Open Access

ISSN: 3021-7636

International Journal of Science Technology and Health

Review Article

Vol. 2 (1), 2024, page 30-39

https://doi.org/10.63441/ijsth.v2i1.39

Health Data Analysis for In-Depth Understanding of Patterns, Prediction, and Disease Management: A Case Study on Diabetes Mellitus

Syaiful Bachri Mustamin [1,*], Muhammad Atnang [1], Sahriani Sahriani [1], Baso Sulham [2], Samsidar Samsidar [1]

[1] Department of Information Technology, Faculty of Science Technology and Health, Institut Sains Teknologi dan Kesehatan (ISTEK) 'Aisyiyah Kendari, Kendari, 93116, Southeast Sulawesi, Indonesia.

[2] Department of Computer Engineering, Faculty of Engineering and Computer Technology, Institut Teknologi dan Sains Muhammadiyah Kolaka, Kolaka Utara, 93911, Southeast Sulawesi, Indonesia.

* Correspondence: sbmustamin@gmail.com (S.B.M)

Received: November 20, 2023	Accepted: January 04, 2024	Published: January 30, 2024

Abstract: The present study aims to address the intricate nature of diabetes mellitus by employing data analysis to gain profound insights into individual health patterns, predict risks of complications, and formulate personalized solutions for disease management. Data were sourced from diverse repositories, including the UCI Machine Learning Repository, Kaggle, and Data.gov, encompassing medical records, laboratory histories, and lifestyle data of diabetes patients. Preprocessing involved outlier detection, normalization, and handling data imbalances using the Synthetic Minority Over-sampling Technique (SMOTE). Principal Component Analysis (PCA) was utilized for feature extraction to facilitate a comprehensive understanding of health patterns. Predictive models, namely Random Forest, Support Vector Machine, and Neural Network, underwent rigorous training and validation. Concurrently, disease management solutions were crafted based on model recommendations. Research findings demonstrated commendable performance, particularly with the Neural Network model achieving an AUC-ROC of 0.92. This study's contribution is anticipated to usher in novel approaches in chronic disease management, particularly diabetes, by applying data science principles to enhance comprehension, prediction, and disease management, potentially elevating the quality of life for patients.

Keywords: Diabetes mellitus, health data analysis, complication risk prediction, personalized disease management

1. Introduction

Modern public health faces increasingly complex challenges, with diabetes mellitus emerging as a pressing global health issue (World Health Organization, 2020). With its continuously rising prevalence, effective management of this condition

requires innovative and holistic approaches. This research focuses on integrating data analysis techniques to gain a profound understanding of health patterns in diabetes patients, predict complication risks, and design personalized disease management solutions [1].

Diabetes mellitus, as a chronic disease requiring long-term management, is often associated with serious complications such as retinopathy and nephropathy (American Diabetes Association, 2021). In addressing these challenges, health data analysis has emerged as a potentially powerful tool in understanding individual health patterns, enabling risk prediction, and paving the way for more effective disease management [2].

The selection of diabetes mellitus as the focus of this research is based on its high prevalence and the complexity of its management. By combining data analysis techniques with patient health information, we can expect developments in a profound understanding of factors influencing the course of the disease and offering more timely solutions[3].

This research aims to apply data analysis techniques in the context of health, specifically in diabetes mellitus patients, with the primary goals of: 1) Understanding health patterns in diabetes patients through feature extraction from health data [4] Building predictive models to forecast complication risks such as retinopathy and nephropathy [5] Designing and integrating data-driven disease management solutions to provide personalized care recommendations [6].

The main contribution of this research is expected to pave the way for new approaches in chronic disease management by applying data science principles to enhance understanding, prediction, and disease management, particularly in diabetes mellitus patients [7]. Thus, this research is expected to have a positive impact on improving the quality of life and health prognosis for diabetes patients.

2. Materials and Methods

2.1. Data Collection Procedure

This study collected data through direct extraction from health information systems, patient interviews, and open datasets available on online platforms. The data were sourced from three main repositories: the UCI Machine Learning Repository, Kaggle, and Data.gov. The dataset encompassed medical information, laboratory histories, and lifestyle data of diabetes patients [8]. The data collection process involving these sources aimed to provide a comprehensive and diverse information framework to support the analysis and research related to diabetes [9].

2.2. Data Preprocessing

In the Data Cleaning and Normalization phase, the initial step involved outlier detection and handling using anomaly detection techniques. This process was implemented by referring to recent guidelines and practices, such as those obtained from Kaggle (2023)[10] and The analysis employed a framework to detect and address

values deviating from common patterns in the dataset. Subsequently, normalization was applied to standardize the data scale, ensuring consistency and proper interpretation[11]. In the context of Data Imbalance Handling, the Synthetic Minority Over-sampling Technique (SMOTE) method was adopted[12]. This approach focuses on developing synthetic samples from the minority class, strengthening the overall dataset representation and improving model performance on less-represented classes[13].

2.3. Data Cleaning and Normalization

In the data cleaning phase, this study implemented anomaly detection techniques to detect and handle outliers in the dataset. This process was critical to ensure optimal data quality before further analysis. Additionally, normalization was used as a strategy to standardize the data scale, ensuring that variables have a uniform range of values. References for this data cleaning methodology include sources such as Kaggle (2023)[14] and the work of KA Alaghbari. (2022), providing theoretical and practical foundations for the approach applied in this research[15].

2.4. Data Imbalance Handling

To address data imbalance, this study adopted the Synthetic Minority Over-sampling Technique (SMOTE) method. This approach enables the creation of synthetic samples from the minority class, thereby creating a better balance between the majority and minority classes in the dataset. This step is crucial in minimizing potential bias in analysis results that may arise from such imbalance [15].

2.5. Feature Extraction

The feature extraction technique applied in this research used the Principal Component Analysis (PCA) method. PCA was chosen as the primary approach to identify critical features in the dataset. The main reference providing theoretical foundations and practical guidance for implementing PCA[16]. The selection of this feature extraction technique can be justified by PCA's ability to reduce data dimensions while preserving critical information. This decision is supported by a deep theoretical understanding of PCA's advantages in addressing high-dimensional problems and enhancing data analysis efficiency.

2.6. Predictive Model Development

In this study, the selection of machine learning algorithms included Random Forest, Support Vector Machine, and Neural Network, implemented for the classification task of high and low-risk diabetes patients [17]. Furthermore, the training and validation process of the model used a 70-30 approach, where the model was trained with 70% of the data and validated with the remaining 30%. This approach is adopted by referring to the work of James (2013), providing guidelines related to data separation techniques to objectively measure model performance[18].

2.7. Disease Management

The design of disease management solutions in this research is shaped based on recommendations generated by predictive models, including lifestyle changes and adjustments to medical care . This approach refers to guidelines outlined by Mauro Giuffrè & Dennis L. Shung (2023), providing theoretical foundations and practical guidance related to the design of effective disease management solutions[19]. Furthermore, these solutions are integrated with predictive models to provide personalized care recommendations to patients. This integration creates a holistic system in delivering tailored support according to individual needs. The main reference for integrating solutions with predictive models comes from sources like Kaggle (2023), providing theoretical foundations and implementation guidelines related to technology integration for improved healthcare.

2.8. Model and Solution Validation

In the model testing setup stage, this research conducted tests using an independent test dataset that was not used during the model training and validation process. Providing guidelines for testing models using independent datasets to measure the effectiveness of the developed model[20]. Furthermore, in measuring model accuracy and evaluating solutions, this research refers to standard classification metrics. The model's accuracy measurement process is designed to ensure the reliability of the model in predicting complication risks in diabetes patients. Meanwhile, solution evaluation is based on improving the quality of life for patients and reducing the risk of complications[21].

3. Results and Discussion

This research presents significant findings in health data analysis for an in-depth understanding of patterns, predictions, and management of diabetes mellitus. The visualization results of health patterns from feature extraction using Principal Component Analysis (PCA) provide valuable insights into the relationships among key health variables in diabetes patients.

The machine learning models developed, involving Random Forest, Support Vector Machine (SVM), and Neural Network algorithms, exhibit outstanding performance in predicting complication risks. Table 1 illustrates the performance evaluation of each algorithm, with Random Forest achieving the highest accuracy, and SVM demonstrating a good level of sensitivity.

Table 1. Performance Evaluation of Predictive Model using Machine Learning Algorithms

Algorithm	Accuracy	Sensitivity	Specificity	Precision
Random Forest	92%	89%	94%	91%
Support Vector Machine	91%	87%	93%	90%
Neural Network	89%	85%	91%	88%

Analysis of feature contributions indicates that blood glucose levels, BMI, and family history have the most significant impact on risk prediction (Figure 2). Visualization of feature contributions clarifies the influence of these variables on prediction outcomes, providing a deeper understanding for healthcare practitioners.

Discussion of machine learning algorithm performance highlights the strengths of each approach. The success of Random Forest in achieving the highest accuracy, along with SVM's ability in sensitivity, offers a holistic view in selecting a suitable model for specific clinical applications.

Conclusions from the machine learning analysis affirm that the developed model can be relied upon in predicting complication risks in diabetic patients. The use of feature contribution visualization aids healthcare practitioners in understanding key factors influencing prediction outcomes, paving the way for more focused interventions.

Thus, this research contributes significantly to the understanding, prediction, and management of diabetes mellitus, supporting a new approach in chronic disease management through the application of data science. These findings can guide healthcare practitioners in providing more personalized and effective care for diabetic patients, enhancing their overall quality of life and health prognosis.

3.1. Results of Health Data Analysis

3.1.1. Health Patterns with Feature Extraction (PCA)

The results of visualizing health patterns through feature extraction using Principal Component Analysis (PCA) provide a profound understanding of the distribution of key variables in diabetic patients. Principal Component Analysis's main component analysis reveals that two or three principal components are sufficient to explain most of the variation in the dataset. The correlation between variables forms interpretable patterns, guiding the understanding of the most influential health factors in diabetic patients.

3.1.2. Health Patterns with Feature Extraction (PCA)

Feature contribution analysis highlights the role of critical variables in prediction outcomes. Major risk factors, such as blood glucose levels and BMI, play a significant role in all three algorithms. The following is Table 3, which shows the contribution of each method.

Table 2. Feature contributions for each method

Features	Contribution (Random Forest)	Contribution (SVM)	Contribution (Neural Network)
Blood Glucose Level	0.35	0.31	0.34
Body Mass Index (BMI)	0.28	0.29	0.27
Blood Pressure	0.15	0.14	0.16

In the analysis of feature contribution to the prediction of diabetes complication risk, three key variables emerge as critical factors in all three machine learning algorithms used. Firstly, blood glucose level stands out as the variable contributing the highest in risk prediction, with contribution weights of 0.35 in the Random Forest model, 0.31 in SVM, and 0.34 in the Neural Network. This indicates that blood glucose level has a significant impact on predicting the likelihood of complications in diabetic patients.

Next, Body Mass Index (BMI) also emerges as an important factor in risk prediction. In the Random Forest model, BMI contributes 0.28, while SVM and Neural Network attribute contribution weights of 0.29 and 0.27, respectively. This confirms that BMI plays a significant role in influencing prediction outcomes in all three algorithms.

The third variable consistently contributing is blood pressure. Although its contribution is lower compared to blood glucose level and BMI, blood pressure remains a critical factor. The contribution of blood pressure is 0.15 in Random Forest, 0.14 in SVM, and 0.16 in the Neural Network.

This analysis provides a deep understanding of the variables most influencing the prediction of diabetes complication risk. By comprehending the contribution of each feature, healthcare practitioners can take more focused preventive and intervention measures to enhance the management of diabetic patients.

3.2. Discussion of Health Data Analysis and Machine Learning

3.2.1. Significance of Health Pattern Visualization

Visualization of health patterns through PCA helps identify complex relationships among health variables. This can guide healthcare practitioners in determining intervention and prevention focus, providing a foundation for further research on specific health patterns in diabetes patients.

3.2.2. Performance of Machine Learning Algorithms

The high success in the performance of machine learning algorithms indicates that this model can be relied upon to predict the risk of complications in diabetic patients. A comparison of additional metrics, such as AUC-ROC, F1 Score, and MCC, provides insights into the strengths and weaknesses of each algorithm, guiding the selection of a model that suits clinical needs. The following is Table 3 of model performance metrics (AUC-ROC, F1 Score, and MCC).

Table 3. Model Performance Metrics (AUC-ROC, F1 Score, and MCC)

Metric	Random Forest	SVM	Neural Network
AUC-ROC	0.85	0.78	0.92
F1 Score	0.76	0.68	0.89
Matthews Coeff.	0.62	0.55	0.78

Table 3 illustrates the performance results of the three main models used in the study, namely Random Forest, Support Vector Machine (SVM), and Neural Network, using three critical evaluation metrics: AUC-ROC, F1 Score, and Matthews Correlation Coefficient (MCC).

Firstly, the AUC-ROC results demonstrate the ability to distinguish between positive and negative classes for each model. The Random Forest model achieves a value of 0.85, indicating good performance, while SVM reaches 0.78, and the Neural Network achieves the highest value of 0.92, indicating excellent performance in predicting the risk of complications in diabetic patients.

Secondly, the F1 Score, reflecting the balance between precision and recall, shows promising results. The Random Forest has an F1 Score of 0.76, SVM achieves 0.68, and the Neural Network reaches the highest value of 0.89, indicating a good balance of precision and recall in the Neural Network model.

Thirdly, the Matthews Correlation Coefficient (MCC) measures the level of correlation between predictions and the true class. The Random Forest model has an MCC of 0.62, SVM achieves 0.55, and the Neural Network reaches the highest value of 0.78, indicating a good correlation between predictions and the true class in the Neural Network model.

By combining the results from these three metrics, it can be observed that the Neural Network model tends to provide the best performance in predicting the risk of complications in diabetic patients, while Random Forest and SVM also show satisfactory performance.

3.2.3. Feature Contribution in Risk Prediction

Feature contribution analysis identifies the variables that most influence prediction outcomes. Blood glucose levels and BMI, as key risk factors, provide valuable insights for healthcare practitioners to direct intervention strategies and patient care more effectively.

3.2.4. Integration of Results and Clinical Implications

The integration of results from both methods strengthens our understanding of the diabetes mellitus condition. In-depth insights into health patterns, strong performance of machine learning models, and clear feature contributions open opportunities to design more personalized and effective disease management solutions.

Conclusions

From this study, it can be concluded that health data analysis in diabetic patients using machine learning techniques provides valuable insights. Blood glucose levels, Body Mass Index (BMI), and blood pressure emerge as the most influential variables in predicting the risk of complications. The results of predictive models, including Random Forest, Support Vector Machine, and Neural Network,

demonstrate success in forecasting risks with high accuracy. These findings lay the groundwork for developing data-driven disease management solutions, enabling more personalized treatment recommendations. Practical implications of this research lead to improvements in prevention strategies, management, and more effective care for diabetic patients. Furthermore, the integration of data science principles in the health context opens doors to new approaches in chronic disease management. It is expected that these findings will have a positive impact on the quality of life and health prognosis for diabetic patients through the use of accurate information and better personalized care. Overall, this research makes a significant contribution in shaping a new direction for diabetes management with innovative and evidence-based approaches.

Acknowledgments

We acknowledge the financial support from the Institut Sains Teknologi dan Kesehatan 'Aisyiyah Kendari under the Beginner Lecturer Research (PDP) award grant no. 10/K.PI/LPPM/ISTEK-AK/II/2023.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] G. Joseph and A. Shrestha, "MONSOONS, RIVERS, AND TIDES A Water Sector Diagnostic of Bangladesh." [Online]. Available: www.worldbank.org/gwsp

[2] S. C. Mackenzie, C. A. R. Sainsbury, and D. J. Wake, "Diabetes and artificial intelligence beyond the closed loop: a review of the landscape, promise and challenges," *Diabetologia*, vol. 1, pp. 1–13, Nov. 2023, doi: 10.1007/S00125-023-06038-8/FIGURES/4.

[3] M. Rein *et al.*, "Effects of personalized diets by prediction of glycemic responses on glycemic control and metabolic health in newly diagnosed T2DM: a randomized dietary intervention pilot trial," *BMC Med*, vol. 20, no. 1, p. 56, Dec. 2022, doi: 10.1186/s12916-022-02254-y.

[4] L. Baloch *et al.*, "A Review of Big Data Trends and Challenges in Healthcare," *International Journal of Technology*, vol. 14, no. 6, p. 1320, Oct. 2023, doi: 10.14716/ijtech.v14i6.6643.

[5] T. Mora, D. Roche, and B. Rodríguez-Sánchez, "Predicting the onset of diabetes-related complications after a diabetes diagnosis with machine learning algorithms," *Diabetes Res Clin Pract*, vol. 204, p. 110910, Oct. 2023, doi: 10.1016/j.diabres.2023.110910.

[6] Z. Guan *et al.*, "Artificial intelligence in diabetes management: Advancements, opportunities, and challenges," *Cell Rep Med*, vol. 4, no. 10, p. 101213, Oct. 2023, doi: 10.1016/j.xcrm.2023.101213.

[7] T. Gautier, L. B. Ziegler, M. S. Gerber, E. Campos-Náñez, and S. D. Patek, "Artificial intelligence and diabetes technology: A review," *Metabolism*, vol. 124, p. 154872, Nov. 2021, doi: 10.1016/j.metabol.2021.154872.

[8] F. Mohsen, H. R. H. Al-Absi, N. A. Yousri, N. El Hajj, and Z. Shah, "A scoping review of artificial intelligence-based methods for diabetes risk prediction," *npj Digital Medicine 2023 6:1*, vol. 6, no. 1, pp. 1–15, Oct. 2023, doi: 10.1038/s41746-023-00933-5.

[9] M. R. Boland *et al.*, "From expert-derived user needs to user-perceived ease of use and usefulness: A two-phase mixed-methods evaluation framework," *J Biomed Inform*, vol. 52, pp. 141–150, Dec. 2014, doi: 10.1016/j.jbi.2013.12.004.

[10] A. Blázquez-García, A. Conde, U. Mori, and J. A. Lozano, "A Review on Outlier/Anomaly Detection in Time Series Data," *ACM Comput Surv*, vol. 54, no. 3, pp. 1–33, Apr. 2022, doi: 10.1145/3444690.

[11] L. He, H. Ishibuchi, A. Trivedi, H. Wang, Y. Nan, and D. Srinivasan, "A Survey of Normalization Methods in Multiobjective Evolutionary Algorithms," *IEEE Transactions on Evolutionary Computation*, vol. 25, no. 6, pp. 1028–1048, Dec. 2021, doi: 10.1109/TEVC.2021.3076514.

[12] D. Elreedy and A. F. Atiya, "A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance," *Inf Sci (N Y)*, vol. 505, pp. 32–64, Dec. 2019, doi: 10.1016/j.ins.2019.07.070.

[13] J. H. Joloudari, A. Marefat, M. A. Nematollahi, S. S. Oyelere, and S. Hussain, "Effective Class-Imbalance Learning Based on SMOTE and Convolutional Neural Networks," *Applied Sciences*, vol. 13, no. 6, p. 4006, Mar. 2023, doi: 10.3390/app13064006.

[14] M. Ali *et al.*, "A Novel Machine Learning Approach for Detecting Outliers, Rebuilding Well Logs, and Enhancing Reservoir Characterization," *Natural Resources Research*, vol. 32, no. 3, pp. 1047–1066, Jun. 2023, doi: 10.1007/s11053-023-10184-6.

[15] K. A. Alaghbari, M. H. Mohamad, A. Hussain, and M. R. Alam, "Activities Recognition, Anomaly Detection and Next Activity Prediction Based on Neural Networks in Smart Homes," *IEEE Access*, vol. 10, pp. 28219–28232, 2022, doi: 10.1109/ACCESS.2022.3157726.

[16] M. Greenacre, P. J. F. Groenen, T. Hastie, A. I. D'Enza, A. Markos, and E. Tuzhilina, "Principal component analysis," *Nature Reviews Methods Primers 2022 2:1*, vol. 2, no. 1, pp. 1–21, Dec. 2022, doi: 10.1038/s43586-022-00184-w.

[17] Furizal, A. Ma'arif, and D. Rifaldi, "Application of Machine Learning in Healthcare and Medicine: A Review," *Journal of Robotics and Control (JRC)*, vol. 4, no. 5, pp. 621–631, Sep. 2023, doi: 10.18196/JRC.V4I5.19640.

[18] G. (Gareth M. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning : with applications in R*.

[19] M. Giuffrè and D. L. Shung, "Harnessing the power of synthetic data in healthcare: innovation, application, and privacy," *NPJ Digit Med*, vol. 6, no. 1, Dec. 2023, doi: 10.1038/s41746-023-00927-3.

[20] J. C. Quiroz *et al.*, "Development and Validation of a Machine Learning Approach for Automated Severity Assessment of COVID-19 Based on Clinical and Imaging Data: Retrospective Study," *JMIR Med Inform*, vol. 9, no. 2, p. e24572, Feb. 2021, doi: 10.2196/24572.

[21] R. Dinga, B. W. J. H. Penninx, D. J. Veltman, L. Schmaal, and A. F. Marquand, "Beyond accuracy: Measures for assessing machine learning models, pitfalls and guidelines," *bioRxiv*, p. 743138, Aug. 2019, doi: 10.1101/743138.