation and a comparative analysis of their accuracy and performance. The results reveal that *Logistic Regression* achieves the highest accuracy at 88%, demonstrating machine learning's potential in proactive healthcare diagnostics and early intervention strategies.

Index Terms —Machine Learning, Diabetes, KNN, Decision Tree, Random Forests, Logistic Regression, Intervention Strategies, Comparative Analysis

1 Introduction

Diabetes is a disease that might occur when human pancreas cannot produce enough insulin. Sometimes, there might be conditions where the body might not be able to use the insulin produced. It is a common disease and is a major contributor to various other conditions or diseases, ranging from severe effects like kidney failures to limb amputation. As of 2019, it was the direct cause of 1.5 million deaths worldwide. [1] In a world where we must live with conditions such as this, it is possible to leverage machine learning algorithms, to predict the likeliness of onset of Diabetes and take the required preventive measures. Our approach demonstrates the use of such ML algorithms. It uses a popular dataset, called as PIMA dataset [2] in order to assess the onset of Diabetes based on several factors.

2 Literature Review

2.1 Traditional Machine Learning Algorithms for Diabetes Prediction:

Numerous studies have investigated the effectiveness of traditional ML algorithms for diabetes prediction. These algorithms, known for their simplicity and interpretability, have shown promising results in classifying individuals as diabetic or non-diabetic.

Khanam and Foo [3], for example, compared 7 ML algorithms on the PIMA Indian Diabetes (PID) dataset and found that Logistic Regression (LR) and Support Vector Machines (SVM) performed particularly well. Sisodia and Sisodia [4] evaluated

Decision Trees (DT), SVM, and Naive Bayes (NB) on the PIDD dataset, with NB achieving the highest accuracy. Mujumdar and V [5] highlighted the potential of boosting classification accuracy by including external factors in the dataset.

Expanding on this, Soni and Varma [6] explored a wider range of techniques, including K-Nearest Neighbors (KNN), LR, DT, SVM, Gradient Boosting (GB), and Random Forest (RF). They reported that RF consistently outperformed other algorithms. In a similar vein, Yahyaoui and Jamil [7] compared SVM, RF, and a Convolutional Neural Network (CNN), finding RF to be the most effective. Rani [8] emphasized the importance of early prediction and explored combining different ML techniques for improved accuracy. These findings suggest that RF, a powerful ensemble learning method, is a

valuable tool for building accurate diabetes prediction models.

2.2 Ensemble Methods for Enhanced Prediction Accuracy:

Ensemble methods, which combine multiple classifiers, offer a way to enhance prediction accuracy. They capitalize on the strengths of individual classifiers to overcome their limitations.

Hasan et al. [9], for instance, proposed a weighted ensemble approach using kNN, DT, RF, AdaBoost, NB, XGBoost, and Multilayer Perceptron (MLP), achieving a remarkable AUC of 0.950. This highlights the potential of ensembles to improve the robustness and accuracy of diabetes prediction.

Building on this, Nahzat and Yağanoğlu [10] compared KNN, RF, SVM, ANN, and DT, concluding that RF provided the highest accuracy. Refat and Al Amin [11] conducted a broad comparison of ML and deep learning techniques, including ensembles, and reported that XGBoost achieved near-perfect accuracy. These studies underscore the power of ensemble methods, especially XGBoost, in achieving high accuracy for diabetes diagnosis.

2.3 Feature Selection and Dimensionality Reduction:

Selecting relevant features and reducing dimensionality are critical steps in building robust and efficient prediction models. High-dimensional datasets can lead to overfitting and reduced generalizability.

Zou and Qu [12] employed Principal Component Analysis (PCA) and minimum Redundancy Maximum Relevance (mRMR) for dimensionality reduction, achieving high accuracy with RF. Dutta and Paul [13] focused on identifying the most influential factors for diabetes prediction. Maniruzzaman [14] used LR to identify risk factors and then employed other algorithms for prediction. Jaiswal et al. [15], in their review, stressed the need for better feature selection and data quality to enhance prediction accuracy. These studies highlight the importance of carefully choosing and refining the input features for optimal model performance.

2.4 Deep Learning Approaches for Diabetes Prediction:

Deep learning, with its ability to learn complex patterns from data, has emerged as a powerful tool for various tasks, including diabetes prediction.

Ayon and Islam [16], for instance, used a deep neural network on the PID dataset, achieving an impressive accuracy of 98.35%. Butt and Letchmunan [17] explored RF, MLP, and LR for classification and LSTM, MA, and LR for prediction. Their results showed the potential of MLP and LSTM for accurate diabetes prediction. In a practical application, Shin and Kim [18] developed diabetes prediction models using gradient boosting and random forest, demonstrating good performance with easily accessible health screening data. These findings point to the promising capabilities of deep learning for accurate diabetes diagnosis.

2.5 Explainable AI and Real-World Applications:

As ML models become more complex, the need for explainability arises. Explainable AI (XAI) aims to make the decision-making process of these models more transparent. Tasian et al. [19] utilized XAI techniques like LIME and SHAP to understand their XGBoost model's predictions. They also developed a website and Android application for real-time prediction, showcasing the potential of translating research into practical tools. Ramesh et al. [20] proposed a remote healthcare framework using SVM for diabetes prediction, emphasizing the role of ML in remote monitoring.

Mir and Dhage [21] explored diabetes prediction on big data, comparing various algorithms and finding SVM to be the most accurate. Suresh and Obulesu [22] focused on building an effective prediction model using R-Studio and the PIMA Indian database. El Massari and Sabouri [23] incorporated ontology in their prediction models, demonstrating the benefits of domain knowledge.

Tigga and Garg [24] concentrated on predicting Type 2 diabetes using questionnaires and machine learning. Ahmed and Issa [25] proposed a fused ML approach combining SVM, ANN, and fuzzy logic, achieving high accuracy. Hasan and Rabbi [26] developed a prediction system using feature selection and AdaBoost, achieving excellent results with the Extra Tree algorithm. Birjais et al. [27] explored predicting future diabetes risk, comparing different algorithms. Farajollahi and Mehmannavaz [28] compared various classifiers for diabetes diagnosis. Panda and Mishra [29] highlighted the importance of feature selection in building accurate models. Dagliati and Marini [30] focused on predicting diabetes complications using a data mining pipeline. Finally, Yuvaraj and SriPreethaa [31] explored diabetes prediction on a Hadoop cluster, demonstrating the potential of big data analytics.

These studies demonstrate the growing interest in developing real-world applications of ML for diabetes, paving the way for more personalized and effective healthcare. Integrating feature selection, dimensionality reduction, and XAI techniques can further enhance model robustness and interpretability. Future research should focus on larger and more diverse datasets, personalized prediction models, and integrating ML into realworld healthcare systems. The continued advancement of ML holds immense promise for improving the lives of individuals affected by diabetes.

3 Methodologies

3.1 About the Dataset

We use the PIMA Indian Diabetes Dataset in this paper, which is a popular dataset used to predict diabetes among patients. Originally derived from the National Institute of Diabetes and Digestive and Kidney Diseases, the dataset is used to predict, if a patient has diabetes or not, based on certain key indicators included in the dataset. All patients here are females at least 21 years old of Pima Indian heritage. The dataset contains the following attributes:

- **Pregnancies**: The number of times a person has been pregnant
- **Glucose**: Glucose concentration in blood plasma in the last 2 hrs
- **BloodPressure**: Blood Pressure (Diastolic) in mm Hg

- **SkinThickness**: Thickness of skin in the tricep fold (in mm)
- **Insulin**: Insulin levels in a 2 hr window in the blood serum. (in mu U/ml)
- **BMI**: Body Mass Index
- **DiabetesPedigreeFunction**: Diabetes pedigree function
- **Age**: Age (in number of years)
- **Outcome**: Dependent Variable or Label. 0 represents no Diabetes, 1 means Diabetes.

The dataset is also not greatly imbalanced. With an ample availability of both Classes 0 and 1, Fig. 1 Shows the distribution:

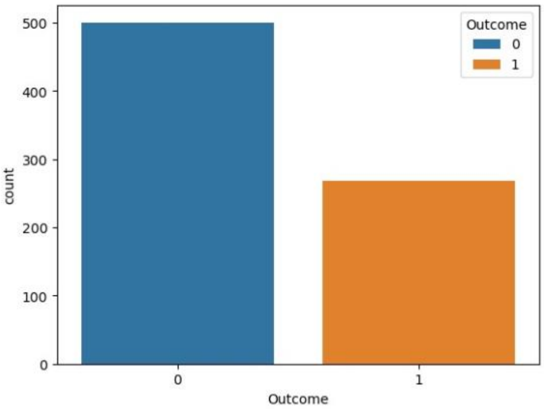

Fig. 1. Class Availablity

3.2 Exploratory Data Analysis

Exploratory Data Analysis, or EDA for short, is a process through which the given data is explored and understood for further use. This stage involves using data visualization methodologies, as well as textual analysis through quantitative aggregate functions such as sum, count, and so on.

The attributes are distributed normally and binormally, as shown in Fig. 2 & 3:

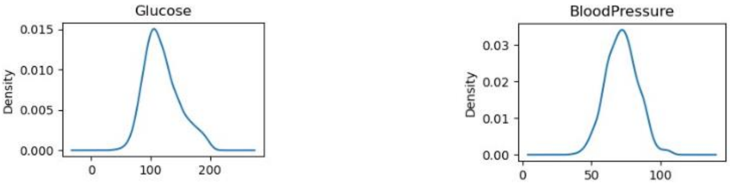

Fig. 2. EDA: Distribution of Attributes

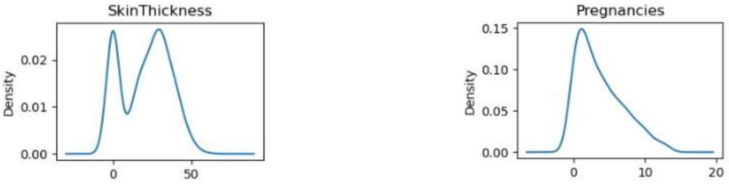

Fig. 3. EDA: Distribution of Attributes II

Along with this, we also explored a correlation heatmap (Fig. 4), which brought out the following observations:

- **Age and Pregnancies** have a moderate positive correlation (0.54)
- **Glucose and Outcome** have a strong positive correlation (0.47).
- **BMI and SkinThickness** have a moderate positive correlation (0.39).
- **Insulin and Glucose** have a moderate positive correlation (0.33).

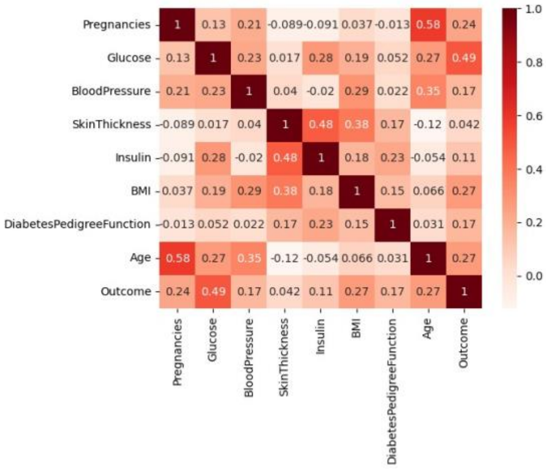

Fig. 4. EDA: Correlation Heatmap

Other basic EDA was also performed, such as shape, checking attribute types, etc.

3.3 Pre Processing

In machine learning, Preprocessing is the process of removal of incorrect or irrelevant data. This might include noise, outliers and other unwanted inconsistencies that might impact the performance of our models.

Minimal Preprocessing was employed on the dataset, specifically, we have removed the outlier classes using Inter Quartile Range method, the formula for the same is as mentioned below:

$$1.5 * Q3 - Q1 \tag{1}$$

Table 1 shows outlier counts on a per-column basis.

Table 1. Attribute Information

Attribute	Number of Outliers
Pregnancies	4
Glucose	5
BloodPressure	45
SkinThickness	1

Insulin	34
BMI	19
DiabetesPedigreeFunction	29
Age	9
Outcome	0

This ensures two things mainly:

1) The dataset contains outliers, these outliers are now removed.
2) Some columns had empty values (i.e zeroes) they are also removed.

After the removal of these rows, our dataset is now ready for training purposes.

3.4 Models Used

Here, we have vividly demonstrated the use of KNN, Decision Trees, Random Forest and Logistic Regression. Where KNN and Logistic Regression are mainly used for classification, while, Decision Trees and Random Forests on the other hand can be used for classification tasks as well, as here we are required to classify based on the given attributes as mentioned above.

For each model, we have chosen a list of hyperparameters (if needed) to reduce the computational complexity. While the tests are not exhaustive, they are representative of a real-world scenario with limited computing capabilities.

For all of our estimators, we have used the Scikit-Learn library's [32] implementation to handle, train and assess the models.

K - Nearest Neighbours (KNN)

The KNN algorithm is a classic and popular algorithm used for classification tasks. The idea behind the algorithm is to label given n attributes based on the k-nearest points to the position of that data point. This is an intuitive and simple algorithm that is extensively used in the field of machine learning.

For this estimator, the (hyper) parameters chosen as Table 2

Table 2. Parameters: KNN

K-Size	**Test Set Size**
7, 9, 19, 21	0.1, 0.2

Results

Here's a snippet of the classification report (Table 3)

Table 3. Results: KNN

KNN Model	precision	recall	f1score	support
0	0.83	0.94	0.88	48
1	0.70	0.44	0.54	16
accuracy	-	-	0.81	64

macro avg	0.77	0.69	0.71	64
weighted avg	0.80	0.81	0.80	64

The above results have come with the following hyper parameters:

1) K-Size: 7
2) Test-size: 0.1

Decision Tree Classifier

Decision Tree classifier is another classic algorithm that is simple yet robust. They are non-parametric in nature, which means the performance of the model does not depend on the underlying distribution of the data (or attributes). This means that decision trees are generally better performing in most instances compared to other model which require extensive preprocessing and understanding of the underlying distributions. One more advantage of this dataset is that it is possible to visualize the tree with ease, which increases model understanding.

Parameters chosen

For this model, we have chosen the following hyperparameters as shown in Table 4:

Table 4. Parameters: Decision Tree

Test Set Size	Scaling Method
0.1 & 0.2	MinMax Scaler, Robust Scaler, Normalizer, Standard Scaler

Results

Table 5. Results: Decision Tree

Decision Tree	precision	recall	f1score	support
0	0.95	0.81	0.88	52
1	0.50	0.83	0.62	12
accuracy	-	-	0.81	64
macro avg	0.73	0.82	0.75	64
weighted avg	0.87	0.81	0.83	64

We see similar results from the decision tree classifier. These results in Table 5 are under the following conditions: 1) MinMax Scaling
2) Test Set: 0.1

Random Forest Classfier

The Random Forest Classifier is what we class as an “Ensemble Method”, this is is a meta estimator that in actuality, trains a number of DecisionTreeClassifiers. It does this on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. This is also an extremely resilient and robust classification estimator that performs generally well due to it’s non parametric nature.

Parameters Considered

For this, we take a similar approach as the Decision Tree Classifier (Table 6).

Table 6. Parameters: Random Forest Classifier

Test Set Size	Scaling Method
0.1 & 0.2	MinMax Scaler, Robust Scaler, Normalizer, Standard Scaler

It is worth noting that more hyperparameters that is tuned are:

1) ‘criterion’
2) ‘min_sample_split’
3) ‘max_depth’ among others

Results

Here, we see a deviation from the above Decision Tree’s results (See Table 7).

Table 7. Results: Random Forest Classifier

Random Forest	precision	recall	f1score	support
0	0.90	0.93	0.91	57
1	0.78	0.70	0.74	20
accuracy	-	-	0.87	77
macro avg	0.84	0.81	0.83	77
weighted avg	0.87	0.87	0.87	77

These results are with the following hyper parameters:

1) Test Size: 0.1
2) StandardScaler()

Using the above settings, we get out highest yet accuracy score across all estimators being used in this paper.

Logistic Regression

Logistic Regression, also called the “LogIt” Estimator is another commonly used estimator for classification tasks. Usually, the logit model is used for binary classification tasks, which is our case is suitable. While it is possible to use the Logistic Regression Model in multi-class scenarios, it usually does not perform and scale well.

Parameters

We only take the test size as the parameter here (Table 8).

Table 8. Parameters: Logistic Regression

Test Set Size
0.1 & 0.2

Results

Since Logistic Regression is an estimator suitable for binary classification, we see that the model performs exceptionally well on our pre-processed dataset. Results in Table 9

Table 9. Result: Logistic Regression

Logistic Regression	precision	recall	f1score	support
0	0.88	0.96	0.92	51
1	0.90	0.73	0.81	26
accuracy	-	-	0.88	77
macro avg	0.89	0.85	0.86	77
weighted avg	0.89	0.88	0.88	77

It is also worth noting that the model can perform even better by choosing the appropriate penalty strategy.

4 Results

Table 10 shows results extracted from the above Classification Reports are as follows. While additional information is available in the tables, here, we will be focusing on the accuracy scores of each model in it's best-performing instance. We see the following results (Fig. 5):

Table 10. Performance of Estimators:

Model	Accuracy Score
KNN Classifier	81%
Decision Tree Classifier	81%
Random Forest Classifier	87%
Logistic Regression	88%

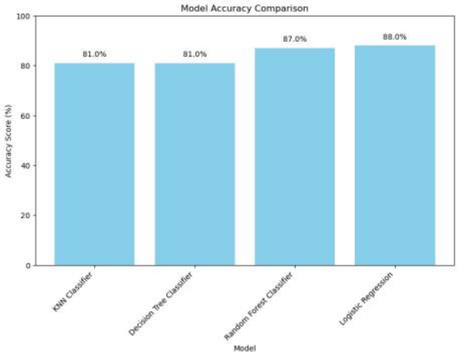

Fig. 5. Visualization of Estimator Performance

5 Further Improvements

While this paper has explored a series of popular classification algorithms, the ocean of machine learning is large, some improvements that could be made are:

1) **Testing with more complex and accurate models**, such as Perceptron-based models, and neural networks
2) **Hyperparameter Tuning**: These models can be further improved for a specific dataset through hyperparameter tuning.

3) **Feature Engineering**: Through the use of existing attributes, it is possible to derive more attributes which may further help in increasing the model's performance.

6 Conclusion

Our work confirms that it is possible to predict the onset of Diabetes Mellitus with great accuracy. We have explored a detailed literature study along with the comparative analysis of four prominent machine learning models; K-Nearest Neighbors, Decision Trees, Random Forest, and Logistic Regression revealing nuanced insights into predictive performance and model optimization strategies. By meticulously preprocessing the PIMA Indian Diabetes Dataset and employing rigorous scaling and parameter tuning techniques, researchers can substantially enhance predictive accuracy and model reliability.

The empirical results demonstrate significant variations in model performance, with Logistic Regression achieving the highest accuracy at 88% with Standard Scaler as the scaling technique and the K-size being "7", closely followed by Random Forest at 87%, and then the significantly lower accuracy category; KNN and Decision Tree Classifier, both at 81%. These findings underscore the importance of systematic model selection, preprocessing, and optimization in medical predictive analytics. The strategic use of techniques like Interquartile Range outlier removal and advanced feature scaling emerged as pivotal in improving model performance, highlighting the intricate relationship between data preparation and predictive efficacy.

Beyond mere statistical metrics, this research illuminates the broader implications of machine learning in proactive healthcare intervention. By developing robust predictive models capable of early diabetes detection, medical professionals can transition from reactive to preventative diagnostic strategies. The research suggests that with continued methodological refinement, increasingly sophisticated machine learning approaches such as advanced neural networks or sophisticated ensemble boosting methods could further enhance predictive precision and clinical applicability.

The scalability and potential of these predictive models extend far beyond the current study. As healthcare continues to embrace data-driven technologies, machine learning stands poised to revolutionize diagnostic processes, enabling more personalized, timely, and effective medical interventions. Future research should focus on expanding dataset diversity, integrating more complex feature engineering techniques, and exploring cutting-edge algorithmic approaches to push the boundaries of predictive healthcare analytics.

References

1. World Health Organization: Diabetes, (2023).
2. UCI Machine Learning Repository: Pima Indians Diabetes Database, (2016).
3. Khanam, J.J., Foo, S.Y.: A comparison of machine learning algorithms for diabetes prediction. Ict Express. 7, 432–439 (2021).
4. Sisodia, D., Sisodia, D.S.: Prediction of diabetes using classification algorithms. Procedia computer science. 132, 1578–1585 (2018).
5. Mujumdar, A., Vaidehi, V.: Diabetes prediction using machine learning algorithms. Procedia Computer Science. 165, 292–299 (2019).

6. Soni, M., Varma, S.: Diabetes prediction using machine learning techniques. International Journal of Engineering Research & Technology (IJERT). 9, 2278–2181 (2020).
7. Yahyaoui, A., Jamil, A., Rasheed, J., Yesiltepe, M.: A decision support system for diabetes prediction using machine learning and deep learning techniques. In: 2019 1st International informatics and software engineering conference (UBMYK). pp. 1–4 (2019).
8. Rani, K.: Diabetes prediction using machine learning. International Journal of Scientific Research in Computer Science, Engineering and Information Technology. 6, 294–305 (2020).
9. Hasan, M.K., Alam, M.A., Das, D., Hossain, E., Hasan, M.: Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access. 8, 76516–76531 (2020).
10. Nahzat, S., Yağanoğlu, M.: Diabetes prediction using machine learning classification algorithms. Avrupa Bilim ve Teknoloji Dergisi. 53–59 (2021).
11. Refat, M.A.R., Al Amin, M., Kaushal, C., Yeasmin, M.N., Islam, M.K.: A comparative analysis of early stage diabetes prediction using machine learning and deep learning approach. In: 2021 6th International Conference on Signal Processing, Computing and Control (ISPCC). pp. 654–659 (2021).
12. Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., Tang, H.: Predicting diabetes mellitus with machine learning techniques. Frontiers in genetics. 9, 515 (2018).
13. Dutta, D., Paul, D., Ghosh, P.: Analysing feature importances for diabetes prediction using machine learning. In: 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON). pp. 924–928 (2018).
14. Maniruzzaman, M., Rahman, M.J., Ahammed, B., Abedin, M.M.: Classification and prediction of diabetes disease using machine learning paradigm. Health information science and systems. 8, 1–14 (2020).
15. Jaiswal, V., Negi, A., Pal, T.: A review on current advances in machine learning based diabetes prediction. Primary Care Diabetes. 15, 435–443 (2021).
16. Ayon, S.I., Islam, M.M.: Diabetes prediction: a deep learning approach. International Journal of Information Engineering and Electronic Business. 13, 21 (2019). 17. Butt, U.M., Letchmunan, S., Ali, M., Hassan, F.H., Baqir, A., Sherazi, H.H.R.: Machine learning based diabetes classification and prediction for healthcare applications. Journal of healthcare engineering. 2021, 9930985 (2021).
18. Shin, J., Kim, J., Lee, C., Yoon, J.Y., Kim, S., Song, S., Kim, H.-S.: Development of various diabetes prediction models using machine learning techniques. Diabetes & Metabolism Journal. 46, 650–657 (2022).
19. Tasin, I., Nabil, T.U., Islam, S., Khan, R.: Diabetes prediction using machine learning and explainable AI techniques. Healthcare Technology Letters. 10, 1–10 (2023).
20. Ramesh, J., Aburukba, R., Sagahyroon, A.: A remote healthcare monitoring framework for diabetes prediction using machine learning. Healthcare Technology Letters. 8, 45–57 (2021).
21. Mir, A., Dhage, S.N.: Diabetes disease prediction using machine learning on big data of healthcare. In: 2018 fourth international conference on computing communication control and automation (ICCUBEA). pp. 1–6 (2018).

22. Suresh, K., Obulesu, O., Ramudu, B.V.: Diabetes prediction using machine learning techniques. Helix-The Scientific Explorer| Peer Reviewed Bimonthly International Journal. 10, 136–142 (2020).
23. El Massari, H., Sabouri, Z., Mhammedi, S., Gherabi, N.: Diabetes prediction using machine learning algorithms and ontology. Journal of ICT Standardization. 10, 319–337 (2022).
24. Tigga, N.P., Garg, S.: Prediction of type 2 diabetes using machine learning classification methods. Procedia Computer Science. 167, 706–716 (2020).
25. Ahmed, U., Issa, G.F., Khan, M.A., Aftab, S., Khan, M.F., Said, R.A., Ghazal, T.M., Ahmad, M.: Prediction of diabetes empowered with fused machine learning. IEEE Access. 10, 8529–8538 (2022).
26. Hasan, S.M., Rabbi, M.F., Champa, A.I., Zaman, M.A.: An effective diabetes prediction system using machine learning techniques. In: 2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT). pp. 23–28 (2020).
27. Birjais, R., Mourya, A.K., Chauhan, R., Kaur, H.: Prediction and diagnosis of future diabetes risk: a machine learning approach. SN Applied Sciences. 1, 1–8 (2019).
28. Farajollahi, B., Mehmannavaz, M., Mehrjoo, H., Moghbeli, F., Sayadi, M.J.: Diabetes diagnosis using machine learning. Frontiers in Health Informatics. 10, 65 (2021).
29. Panda, M., Mishra, D.P., Patro, S.M., Salkuti, S.R.: Prediction of diabetes disease using machine learning algorithms. IAES International Journal of Artificial Intelligence. 11, 284 (2022).
30. Dagliati, A., Marini, S., Sacchi, L., Cogni, G., Teliti, M., Tibollo, V., De Cata, P., Chiovato, L., Bellazzi, R.: Machine learning methods to predict diabetes complications. Journal of diabetes science and technology. 12, 295–302 (2018).
31. Yuvaraj, N., SriPreethaa, K.: Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster. Cluster Computing. 22, 1–9 (2019).
32. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine Learning in Python, (2011).