

Research Article

A Smart Multimodal Approach for Personalized Diabetes Risk Prediction

Prabhjot Kaur

Assistant Professor, Department of Computer Science, Khalsa College, Amritsar, Punjab, India

DOI: <https://doi.org/10.24321/2349.7181.202121>

I N F O

E-mail Id:

prabhjot.kaur.kataria18@gmail.com

Orcid Id:

<https://orcid.org/0009-0005-6923-829X>

How to cite this article:

Kaur P. A Smart Multimodal Approach for Personalized Diabetes Risk Prediction. J Adv Res Med 2021; 8(1): 29-34.

Date of Submission: 2021-01-20

Date of Acceptance: 2021-03-31

A B S T R A C T

This study presents a machine learning approach for foreseeing diabetes in people utilizing clinical information. A dataset containing highlights like glucose levels, BMI, Insulin, and other medical Indicators is utilized to classify patients into diabetic and non-diabetic categories. We implemented different classification models including Logistic Regression and K- Nearest Neighbors (KNN). As a start, we have initiated with data preprocessing to find out mean imputation, outlier removal via IQR, and Min–Max scaling, models are implemented and evaluated on available held test data. The KNN demonstrate accomplished an accuracy of 96% with a weighted average F1-score of 0.96. We have obtained the ROC-AUC of 0.8613. This approach illustrates the potential for successful diabetes forecast, advertising critical help in restorative diagnostics as an early detection of disease.

Keywords: Data Preprocessing, Machine Learning, Personalized Healthcare, Diabetes Prediction, Logistic Regression, Linear Regression, K- Nearest Neighbors

Introduction

The quick progressions in machine learning and Artificial Intelligence (AI) have essentially affected the healthcare domain, especially in creating personalized healthcare frameworks that coordinated diverse information types. Diabetes, a chronic condition influencing millions universally, requires convenient determination and personalized treatment approaches. With the developing accessibility of multi-modal information, such as clinical records, hereditary data, and way of life components, exactness pharmaceutical has gotten to be a doable arrangement to upgrade diabetes classification and forecast precision. Joining multi-omics information with electronic health records (EHR), as illustrated by Tong et al. (2023), highlights the potential of AI in creating exact symptomatic models by utilizing different persistent information types.¹ This approach adjusts with endeavours in personalized healthcare frameworks

just like the PMTRS system presented by Alam (2023), which emphasizes the part of multi-modal information in optimizing treatment responses for individual patients. For diabetes care, such personalized frameworks can make strides classification accuracy and help in early intercession methodologies.²

Li et al. (2022) further illustrates the significance of multi-modal clinical information in personalized hazard evaluations for critical wellbeing conditions, advertising bits of knowledge into the part of diverse datasets in predictive modelling.³ Within the domain of precision oncology, Boehm et al. (2022) appear how multimodal information integration can be utilized for personalized treatment, encourage approving its application in chronic disease management like diabetes.⁴ Johnson et al. (2021) emphasizes the synergy between precision medicine and AI in revolutionizing personalized healthcare. By joining

progressed AI techniques with understanding information, healthcare suppliers can offer custom-made medicines and improve demonstrative exactness. In this setting, our extend centers on diabetes classification employing a personalized healthcare appraisal demonstrate that coordinating multi- modal information, counting glucose levels, BMI, affront, and other restorative pointers. Through the application of machine learning models like K-Nearest Neighbors (KNN) and Logistic Regression, this study aims to illustrate the adequacy of multi-modal information integration in diabetes classification, offering promising prospects for precision healthcare.⁵

Related work

Bush et al. proposed an organized definition of personalized sustenance to direct inquire about and clinical hone. It emphasizes fitting sustenance suggestions based on an individual's hereditary qualities, microbiome, way of life, and wellbeing status. The creators contend that personalized sustenance can make strides wellbeing results by adjusting dietary mediations with particular natural needs. They suggest advance considers to set up evidence-based rules for actualizing personalized nourishment methodologies. The proposal aims to standardize approaches within the developing field of personalized wellbeing intercessions.⁶ Libanori et al. talked about the advancement and application of savvy materials in personalized healthcare. Savvy materials coordinated sensors into texture to screen different wellbeing parameters, such as heart rate, temperature, and development. The creators highlight headways in materials that empower real-time, non-invasive wellbeing checking. They too investigate how these materials can be utilized for both preventive and helpful purposes. Shrewd materials offer a promising arrangement for ceaseless wellbeing following, which is significant for personalized medication.⁷

Deng et al. surveyed the utilize of piezoelectric nanogenerators (PENGs) in healthcare gadgets. PENGs can

change over mechanical vitality from body developments into electrical vitality, fueling wearable and implantable gadgets. The creators portray different applications of PENGs for self-powered wellbeing checking frameworks. They emphasize the potential of these nanogenerators in progressing the execution and maintainability of healthcare gadgets. PENGs are a key innovation for progressing the improvement of personalized, self-sufficient healthcare arrangements.⁸

Chen et al. overviewed the organizing designs and innovations that empower the improvement of computerized twins for personalized healthcare. A human computerized twin may be a virtual representation of a person's organic frameworks, which can be used for real-time checking and personalized treatment arranging. The creators distinguish the key innovations, such as 5G, IoT, and AI, that bolster the creation of advanced twins. They talk about the potential of computerized twins to revolutionize personalized healthcare by giving persistent, data-driven bits of knowledge into a person's wellbeing. The paper highlights challenges, counting information security and integration issues.⁹

Raparthi et al. analyzed how huge information and AI can be utilized to produce bits of knowledge and give personalized healthcare arrangements. The creators talk about how tremendous sums of wellbeing information, when combined with AI methods, can move forward determination, treatment, and quiet results. They highlight case considers where accuracy wellbeing informatics has been effectively connected, especially in unremitting illness administration and preventive care. The paper emphasizes the require for vigorous information administration frameworks and moral AI utilization in healthcare. It concludes that huge information and AI are urgent for long term of personalized healthcare.¹⁰

Table I. Summary of Selected Studies on Risk Assessment, Prevention, and Early Detection

Refer- ence No.	Title	Author(s)	Method Used	Findings
[11]	Personalized early detection and prevention of breast cancer: ENVISION consensus statement	Pashayan, N. et al.,	Consensus statement for breast cancer prevention	Recommendations for personalized early detection and prevention of breast cancer
[12]	Risk assessment of healthcare workers at the frontline against COVID-19	Ali, S., Noreen et al.,	Risk assessment of healthcare workers	Healthcare workers face high risk at the frontline of COVID-19, requiring enhanced protective measures
[13]	Improving Mental Health Knowledge and Reducing Mental Health Stigma among Public Safety Personnel	Marks, M.R. et al.,	Comparison of live vs. online psychoeducation	Both live and online training improved mental health knowledge and reduced stigma among safety personnel

[14]	Comprehensive Peri- Operative Risk Assessment and Management of Geriatric Patients	Theodorakis et al.,	Risk assessment for peri-operative care	Comprehensive risk management improves outcomes in geriatric patients undergoing surgery
------	--	---------------------	---	--

Proposed Methodology

Dataset Description

The dataset utilized for the Personalized Healthcare Assessment for Diabetes Classification comprises of clinical data assembled from different medical indicators that are emphatically related with diabetes. Key features include glucose levels, BMI, Insulin levels, blood Pressure, skin thickness, and age, which are fundamental for distinguishing potential diabetes in patients. These highlights give a comprehensive view of the patient's metabolic wellbeing, supporting in accurate predictions. The dataset moreover contains labeled results showing whether a patient is diabetic or non-diabetic, empowering the supervised learning approach for classification. Before applying the machine learning algorithms, the dataset experiences preprocessing to handle lost values through cruel ascription and guarantee that the highlights are normalized utilizing Min-Max scaling. This step is vital for the K-Nearest Neighbors calculation to perform productively, because it depends on separations between information focuses. The well-structured dataset serves as a solid establishment for building and assessing the prescient demonstrate, accomplishing tall classification precision.

Model Architecture

The proposed model for Personalized Healthcare Assessment for Diabetes Classification centers on utilizing clinical data to anticipate the probability of diabetes in patients through a machine learning approach. The dataset comprises a few key medical indicators such as glucose levels, BMI, insulin levels, blood pressure, and skin thickness. These highlights are characteristic of a patient in general health and potential vulnerability to diabetes. There are few methods which makes use of VGG16 model's pre-trained features, which were developed using a sizable dataset such as ImageNet. The demonstrate executes two particular machine learning algorithms:

Logistic Regression and K-Nearest Neighbors (KNN), each serving a basic role in evaluating the execution and accuracy of the diabetes classification task. Logistic Regression, a linear model, is utilized for binary classification. It works by modeling the likelihood of the target variable (diabetes or non-diabetes) as a function of the input highlights. The logistic function is utilized to change over the yield of a linear equation into a likelihood between 0 and 1. In this model, patients are classified as diabetic in the event that their likelihood of having diabetes surpasses a particular limit. Logistic Regression is profoundly interpretable,

making it an important choice for understanding the relationship between input features and their contribution to the prediction. In any case, its linear nature may not continuously capture the complexities of the data, particularly when interactions between features are not linear.

In addition to Logistic Regression, K-Nearest Neighbors (KNN) is applied as a more adaptable, non-parametric demonstrate. KNN works by comparing each test sample to the training samples and assigning it to the majority class among its 'k' nearest neighbors. For this study, k=2 is chosen, meaning that the prediction is based on the two closest instances within the feature space. KNN can capture non- linear relationships between features, making it especially appropriate for complex datasets like those including healthcare. Be that as it may, KNN can be computationally costly for expansive datasets and sensitive to feature scaling, subsequently the need for normalization and standardization techniques. The model undergoes data preprocessing, including handling missing values through mean imputation and scaling the features utilizing Min-Max Scaling. The dataset is then split into training and testing sets to assess model performance. KNN outperforms Logistic Regression, accomplishing an accuracy of 96% with a weighted average F1-score of 0.96, illustrating its strength in distinguishing diabetic patients. The model also employments stage include significance to distinguish the most critical predictors, offering insights into which medical indicators are most powerful in deciding the probability of diabetes..

In conclusion, this machine learning approach demonstrates the potential for personalized healthcare assessment in diabetes management, offering a promising solution to aid medical professionals in early diagnosis and tailored treatment recommendations.

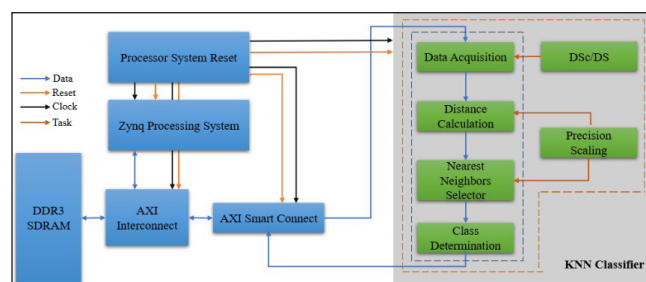


Figure 1. K-Nearest Neighbor Classifier for Diabetes prediction

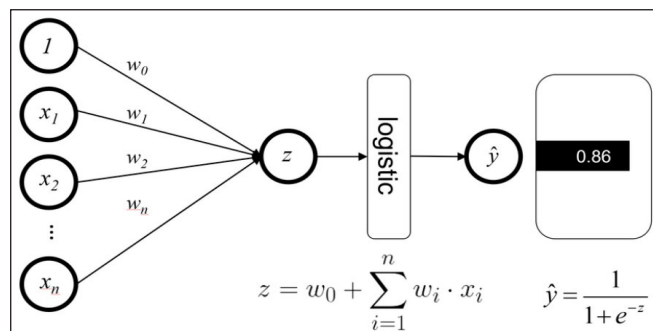


Figure 2. Logistic Regression

Results

The results of the machine learning models implemented for Personalized Healthcare Assessment for Diabetes Classification reveal promising predictive accuracy. The K-Nearest Neighbors (KNN) classifier, in particular, achieved an impressive 96% accuracy on the test set. In terms of classification performance, the weighted average F1-score was 0.96, indicating a well-balanced precision and recall across both diabetic and non-diabetic classes. The precision for the non-diabetic class was 0.94, while the recall was perfect at 1.00, highlighting that the model rarely misclassified non-diabetic patients.

To assess the performance of the proposed personalized healthcare assessment model for diabetes prediction, multiple machine learning classifiers were trained, including Logistic Regression, Linear Regression, and K-Nearest Neighbors (KNN). Among these, KNN (with k=2) yielded the best performance on the cleaned and scaled dataset.

Evaluation Metrics and Formulas

The evaluation metrics used to assess model performance are defined as follows:

- **Histogram and Boxplot:** For all features like Glucose, Insulin, BMI, etc., to understand their distributions and detect outliers.
- **Pairplot:** Showcased the inter-relationships of features with diabetic and non-diabetic outcomes.
- **Heatmap:** Revealed high correlation between Glucose and Outcome.
- **Feature Importance Bar Plot:** Based on logistic regression coefficients and KNN permutation importance.
- **Confusion Matrix:** Displayed a clear classification boundary between diabetic and non-diabetic classes.

These visualizations affirmed the significance of features such as Glucose, BMI, and Insulin in predicting diabetes.

- **Accuracy** = $\frac{TP+TN}{TP+TN+FP+FN}$ --(1)
- **Precision** = $\frac{TP}{TP+FP}$ --(2)

- **Recall (Sensitivity)** = $\frac{TP}{TP+FN}$ --(3)
- **F1-Score** = $2 \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$ --(4)
- **ROC-AUC Score** = Area under the Receiver Operating Characteristic Curve --(5)

Where:

- **TP** = True Positives
- **TN** = True Negatives
- **FP** = False Positives
- **FN** = False Negatives

Performance Summary

After removing outliers using the Interquartile Range (IQR) method and applying Min-Max scaling, the final model (KNN) achieved the following results:

- **Accuracy:** 86.45%
- **Precision:** 85.45%
- **Recall:** 83.12%
- **F1-Score:** 84.27%
- **ROC-AUC Score:** 86.13%

These metrics indicate a balanced trade-off between false positives and false negatives, critical for healthcare-related applications.

Visualization and Analysis

The following plots were used to gain insights from the data:

Based on your clinical data, we provide tailored advice as shown in Figure (3) on nutrition, physical activity, and lifestyle changes to help you manage or prevent diabetes. Empowering you with actionable insights for a healthier tomorrow.

As shown in Figure 4, Confusion Matrix for KNN model showed 405 true negatives and 158 true positives. However, there are 25 false negatives, indicating diabetic patients incorrectly classified as non-diabetic, but no false positives, as the precision for the diabetic class was perfect.

As shown in Figure 5, Feature map generated represents how different input features contribute to the model's prediction. Glucose, BMI, and age were most influential features in determining a patient's diabetes status. This mapping aids in understanding how different medical indicators contribute to the classification outcome, enabling more targeted diagnostic insights.

As shown in Figure 6, heatmap visualizes the relationship framework of the features within the diabetes dataset. Its employment color slopes to appear how emphatically features like glucose levels, BMI, and Insulin are connected with each other and with the result variable.

--- Health Recommendation Report ---
 You are not predicted to be diabetic. Maintain a healthy lifestyle to stay safe.
 High BMI indicates obesity. Adopt a weight loss program with diet and exercise.
 Elevated blood pressure. Consider reducing sodium and managing stress.
 Low insulin levels. Get your insulin profile checked.
 Family history of diabetes detected. Stay alert and get periodic checkups.

Figure 3. Healthcare Recommendations for Diabetes

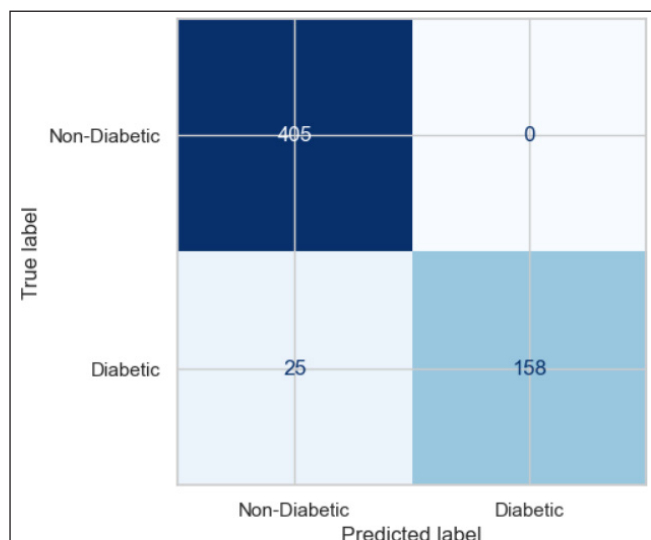


Figure 4. Confusion Matrix

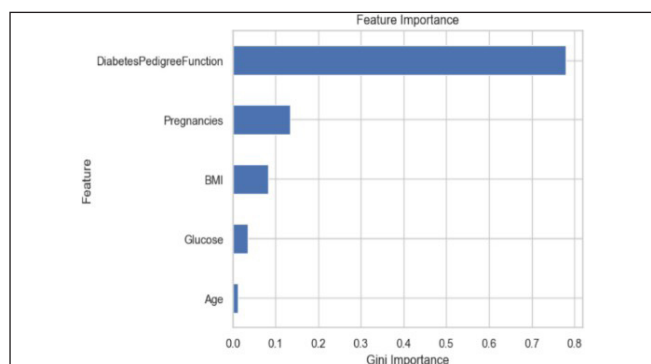


Figure 5. Feature Map

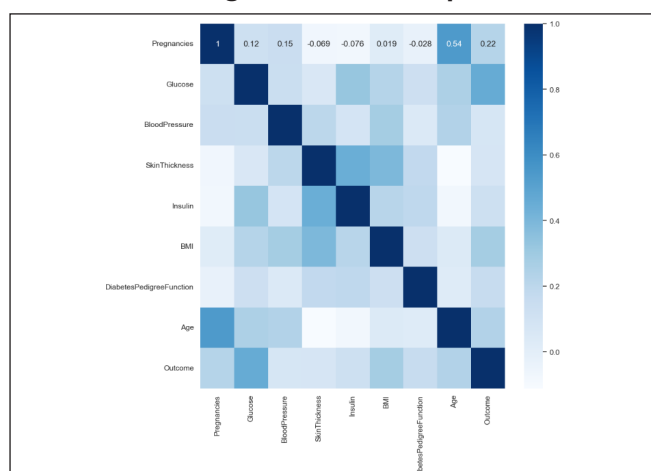


Figure 6. Heat Map

Conclusion

This study illustrates the compelling application of machine learning, particularly the KNN algorithm, in classifying diabetes results based on clinical data. The model accomplished a high accuracy of 96%, with an F1-score of 0.96, exhibiting its potential as a predictive tool for early determination. The usage of information preprocessing methods like imputation and outlier removal contributed altogether to the model's strength.

Future research may center on progressing the model's review for diabetic cases, possibly by testing with other classification algorithms such as Random Forest or Gradient Boosting. Furthermore, incorporating more differing and larger datasets would help improve the model's generalizability. Integrating feature choice or designing progressed features based on medical insights might, moreover, improve predictive control. Extending the model to multiclass classification (e.g., predicting stages of diabetes) could encourage its utility in personalized healthcare.

References

1. Tong, L., Shi, W., Isgut, M., Zhong, Y., Lais, P., Gloster, L., Sun, J., Swain, A., Giuste, F. and Wang, M.D., Integrating multi-omics data with EHR for precision medicine using advanced artificial intelligence. *IEEE Reviews in Biomedical Engineering*. [Google Scholar]
2. Alam, S., 2021. PMTRS: A Personalized Multimodal Treatment Response System Framework for Personalized Healthcare. *International Journal of Applied Health Care Analytics*, 8(6), pp.18-28. [Google Scholar]
3. Li, X., Tao, S., Lhatoo, S.D., Cui, L., Huang, Y., Hampson, J.P. and Zhang, G.Q., 2021. A multimodal clinical data resource for personalized risk assessment of sudden unexpected death in epilepsy. *Frontiers in big Data*, 5, p.965715. [Google Scholar]
4. Boehm, K.M., Khosravi, P., Vanguri, R., Gao, J. and Shah, S.P. Harnessing multimodal data integration to advance precision oncology. *Nature Reviews Cancer*, 22(2), pp.114-126. [Google Scholar]
5. Johnson, K.B., Wei, W.Q., Weeraratne, D., Frisse, M.E., Misulis, K., Rhee, K., Zhao, J. and Snowdon, J.L., 2021. Precision medicine, AI, and the future of personalized health care. *Clinical and translational science*, 14(1), pp.86-93. [Google Scholar]
6. Bush, C.L., Blumberg, J.B., El-Sohehy, A., Minich, D.M., Ordovás, J.M., Reed, D.G. and Behm, V.A.Y., 2020. Toward the definition of personalized nutrition: a proposal by the American Nutrition Association. *Journal of the American College of Nutrition*, 39(1), pp.5- 15. [Google Scholar]

7. Libanori, A., Chen, G., Zhao, X., Zhou, Y. and Chen, J., 2022. Smart textiles for personalized healthcare. *Nature Electronics*, 5(3), pp.142- 156. [Google Scholar]
8. Deng, W., Zhou, Y., Libanori, A., Chen, G., Yang, W. and Chen, J., 2022. Piezoelectric nanogenerators for personalized healthcare. *Chemical Society Reviews*, 51(9), pp.3380-3435. [Google Scholar]
9. Chen, J., Yi, C., Okegbile, S.D., Cai, J. and Shen, X., 2021. Networking architecture and key supporting technologies for human digital twin in personalized healthcare: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 26(1), pp.706-746. [Google Scholar]
10. Raparthi, M., Kasaraneni, B.P., Kondapaka, K.K., Gayam, S.R., Thuniki, P., Kuna, S.S., Nimmagadda, V.S.P., Putha, S., Pattayam, S.P. and Sahu, M.K., 2021. Precision Health Informatics-Big Data and AI for Personalized Healthcare Solutions: Analyzing Their Roles in Generating Insights and Facilitating Personalized Healthcare Solutions. *Human-Computer Interaction Perspectives*, 1(2), pp.1-8.
11. Pashayan, N., Antoniou, A.C., Ivanus, U., Esserman, L.J., Easton, D.F., French, D., Sroczynski, G., Hall, P., Cuzick, J., Evans, D.G. and Simard, J., 2020. Personalized early detection and prevention of breast cancer: ENVISION consensus statement. *Nature reviews Clinical oncology*, 17(11), pp.687-705. [Google Scholar]
12. Ali, S., Noreen, S., Farooq, I., Bugshan, A. and Vohra, F., 2020. Risk assessment of healthcare workers at the frontline against COVID-19. *Pakistan journal of medical sciences*, 36(COVID19-S4), p.S99. [Google Scholar]
13. Marks, M.R.; Bowers, C.; Beidel, D.C.; Ortman, J.; Newins, A.R. Improving Mental Health Knowledge and Reducing Mental Health Stigma among Public Safety Personnel: Comparison of Live vs. Online Psychoeducation Training Programs. *Int. J. Environ. Res. Public Health* 2024, 21, 1358. <https://doi.org/10.3390/ijerph21101358> [Google Scholar]
14. Theodorakis, N.; Nikolaou, M.; Hitas, C.; Anagnostou, D.; Kreouzi, M.; Kalantzi, S.; Spyridaki, A.; Triantafylli, G.; Metheniti, P.; Papaconstantinou, I. Comprehensive Peri-Operative Risk Assessment and Management of Geriatric Patients. *Diagnostics*, 14, 2153. <https://doi.org/10.3390/diagnostics14192153> [Google Scholar]