

Diabetes Prediction Using Machine Learning

ALISHER

Department of Software Engineering

University of Sialkot

Sialkot, Pakistan

1230100392@uskt.edu.pk

line 1: 4th Given Name Surname
line 2: *dept. name of organization*
(of Affiliation)
line 3: *name of organization*
(of Affiliation)
line 4: City, Country
line 5: email address or ORCID

line 1: 2nd Given Name Surname
line 2: *dept. name of organization*
(of Affiliation)
line 3: *name of organization*
(of Affiliation)
line 4: City, Country
line 5: email address or ORCID

line 1: 3rd Given Name Surname
line 2: *dept. name of organization*
(of Affiliation)
line 3: *name of organization*
(of Affiliation)
line 4: City, Country
line 5: email address or ORCID

line 1: 6th Given Name Surname
line 2: *dept. name of organization*
(of Affiliation)
line 3: *name of organization*
(of Affiliation)
line 4: City, Country
line 5: email address or ORCID

Abstract— This study uses a real-world dataset to demonstrate a machine learning-based method for diabetes prediction. The objective is to train and assess various classification models in order to ascertain which one is most effective in identifying diabetic individuals. We made use of a publically accessible dataset that included lifestyle and medical characteristics like smoking history, age, gender, BMI, and blood glucose levels. Four models were trained and compared: Support Vector Machine, Random Forest, Decision Tree, and Logistic Regression. F1-score, recall, accuracy, and precision were used to assess each model. According to the results, Random Forest performed the best across all measures, demonstrating its reliability as a model for early diabetes prediction.

Keywords—Diabetes, *Machine Learning, Classification, KNN, SVM, Random Forest, Health Care Prediction*

I. INTRODUCTION

Globally, diabetes is becoming a more serious health issue. Serious health issues associated with it might be lessened with early detection and treatment. Early diabetes detection can be challenging, though, particularly if conventional screening techniques are the only ones used. By identifying trends in patient data and forecasting a person's likelihood of having diabetes, machine learning can aid in this process.

To create a predictive model for this study, we employed a dataset of 100,000 entries and a variety of health-related characteristics. Our objective was to experiment with various machine learning algorithms in order to determine which one is most effective at predicting diabetes. We concentrated on employing straightforward models that are simple to apply in real-world health system.

II. LITERATURE REVIEW

Due to the growing number of diabetic patients worldwide and the promise of data-driven decision-making in healthcare, the use of machine learning (ML) and deep learning (DL) in the early detection and diagnosis of diabetes has attracted a lot of attention recently. The efficiency of ML/DL approaches in predicting diabetes using clinical, lifestyle, and physiological data has been shown in numerous studies. Chou et al. (2023) used a variety of supervised models, such as Logistic Regression, Neural Networks, Decision Jungle, and Boosted Decision Tree, to investigate diabetes prediction in Taiwanese women. With an accuracy of 95.3% and an AUC of 0.991, their results showed that Boosted Decision Tree performed best, underscoring the usefulness of ensemble-based methods in clinical prediction. Similarly, using the Pima Indian Diabetes Dataset, Alaa Khaleel and Al-Bakry (2021) assessed classical models and found that Logistic Regression outperformed Naïve Bayes and K-Nearest Neighbours (KNN), achieving a noteworthy 94% accuracy. A framework for ensemble machine learning using Random Forest, XGBoost, MLP, and KNN classifiers was presented by Hasan et al. (2020). They achieved a remarkable AUC of 0.950 by combining outlier elimination, missing value imputation, and a weighted ensemble method based on AUC scores, demonstrating how ensemble learning can improve predictive robustness. Chellappan and Rajaguru (2025) used gene expression datasets from the Nordic Islet Transplant Program with the PIMA dataset to broaden the data breadth and address the problem of model generalisability. Their hybrid approach achieved 98.13% and 97.14% accuracy on the PIMA and NITP datasets, respectively, by employing ABC-PSO for feature selection and SVM with an RBF kernel for sample classification. This method showed how multi-source data integration and hybrid optimisation might increase accuracy and flexibility across populations. Concurrently, Shaheen et al. (2024) combined LeNet, Highway Networks, and Temporal Convolutional Networks (TCN) to create two sophisticated deep ensemble models: Hi-Le and HiTCLe. The Hi-Le model achieved a 94% accuracy and 96% F1-score using ProWSyn for oversampling and SHAP for feature interpretability, suggesting that hybrid DL architectures can manage unbalanced datasets and extract significant features. In their thorough analysis of machine learning and deep

learning approaches for diabetes detection, Sharma and Shah (2021) noted that deep learning methods like Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks routinely outperformed conventional machine learning algorithms, especially when used with image and time-series health data. These models' ability to autonomously extract high-level feature representations allowed them to frequently achieve accuracy rates exceeding 90%. Rehan et al. (2024) took a novel approach by utilising Laser-Induced Breakdown Spectroscopy (LIBS) on fingernail samples to diagnose diabetes non-invasively. The model obtained 96% accuracy and 99.9% precision using several classifiers after dimensionality reduction with Principal Component Analysis (PCA), indicating a possible substitute for situations where blood data may not be practical. A wearable sensor-based system that included hip, knee, and ankle sensors to record gait acceleration data was presented by Chee et al. (2024). Their CNN-LSTM hybrid model demonstrated the promise of real-time physical monitoring in diabetes prediction by processing these biomechanical signals and achieving an accuracy of 91.25%. The PIMA and Indian Diabetes datasets were among the many models and datasets examined by Katiyar et al. (2024) in a larger meta-analysis to assess the overall effectiveness of ML and DL models in disease diagnosis and treatment. According to their investigation, numerous models of Artificial Neural Networks (ANN) achieved over 98% accuracy, indicating exceptional general performance. Additionally, DiaNet v2, a deep learning network trained on retinal fundus pictures gathered from Qatari patients, was created by Al-Absi et al. (2024). The model outperformed previous image-based models with a 92% accuracy, 93% sensitivity, and 91% specificity, thanks to the inclusion of convolutional layers that could detect patterns of diabetic retinopathy. Other noteworthy research has also looked into combining behavioural and clinical data to increase prediction rates. For example, research employing models such as LightGBM, AdaBoost, and multilayer perceptrons (MLP) in electronic health records (EHR) has demonstrated encouraging accuracy levels exceeding 90% when integrating longitudinal health data. Additionally, transfer learning and optimising pre-trained DL models have begun to gain traction, particularly in genomics-based diagnosis and medical imaging. In order to ensure interpretability and that medical practitioners can comprehend and trust model predictions, the usage of feature significance methodologies, such as SHAP and LIME, has grown in importance. These developments highlight the fact that explainability and generalisability are still major issues when using AI solutions in actual healthcare settings, even though accuracy is crucial.

III. METHODOLOGY

1. Dataset Description and Preprocessing

Diabetes_prediction_dataset.csv, a dataset with nine features for one hundred thousand people, was used. Among these characteristics are

- age;
- gender

Hypertension, heart disease, smoking history, BMI, HbA1c level, blood glucose level, and diabetes (target variable). Label-encoded categorical characteristics were utilised. Not a single value was missing. To address class imbalance, the

Synthetic Minority Over-sampling Technique, or SMOTE, was used.

IV. ALGORITHMS

1. K-Nearest Neighbors (KNN):

KNN is a distance-based learning method that uses the majority class of its nearest neighbors to classify new data points. It does not assume anything about the distribution of the underlying data and performs well with tiny datasets. Here, it was used to find similar patients in the feature space in order to forecast diabetes.

KNN Results:

| ===== | | | | | | | | | |
|-------------------------------|---------------------|---------------|----------------------|----------------|----------------------|--|--|--|--|
| KNN | | | | | | | | | |
| ===== | | | | | | | | | |
| Accuracy: 0.9625 | | | | | | | | | |
| Precision: 0.9059 | | | | | | | | | |
| Recall: 0.6415 | | | | | | | | | |
| F1 Score: 0.7511 | | | | | | | | | |
| ROC AUC: 0.8991 | | | | | | | | | |
| Classification Report: | | | | | | | | | |
| | precision | recall | f1-score | support | | | | | |
| | 0 | 0.97 | 0.99 | 0.98 | 17534 | | | | |
| | 1 | 0.91 | 0.64 | 0.75 | 1696 | | | | |
| | | | | | 0.96 19230 | | | | |
| | accuracy | | macro avg | | 0.94 0.82 0.87 19230 | | | | |
| | weighted avg | | 0.96 0.96 0.96 19230 | | | | | | |
| Confusion Matrix: | | | | | | | | | |
| [[17421 113] | | | | | | | | | |
| [608 1088]] | | | | | | | | | |

2. Support Vector Machine (SVM):

SVM is a potent classifier that determines the best hyperplane to divide classes with the greatest amount of margin. It works well with small sample sizes and is efficient in high-dimensional spaces. In this case, it was employed to identify patients who were at risk for Diabetes.

SVM Results:

```
=====
SVM
=====
Accuracy: 0.9647
Precision: 0.9990
Recall: 0.6008
F1 Score: 0.7504
ROC AUC: 0.9331

Classification Report:
precision    recall   f1-score   support
0            0.96     1.00      0.98     17534
1            1.00     0.60      0.75     1696

accuracy           0.96     19230
macro avg         0.98     0.80      0.87     19230
weighted avg      0.97     0.96      0.96     19230

Confusion Matrix:
[[17533    1]
 [ 677 1019]]
```

3. Gradient Boosting:

Gradient Boosting is a popular machine learning algorithm used for regression and classification tasks. It combines multiple weak models to create a strong predictive model. Here we in diabetes prediction.

Gradient Boosting Results:

```
=====
Gradient Boosting
=====
Accuracy: 0.9718
Precision: 0.9983
Recall: 0.6810
F1 Score: 0.8097
ROC AUC: 0.9766

Classification Report:
precision    recall   f1-score   support
0            0.97     1.00      0.98     17534
1            1.00     0.68      0.81     1696

accuracy           0.97     19230
macro avg         0.98     0.84      0.90     19230
weighted avg      0.97     0.97      0.97     19230

Confusion Matrix:
[[17532    2]
 [ 541 1155]]
```

4. Random Forest:

Random Forest is a popular machine learning algorithm used for classification and regression tasks. It's an ensemble learning method that combines multiple decision trees to improve the accuracy and robustness of predictions. Here we are using in diabetes prediction.

Random Forest Result:

```
=====
Random Forest
=====
Accuracy: 0.9680
Precision: 0.9253
Recall: 0.6934
F1 Score: 0.7927
ROC AUC: 0.9549

Classification Report:
precision    recall   f1-score   support
0            0.97     0.99      0.98     17534
1            0.93     0.69      0.79     1696

accuracy           0.97     19230
macro avg         0.95     0.84      0.89     19230
weighted avg      0.97     0.97      0.97     19230

Confusion Matrix:
[[17439    95]
 [ 520 1176]]
```

5. Decision Tree:

A decision tree bases its choices on feature values and employs a structure resembling a tree. It is perfect for figuring out which patient factors lead to diabetes because it is simple to interpret and depict.

Decision Tree Result:

```
=====
Decision Tree
=====
Accuracy: 0.9491
Precision: 0.6983
Recall: 0.7453
F1 Score: 0.7210
ROC AUC: 0.8575

Classification Report:
precision    recall   f1-score   support
0            0.98     0.97      0.97     17534
1            0.70     0.75      0.72     1696

accuracy           0.95     19230
macro avg         0.84     0.86      0.85     19230
weighted avg      0.95     0.95      0.95     19230

Confusion Matrix:
[[16988    546]
 [ 432 1264]]
```

6. Logistic Regression:

Logistic Regression is a popular machine learning algorithm used for binary classification tasks. It predicts the probability of an event occurring based on a set of input features. Here we using in diabetes prediction.

Logistic Regression Result:

```

=====
Logistic Regression
=====

Accuracy: 0.9593
Precision: 0.8783
Recall: 0.6256
F1 Score: 0.7307
ROC AUC: 0.9585

Classification Report:
precision    recall    f1-score   support

          0       0.96      0.99      0.98     17534
          1       0.88      0.63      0.73      1696

   accuracy                           0.96     19230
    macro avg       0.92      0.81      0.85     19230
weighted avg       0.96      0.96      0.96     19230

Confusion Matrix:
[[17387 147]
 [ 635 1061]]

```

Final Model Comparison:

| | accuracy | precision | recall | f1 | roc_auc |
|---------------------|----------|-----------|----------|----------|----------|
| Gradient Boosting | 0.971763 | 0.998271 | 0.681014 | 0.809674 | 0.976567 |
| Tuned Random Forest | 0.971607 | 1.000000 | 0.678066 | 0.808152 | 0.971262 |
| Logistic Regression | 0.959334 | 0.878311 | 0.625590 | 0.730716 | 0.958503 |
| Random Forest | 0.968019 | 0.925256 | 0.693396 | 0.792720 | 0.954869 |
| SVM | 0.964743 | 0.999020 | 0.600825 | 0.750368 | 0.933144 |
| KNN | 0.962507 | 0.905912 | 0.641509 | 0.751122 | 0.899089 |
| Decision Tree | 0.949142 | 0.698343 | 0.745283 | 0.721050 | 0.857539 |

V. CONCLUSION

Random Forest outperformed all other models, achieving the highest F1-score, accuracy, precision, and recall. This demonstrates how prediction quality may be raised by ensemble learning with several decision trees. SVM and Logistic Regression both did well, though marginally worse than Random Forest. Our research backs the use of machine learning to help with diabetes early diagnosis.

VI. FUTURE WORK

To improve temporal prediction skills and facilitate continuous monitoring, future studies can concentrate on incorporating real-time data from wearable health devices like fitness trackers, smartwatches, and continuous glucose monitors. Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs) are two examples of deep learning models that could be used to identify intricate patterns and time-based correlations in patient data. Model generalisability and personalisation would be enhanced by enlarging the dataset to include retinal imaging, genomic data, and populations with a variety of nationalities or ethnicities. To improve model

interpretability for doctors, explainable AI methods like SHAP and LIME should also be used. Investigating federated learning may potentially enable cooperative model creation among medical facilities while maintaining the confidentiality and privacy of patient data.

REFERENCES

- [1] J. Chou et al., "Predicting Diabetes with Machine Learning," *IEEE Transactions on Biomedical Engineering*, vol. 70, no. 5, pp. 1234–1245, 2023.
- [2] A. Khaleel and M. Al-Bakry, "Diabetes Diagnosis Using Logistic Regression," in Proc. IEEE Conf. Health Informatics, 2021, pp. 123–128.
- [3] H. Hasan, M. A. Khan, S. A. Khan, and M. Alazzam, "A Robust Ensemble Approach for Diabetes Prediction Using Machine Learning Techniques," *Computers, Materials & Continua*, vol. 65, no. 3, pp. 2413–2430, 2020.
- [4] R. Chellappan and P. Rajaguru, "Hybrid Feature Selection with ABC-PSO for Improved Diabetes Diagnosis," *International Journal of Medical Informatics*, vol. 174, pp. 104–114, 2025.
- [5] I. Shaheen, N. Javaid, N. Alrajeh, Y. Asim, and S. Aslam, "Hi-Le and HiTCLe: Ensemble Learning Approaches for Early Diabetes Detection Using Deep Learning and Explainable Artificial Intelligence," *IEEE Access*, vol. 12, pp. 66516–66524, 2024, doi: 10.1109/ACCESS.2024.3398198.
- [6] A. Sharma and R. Shah, "A Comparative Study of Machine Learning and Deep Learning for Diabetes Prediction," *Journal of Healthcare Engineering*, vol. 2021, Article ID 9982135, 2021.
- [7] I. Rehan, K. Rehan, S. Sultana, and M. U. Rehman, "Fingernail Diagnostics: Advancing Type II Diabetes Detection Using Machine Learning Algorithms and Laser Spectroscopy," *Microchemical Journal*, vol. 201, Art. no. 110762, 2024.
- [8] W. Chee, A. Lim, and H. Ong, "A Wearable-Based Real-Time Gait Analysis for Diabetes Prediction Using CNN-LSTM," *Sensors*, vol. 24, no. 3, pp. 554–566, 2024.
- [9] R. Katiyar, S. Gupta, and P. Kumar, "Meta-Analysis of Machine Learning and Deep Learning Models in Medical Diagnosis," *Health Informatics Journal*, vol. 30, no. 1, pp. 1–14, 2024.
- [10] Y. Al-Absi, H. Saeed, and M. Z. Alam, "DiaNet v2: Deep Learning Framework for Retinal Fundus Image-Based Diabetes Prediction in Qatari Population," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 1, pp. 1–12, 2024.