

RESEARCH ARTICLE

Unveiling diabetes onset: Optimized XGBoost with Bayesian optimization for enhanced prediction

Muhammad Rizwan Khurshid¹, Sadaf Manzoor^{1*}, Touseef Sadiq^{2*}, Lal Hussain^{3,4*}, Mohammed Shahbaz Khan⁵, Ashit Kumar Dutta⁶

1 Department of Statistics, Islamia University College, Peshawar, Khyber Pakhtunkhwa, Pakistan, **2** Centre for Artificial Intelligence Research (CAIR), Department of Information and Communication Technology, University of Agder, Kristiansand, Grimstad, Norway, **3** Department of Computer Science & IT, Neelum Campus, The University of Azad Jammu and Kashmir, Athmuqam, Azad Kashmir, Pakistan, **4** Department of Computer Science & IT, King Abdullah Campus, The University of Azad Jammu and Kashmir, Muzaffarabad, Azad Kashmir, Pakistan, **5** Children's National Hospital, Washington, DC, United States of America, **6** Department of Computer Science and Information Systems, College of Applied Sciences, AlMaarefa University, Ad Diriyah, Riyadh, Kingdom of Saudi Arabia

* touseef.sadiq@uia.no (TS); lal_hussain2008@live.com (LH); sadaf@icp.edu.pk (SM)



OPEN ACCESS

Citation: Khurshid MR, Manzoor S, Sadiq T, Hussain L, Khan MS, Dutta AK (2025) Unveiling diabetes onset: Optimized XGBoost with Bayesian optimization for enhanced prediction. PLoS ONE 20(1): e0310218. <https://doi.org/10.1371/journal.pone.0310218>

Editor: Hashim Ali, Nazarbayev University, KAZAKHSTAN

Received: July 8, 2024

Accepted: August 27, 2024

Published: January 24, 2025

Copyright: © 2025 Khurshid et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The publicly available Kaggle dataset (<https://www.kaggle.com/datasets/mathchi/diabetes-data-set>) on diabetes prediction offers a valuable resource for both healthcare professionals and researchers. Consent for publication.

Funding: Ashit Kumar Dutta would like to express sincere gratitude to AlMaarefa University, Riyadh, Saudi Arabia, for providing funding to conduct this research.

Abstract

Diabetes, a chronic condition affecting millions worldwide, necessitates early intervention to prevent severe complications. While accurately predicting diabetes onset or progression remains challenging due to complex and imbalanced datasets, recent advancements in machine learning offer potential solutions. Traditional prediction models, often limited by default parameters, have been superseded by more sophisticated approaches. Leveraging Bayesian optimization to fine-tune XGBoost, researchers can harness the power of complex data analysis to improve predictive accuracy. By identifying key factors influencing diabetes risk, personalized prevention strategies can be developed, ultimately enhancing patient outcomes. Successful implementation requires meticulous data management, stringent ethical considerations, and seamless integration into healthcare systems. This study focused on optimizing the hyperparameters of an XGBoost ensemble machine learning model using Bayesian optimization. Compared to grid search XGBoost (accuracy: 97.24%, F1-score: 95.72%, MCC: 81.02%), the XGBoost with Bayesian optimization achieved slightly improved performance (accuracy: 97.26%, F1-score: 95.72%, MCC: 81.18%). Although the improvements observed in this study are modest, the optimized XGBoost model with Bayesian optimization represents a promising step towards revolutionizing diabetes prevention and treatment. This approach holds significant potential to improve outcomes for individuals at risk of developing diabetes.

1. Introduction

Diabetes is a chronic condition characterized by high blood sugar levels. It primarily manifests as Type 1 (immune system attacks insulin-producing cells) or Type 2 (insulin resistance or