Sciences Methods and Technologies International Journal (SciMeTech)

ISSN: 3085-5284, journal homepage: www.scimetech.com

Improving Diabetes Prediction Using a Hybrid Model: Machine Learning and Rule-Based Logic

Chaimae LAHRIRE1, Anas ABOU EL KALAM1, Wissam ABBASS1

1Cadi Ayyad University, National School of Applied Sciences, Marrakech,Research Laboratory for Smart and Sunstainable Technologies,Marrakesh, Morocco

Email: C.lahrire.ced@uca.ac.ma, a.abouelkalam@uca.ac.ma, wissam.abbass@edu.uca.ac.ma

How to cite:

Chaimae LAHRIRE, Anas ABOU EL KALAM, Wissam ABBASS, "Improving Diabetes Prediction Using a Hybrid Model: Machine Learning and Rule-Based Logic", Sciences Methods and Technologies International Journal (SciMeTech), Vol 2, Issue 2, p 192-196

Keywords:

Diabetes prediction
Hybrid model
Machine learning
Rule-based system

Abstract

The objective of this paper is to propose a hybrid approach that combines machine learning techniques with predefined medical rules-based logic in order to improve prediction performance. The dataset used in this study concerning the pathology of diabetes is drawn from the Indian database Pima, in addition to several developed models namely: logistic regression, decision tree and random forest, are evaluated.

The results obtained show that the hybrid model is even more efficient in terms of patient identification, accuracy and memory retention. This proves the importance of combining data-driven methods with rules and medical knowledge to achieve increasingly reliable medical decision-making systems.

I. Introduction

According to the World Health Organization, diabetes remains a major chronic disease affecting millions of people worldwide. Its prevalence is continuously increasing, which justifies the importance of early detection to prevent any kind of severe complication [1].

The prediction performance of this pathology is revealed from machine learning techniques and databases such as the Pima Indians Diabetes Dataset [2]. However, the lack of interpretability still limits the use of these models in medical practice [3]. The hybrid approach proposed by this study will contribute to the improvement of prediction and interpretability.

II. Related work

Several studies have demonstrated high performance in classifying medical data for the prediction of diabetes disease, using machine learning models such as logistic regression, decision trees and random forests.

The use of supervised learning approaches and the Pima Indian dataset, commonly used to assess diabetes prediction, have yielded promising results [4]. Better use of data and better accuracy of predictions are achieved by applying these methods.

However, the main limitation presented by this approach is the lack of interpretability [5]. In order to improve the transparency of decisions in clinical practice, rules-based methods and medical knowledge have been proposed [6]. The combination of machine learning with logical rules was developed to improve both performance and interpretability.

III. Methodology

a. Dataset

The dataset on Indian Pima diabetes used in this study includes a range of medical indicators such as glucose level, body mass index (BMI), age and blood pressure. These variables are essential in predicting diabetes.

Data preprocessing is an important step to improve the model's performance; during this process, all missing values are processed and numerical values are standardized to avoid errors during learning using z-score normalization [7].

In order to make a consistent comparison with previous work, we adopted the Pima Indian dataset as a reference dataset for diabetes prediction. Despite its small size and limited population coverage, it remains suitable for the evaluation of the proposed hybrid framework. The main objective of this study is to validate the proposed hybrid framework rather than deploy it clinically in the real world.

b. Machine learning models

Several supervised learning algorithms were treated in this study, namely logistic regression, decision tree and random forest, in order to compare their performance in the prediction of diabetes.

Logistic regression is discussed in more detail because of its simplicity and interpretability. Relying on the sigmoid function makes it possible to estimate the probability of diabetes, which transforms values into probabilities between 0 and 1, and then compares them with a fixed threshold, generally equal to 0.5. In order to be able to make a decision whether the patient is classified as diabetic or not [8].

The probability of diabetes is estimated using the sigmoid function:

$$(y = 1 \mid X) = \frac{1}{1 + e^{-(wX + b)}}$$

- **P(y = 1 | X): probability that the patient is diabetic**
- **X: patient data (age, glucose, BMI, etc.)**
- **w: weights**
- **b: bias**
- **e: mathematical constant (~2.718)**

c. Hybrid Model (Machine Learning + Rule-Based Logic)

To improve the reliability of predictions, a hybrid model is proposed by combining machine learning predictions with rule-based logic.

The rule-based component is defined by well-known clinical risk factors such as high glucose levels, high body mass index (BMI), blood pressure and age, which are commonly associated with diabetes risk according to medical guidelines [9]. In this study, main rules are implemented defined according to medical thresholds to improve the prediction of the pathology:

• Glucose > 140 → risk of diabetes / BMI > 30 → risk of overweight/ Age > 45 → increased risk

d. Evaluation Metrics

The evaluation of the performance of methods in medical diagnosis is based on several metrics, namely accuracy; precision, recall and F1 score [10].

IV. Experimental setup

The study is conducted using the Python language in addition to machine learning libraries such as Scikit-learn and NumPy. This supervised learning approach follows the practices commonly used in recent work [11].

In order to evaluate the performance of the model, all data is divided into two parts: 80% of the data for training and 20% for testing [12].

Each of the models is trained on the training set by minimizing a loss function, which allows to measure the deviation between the real values and the predictions of the model in order to solve any binary classification problem.

The approach proposed in this study is based on combining the predictions of the machine learning model with a rule-based system derived from data and expert knowledge to improve the reliability of the predictions. Consider (x) the vector comprising the patient's characteristics, y_{ml} (x) $\in \{0, 1\}$ the machine learning model prediction, and

YR(x) $\in \{0,1\}$ the rule-based system that produces binary predictions. The final hybrid prediction is obtained using a

V (yR=1)

logical OR merges: Yfinal = $\begin{cases} 1 \; if \; (yML=1) \\ 0 \; otherwise \end{cases}$

Using models with default hyperparameter settings ensures fair comparison and reproducibility. The main purpose of this study is not to optimize performance, but to properly evaluate the proposed hybrid decision strategy. In addition, the realization of an ablation study will be used to evaluate three configurations: machine learning models only, rules-based system only, and the proposed hybrid approach. To measure the individual and combined contribution of each component to overall predictive performance. Finally, these models are evaluated on the test set according to several performance indicators such as precision, accuracy; recall and F1 score [13].

The reference of the dataset used comes from the UCI machine learning repository, a source for evaluating machine learning algorithms. Thus, the relevance of diabetes-related variables is endorsed by all World Health Organization medical guidelines.

V. Results and discussion

Table 1 Model performance comparison

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.747	0.638	0.673	0.655
Decision Tree	0.747	0.625	0.727	0.672
Random Forest	0.721	0.607	0.618	0.613
Rule-based Model	0.70	0.55	0.80	0.63
Hybrid Model	0.688	0.545	0.764	0.636

The results obtained show that the logistic regression and decision tree models achieve high accuracy (0.747), while the random forest model shows slightly lower performance.

The rule-based model obtained a high recall score (0.80), demonstrating its strong ability to identify diabetic patients, but its accuracy is low (0.55) indicating an increase in the number of false positive predictions. This behavior is predictable because rule-based methods use commonly accepted medical thresholds and favor sensitivity to specificity. However, the hybrid model offers the best ability to correctly identify diabetic patients with a recall rate (0.764). This aspect is still necessary in the medical field, because failing to detect a sick patient (false negative) can have serious consequences on his health.

This improvement in recall is justified by the integration of medical rules into the machine learning model. By combining expert knowledge with machine learning techniques, the hybrid model succeeds in detecting high-risk cases that could be ignored by purely data-driven approaches.

On the other hand, the hybrid model has lower accuracy and precision compared to stand-alone models. This can be explained by the inclusion of general medical rules, which can increase false positives and also through the use of logical OR fusion that allows the hybrid system to identify high-risk patients who might not be detected by autonomous datadriven models.

In short, although traditional machine learning models offer globally balanced performance, the hybrid approach focuses on the sensitivity (recall) and reliability of diabetes case detection, which makes the hybrid approach a promising solution for clinical decision support systems.

VI. CONCLUSION

The strategy proposed in this study, based on a combination of machine learning techniques and medical rules, proved to be an improvement in the detection of diabetic patients. Even if the accuracy is slightly lower than that of stand-alone models, we still get better sensitivity, which is essential in the medical field to reduce false negatives.

However, this study presents some limitations such as the use of a moderate size dataset and the use of default hyperparameters and general medical rules, which could be further optimized in future work.

In future work, it would be interesting to optimize medical rules and explore more advanced hybrid strategies, as well as to evaluate the model on broader datasets in order to improve its performance and facilitate its application in real clinical environments.

References

[1] World Health Organization. (2023). *Diabetes*.

[2] Dua, D., & Graff, C. (2020). *UCI machine learning repository*. University of California, Irvine.

[3] Samek, W., Wiegand, T., & Müller, K.-R. (2021). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.

[4] American Diabetes Association. (2023). Standards of care in diabetes—2023. *Diabetes Care*.

[5] Goodfellow, I., Bengio, Y., & Courville, A. (2022). *Deep learning*. MIT Press.

[6] Chollet, F. (2021). *Deep learning with Python* (2nd ed.). Manning Publications.

[7] Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras & TensorFlow* (3rd ed.). O'Reilly Media.

[8] Zhou, Z. (2021). *Machine learning*. Springer.

[9] Li, X., et al. (2022). Machine learning-based approaches for diabetes prediction: A systematic review. *Computers in Biology and Medicine*.

[10] Esteva, A., et al. (2021). A guide to deep learning in healthcare. *Nature Medicine*.

[11] Lundberg, S. M., et al. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*.

[12] Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence (XAI). *IEEE Access*.

[13] Ahmad, M., et al. (2021). Explainable AI in healthcare: Applications, challenges and future trends. *Artificial Intelligence in Medicine*.