International Science and Technology Journal
المجلة الدولية للعلوم والتقنية
Volume 38 العدد
Part 2 المجلد
المجلة الدولية للعلوم والتقنية
International Science and Technology Journal
ISTJ

Advancing diabetes disease prediction:
a comprehensive approach using machine learning and expert knowledge

http://www.doi.org/10.62341/istj-vol38-2-jr46

Received	**2026/05/22**	**تم استلام الورقة العلمية في**
Accepted	**2026/06/15**	**تم قبول الورقة العلمية في**
Published	**2026/06/17**	**تم نشر الورقة العلمية في**

Advancing diabetes disease prediction: a comprehensive approach using machine learning and expert knowledge

Jazya Moftah Ahmad Amshaher

Sirt University, Libya
jazamoftah@su.edu.ly

Abstract:
The swift expansion of healthcare data has presented significant hurdles for improving clinical outcomes. Identifying the relationships between patient characteristics, medical records, and disease outcomes using large volumes of Electronic Health Records (EHRs) is one of the challenges faced. Although several machine learning algorithms have been suggested to predict diseases, many existing methods are still not sufficiently accurate. The primary disadvantage of these methods is their emphasis on the prediction algorithm, which often overlooks crucial aspects like feature selection and weighting. Due to these shortcomings, predictive models may not be as effective.
To overcome these obstacles, this research seeks to develop an improved framework for predicting diseases by employing machine learning methods and domain knowledge. This paper investigates the utilization of K-Nearest Neighbor's (KNN) algorithm for predicting diseases. It will use domain expertise to help choose and weight relevant features from the large dataset of diabetes data and the resulting performance for accuracy was 72.73%, Precision 62.50%, Recall 55.56%, area under the curve (AUC) 78.31%, and F1-Score 58.82%.
Keywords: Diabetic Disease, Machine Learning, Disease Prediction, feature selection.

International Science and Technology Journal
المجلة الدولية للعلوم والتقنية
Volume 38 العدد
Part 2 المجلد
المجلة الدولية للعلوم والتقنية
International Science and Technology Journal
ISTJ

Advancing diabetes disease prediction:
a comprehensive approach using machine learning and expert knowledge

http://www.doi.org/10.62341/istj-vol38-2-jr46

طريقة شاملة باستخدام التعلم الآلي وخبرة المختصين

جازية مفتاح أحمد أمشهر
جامعة سرت – ليبيا
jazamoftah@su.edu.ly

الملخص:

مع الزيادة القوية في البيانات المتعلقة بالرعاية الصحية، هناك العديد من التحديات التي قد تعترض تحسن النتائج السريرية. من التحديات التي يشكل تحديًا كبيرًا هو تحديد العلاقات بين الخصائص المرضية والأمراض والعواقب السريرية من خلال تحليل كميات ضخمة من السجلات الصحية الإلكترونية. بالرغم من وجود العديد من خوارزميات التعلم الآلي المقترحة للأمراض، الكثير من الأساليب الموجودة حالياً لا تحقق درجة عالية من الدقة. النقطة الرئيسية لجميع هذه الأساليب هي أنها تشترك في الاهتمام بخوارزمية التنبؤ، مما يعني غالباً إهمال بعض الجوانب الهامة مثل اختيار الميزات وترجيحها. نتيجة لوجود هذه العيوب، يكون العديد من النماذج الخاصة بالتنبؤ ضعيف.

وتجسد هذه الورقة جهدنا في التغلب على هذه المعوقات بتطوير نموذج تنبؤ محسّن لمرض السكري باستخدام التعلم الآلي ومعرفة المجال. وتعالج الورقة الاستفادة من خوارزمية أقرب جار (KNN) لأمراض السكري. ستستفيد من الامتياز في مجال معرفة للمساعدة في اختيار. سيستخدم هذا النظام الخبرة في المجال للمساعدة في اختيار وترجيح الميزات ذات الصلة من مجموعة البيانات الكبيرة لبيانات مرض السكري، وكانت النتائج النهائية للدقة 72.73%، والدقة 62.50%، والاستدعاء 55.56%، والمساحة تحت المنحنى 78.31% (AUC)، ودرجة F1 58.82%.

الكلمات المفتاحية: مرض السكري، التعلم الآلي، التنبؤ بالأمراض، اختيار الميزات.

http://www.doi.org/10.62341/istj-vol38-2-jr46

Introduction

The technique of machine learning can be used as an effective means of discovering patterns in the data that are present. It has many uses ranging from healthcare to finance and business analytics, and it improves efficiency and resource management (Kolasa et al,2024). Machine learning has become increasingly important in healthcare, where it is widely applied to disease prediction , personalized medicine and medical image analysis (Alhumaidi et al,2025). Despite these advances, many existing disease prediction approaches still suffer from limitations in predictive accuracy. A major drawback of several studies is their primary focus on prediction algorithms while overlooking other influential factors such as feature selection, feature weighting, feature engineering, and dataset specific challenges, particularly class imbalance (Islam,2023). These factors can significantly affect the accuracy and generalization of predictive models.

Many machine learning models operate as "black-box" systems, making it difficult for healthcare professionals to understand the reasoning behind their predictions. Consequently, there is increasing interest in developing transparent machine learning models that can provide high predictive performance (Bomrah et al,2023).The proposed methodology seeks to integrate domain knowledge with machine learning algorithms to improve prediction accuracy, enhance model interpret ability. By combining expert knowledge with machine learning techniques, the study aims to develop a more reliable and effective framework for early diagnosis (Sarku et al,2023).The main contributions of this work are summarized as follows:

1. A promising Predictive Method: Through the integration of machine learning and domain knowledge, this study presents an enhanced predictive model that surpasses existing methods for early disease detection.
2. Comprehensive Algorithm Evaluation: This research stands out by holistically evaluating algorithm, namely the K-Nearest Neighbors (KNN) algorithm providing insights into their individual and comparative efficiencies.

http://www.doi.org/10.62341/istj-vol38-2-jr46

3. Sophisticated Feature Optimization: Using the power of domain expertise, this research has provided a sophisticated method for optimizing features, which increases the accuracy of predictions.
4. Easy Interpretation of Models: In our research methodology, interpretation is emphasized, making it easier for practitioners to understand how predictions work.

Background

Machine Learning Techniques

Within machine learning techniques, there exists a spectrum of methods, each tailored to extract distinct forms of information from a variety of datasets. Among these techniques, classification algorithms are approaches that involve the allocation of data into classes, based on specific attributes. However, clustering algorithms focus on grouping similar data points together, enabling the identification of patterns within the dataset. Moreover, association rule algorithms serve the purpose of revealing relationships and dependencies among various variables present in the dataset (Pujari, 2001).

K-Nearest Neighbors (KNN) Algorithm

(KNN) algorithm is a supervised machine learning technique. It's a simple but powerful algorithm used in machine learning for tasks like classification and regression, it simply calculated the distance of a new data point to all other training data points. where assigns the data point to the class to which the majority of K data points belong.

Domain Knowledge

Health informatics professionals come from a range of educational and training (Cynthia et al.,2020). Incorporating domain knowledge in machine learning model for diabetes disease prediction can be in details how this can be considered:

http://www.doi.org/10.62341/istj-vol38-2-jr46

- Feature Selection Based on Expertise: Cardiologist's professionals can identify key risk factors and clinical indicators critical for diabetes disease prediction. This includes factors like blood pressure, patient's medical history. Their insights ensure that the most relevant features are included in the model.
- Assigning Weights to Features: Once the relevant features are identified, medical experts can guide how much weight each feature should carry in the predictive model, which can be done as part of the implementation of ML algorithms that allow considering weights.

Literature review

Numerous studies have applied data mining algorithms for disease prediction. (Xie & Lv, 2021) present a novel methodology for predicting Heart Disease. Their approach involves leveraging a pre-trained Deep Neural Network (DNN) to extract essential features, implementing Principal Component Analysis (PCA) to reduce dimensional, and utilizing Logistic Regression (LR) for the prediction task.

Dritsas & Trigka (2023) focuses on the early prediction of disease using machine learning (ML) techniques as its primary objective. In particular, the study examines and compares various machine learning (ML) models and ensemble approaches in relation to predicting liver disease. The performance measures used in making the comparisons include accuracy, precision, recall, F-measure, and area under the curve (AUC). The experimental findings demonstrate that the Voting classifier surpasses the performance of other models, emerging as the most effective in predicting liver disease occurrence.

Ananey-Obiri & Sarku (2020) apply data exploration and mining techniques in order to uncover concealed patterns. The objective is to consider machine learning algorithms, including logistic linear regression, decision tree classifier, and Gaussian Naïve Bayes models, for the purpose of predicting the occurrence of heart diseases in patients.

http://www.doi.org/10.62341/istj-vol38-2-jr46

Gupta et al., (2022) suggest different methods to facilitate the early diagnosis of cardiac or heart disease, aiming to enable timely disease management. they employ several machine learning algorithms, including K-Nearest Neighbour, Decision Tree, Logistic Regression, Naive Bayes, and Support Vector Machine (SVM) models. The findings reveal that Logistic Regression outperformed all other supervised classifiers, exhibiting superior performance based on various performance metrics.

In (El-Sofany,2024). improved the accuracy and transparency of heart disease prediction through combine with explainable artificial intelligence (XAI) where evaluated several machine learning models on clinical datasets and demonstrated that integrating explain ability techniques can enhance both predictive performance.

In Rimal et al (2025) conducted a comparative study of several algorithms, including Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Random Forest, for heart disease prediction. where improved reliability by employed cross-validation techniques improve model and found that ensemble-based approaches achieved superior predictive performance compared with traditional classifiers. Their findings highlighted the importance of selecting appropriate classification techniques and validation strategies to improve diagnostic accuracy.

The Statement of the Problem

1. Current disease prediction models frequently face limitations due to their overemphasis on algorithms, overlooking key aspects like advanced feature selection, proper feature weighting. Such neglect results in difficulties attaining high accuracy and interpret ability in these models.
2. These models often overlook the clinical significance of different features, leading to insufficient consideration of their relevance. They typically fail to assign the right level of importance to various features, impacting their effectiveness.

Advancing diabetes disease prediction:
a comprehensive approach using machine learning and expert knowledge

http://www.doi.org/10.62341/istj-vol38-2-jr46

Research Methodology

Preprocessing

The study was based on the Pima Indians Diabetes Database that was found in the UCI Machine Learning Repository. There were 768 cases of patients in the data set, while the eight predictor variables included the number of pregnancies, glucose level, blood pressure, skin thickness, insulin, BMI, pedigree function of diabetes, and age.

The dataset consists of 500 non-diabetic cases and 268 diabetic cases. The following steps for data preprocessing will encompass as Fig 1:

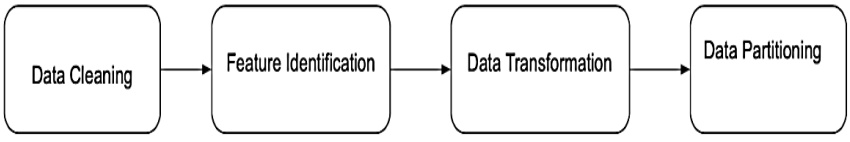

Fig.1: Steps of data prepressing

1. Data Cleaning: This involves the removal of incomplete, redundant, or inconsistent data entries.
2. Feature Identification: The domain knowledge can be used to select features as well as to assign weight for the selected ones.
3. Data Transformation: This involves modifying the raw data into format that can be analyzed. These may include normalizing, discretizing, and encoding of categorical variables.
4. Data Partitioning: t consists of partitioning the data set into training, validation, and test data sets.

Handling the Domain Knowledge

The domain knowledge is knowledge of a specific discipline or field in contrast to general knowledge. Domain knowledge in diabetes disease can be used to select features from the dataset to determine and predict the patient's current condition and expected condition based on his data. The domain knowledge can be used to select features as well as to assign weight for the selected ones. It is intended to engage a group of three practitioners, each possessing a

http://www.doi.org/10.62341/istj-vol38-2-jr46

minimum of three years of professional experience. The experts independently evaluated the importance of each diabetes risk factor. The final feature weights were obtained by averaging by the experts and subsequently normalizing the values so that the total weight equaled one.

Each expert was asked to rate the importance of features on a scale from 1 to 10 according to their clinical relevance. The average score for each feature was calculated and transformed into normalized weights used within the weighted KNN model. The following portion will elucidate on how the K-Nearest Neighbor (KNN) approach can be utilized in applying domain knowledge to predict illnesses:

A. The K-NN Approach (K-Nearest Neighbors)

The simplicity of the KNN method makes it quite an effective tool for predicting illnesses. In healthcare datasets, where each data point generally corresponds to an individual patient's medical profile, K-NN is capable of assessing the probability of a disease by examining the most analogous cases (data points) to the patient being considered. The assessment takes place by measuring certain medical factors like Age, Pregnancies, Glucose, and so forth. The first main strength of using the K-Nearest Neighbors (K-NN) technique in the medical sector involves the non-parametric nature of this method as it makes no assumptions regarding the distribution of the data. In the context of medicine, it helps to account for the deviation of medical data from traditional statistical distributions. Furthermore, k-NN's capability to adjust to alterations in the input data renders it well-suited for the dynamic nature of medical datasets.

However, the effectiveness of KNN is dependent on the meticulous selection of significant features and the correct choice of the number of neighbors (k) to include in the analysis. The use of a weighted k-NN version, which assigns different levels of significance to various attributes, can further refine the algorithm's accuracy. This is especially true in the context of intricate diseases, where some indicators may hold more diagnostic weight than others. The use of a weighted k-NN version, which assigns different levels of

http://www.doi.org/10.62341/istj-vol38-2-jr46

significance to various features. This is true in the context of intricate diseases, where some indicators may hold more diagnostic weight than others. The (K-NN) algorithm, a fundamental component of machine learning, typically processes all features equally in its distance calculations. this approach might not be ideal in intricate domains such as healthcare. For predicting diabetes disease, where elements like age and Insulin play pivotal roles, the integration of weighted features into k-NN proves highly beneficial. This article explores three distinct strategies: direct weighting guided by domain knowledge, indirect weighting through a Gaussian kernel, and an innovative hybrid of these two techniques. Incorporating explicit weight by domain knowledge into KNN weighting involves assigning weights to features, reflecting their established significance in the field. The weighted KNN should be calculated as following:

$$\mathrm{d}(xi,j\ ,\ zl,j\ ) = \sqrt{\sum_{i=1}^{n} w\ |\mathrm{xi,j}\ -\ \mathrm{zl,j}\ |\ 2}$$

Where X, Z are the values of an attribute for two patients, w is the weight for each attribute, which will be given by the professionals as display in table 1. This weighted KNN is a well-known formula that has been introduced by (Shepard, 1968) and many variants of these functions have been proposed that fit in the applied domain (Dudani, 1976).

Evaluation

The evaluation performance of the proposed approach as a subsequent assessment measures will be employed, as each measurement will show a side of the reported results, accuracy, recall, precision, F1-score and confusion matrix. The evaluation of proposed method's effectiveness will proceed in the following manner:

1. Initially, the selected machine learning algorithm without the integration of domain knowledge will be evaluated.
2. Results from step 1 will be gathered and subjected to analysis.

Advancing diabetes disease prediction:
a comprehensive approach using machine learning and expert knowledge

http://www.doi.org/10.62341/istj-vol38-2-jr46

3. the domain knowledge will be integrated, as previously detailed, into the algorithm.
4. The experiment will be executed again, and the outcomes will be documented and examined.
5. A comparative assessment will be performed between the results from both phases.

Results of Experiments

The Diabetes dataset was employed using the KNN model along with the weight-based KNN model based on expertise. The result as follow:

Table 1. Domain Knowledge-Based Feature Weights

#feature	feature	weights
1	Pregnancies	0.05
2	Glucose	0.35
3	BloodPressure	0.10
4	SkinThickness	0.05
5	Insulin	0.15
6	BMI	0.20
7	Diabetes PedigreeFunction	0.05
8	Age	0.05

The experimental results demonstrated that incorporating specialized knowledge improved the predictive performance of the KNN algorithm across all evaluation metrics. In this regard, accuracy of the weighted KNN model was found to be 72.73%, which was slightly higher than that of the KNN model, i.e., 70.13%. Moreover, there is a marked increase in the precision score from 58.33% to 62.50% and the recall score from 51.85%. Likewise, the area under curve (AUC) has increased to 0.7831 from 0.7405. These results suggest that assigning weights to features can enhance the effectiveness of diabetes prediction systems as in table 2.

http://www.doi.org/10.62341/istj-vol38-2-jr46

Table 2. Comparative Performance of Traditional KNN and Weighted KNN Models

Model	Accuracy	Precision	Recall	F1	AUC
Traditional KNN	70.13	58.33	51.85	54.90	0.7405
Weighted KNN	72.73	62.50	55.56	58.82	0.7831

Conclusion

The healthcare sector has witnessed substantial data growth in the past few years, offering opportunities and challenges for enhancing patient care and results. Through the analysis of extensive electronic health records (EHRs) and pertinent data sources, this research revealed concealed patterns and connections among patient attributes, medical backgrounds, and disease results. Existing approaches for disease prediction often exhibit limitations in terms of accuracy. Another notable concern with many existing models is their interpret ability. It is often difficult for practitioners to determine the logic behind particular predictions. The proposed research intends to overcome the weaknesses mentioned above. In this regard, harnessing machine learning algorithm and integrating domain knowledge, it aimed to develop a more sophisticated predictive model, enhancing early disease detection.

The algorithm *K-Nearest Neighbors* (*KNN*) are used to predict the disease by incorporating the domain knowledge. Based on that, the performance achieved an accuracy of 72.73%as a result of incorporating with the domain knowledge assigned. In addition, this approach will help the practitioners and convince them about the effectiveness of the proposed approach, as they will understand the way of how the model is working.

References

Alhumaidi, N. H., Dermawan, D., Kamaruzaman, H. F., & Alotaiq, N. (2025). The use of machine learning for analyzing real-

http://www.doi.org/10.62341/istj-vol38-2-jr46

world data in disease prediction and management: Systematic review. JMIR Medical Informatics, 13.

Ananey-Obiri, D., & Sarku, E. (2020). Predicting the presence of heart diseases using comparative data mining and machine learning algorithms. International Journal of Computer Applications, 176(11), 17–21.

Bomrah, S., Uddin, M., Upadhyay, U., Komorowski, M., & Priya, J. (2024). A scoping review of machine learning for sepsis prediction: Feature engineering strategies and model performance: A step towards explainability. Critical Care, 28, 180.

Cynthia, S., Cynthia, S., Elaine, B., Carla, M., Sandra, G., & Jeffrey, J. (2020). Domains, tasks, and knowledge for health informatics practice: Results of a practice analysis. Journal of the American Medical Informatics Association, 27(6), 845–852.

Dritsas, E., & Trigka, M. (2023). Supervised machine learning models for liver disease risk prediction. Computers, 12(1), Article 19.

Dudani, S. A. (1976). The distance-weighted k-nearest-neighbor rule. IEEE Transactions on Systems, Man, and Cybernetics, SMC-6(4), 325–327.

El-Sofany, H., Bouallegue, B., & Abd El-Latif, Y. M. (2024). A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method. Scientific Reports, 14, 23277.

Gupta, C., Saha, A., Reddy, N. S., & Acharya, U. D. (2022). Cardiac disease prediction using supervised machine learning techniques. Journal of Physics: Conference Series, 2161(1), 012013.

Islam, R., Sultana, A., & Islam, M. R. (2024). A comprehensive review for chronic disease prediction using machine learning algorithms. Journal of Electrical Systems and Information Technology, 11, 27.

Kolasa, K., Admassu, B., Hołownia-Voloskova, M., Kędzior, K. J., Poirrier, J.-E., & Perni, S. (2024). Systematic reviews of

http://www.doi.org/10.62341/istj-vol38-2-jr46

machine learning in healthcare: A literature review. Expert Review of Pharmacoeconomics & Outcomes Research, 24(1), 63–115.

Pujari, A. K. (2001). Data mining techniques. Universities Press.

Rimal, Y., Sharma, N., Paudel, S., Alsadoon, A., Koirala, M. P., & Gill, S. (2025). Comparative analysis of heart disease prediction using logistic regression, SVM, KNN, and random forest with cross-validation for improved accuracy. Scientific Reports, 15, Article 13444.

Sarku, E. (2025). Artificial intelligence in predictive healthcare: A systematic review. Journal of Clinical Medicine, 14(19), 6752.

Shepard, D. (1968). A two-dimensional interpolation function for irregularly-spaced data. In Proceeding of the 1968 23rd ACM national conference ,pp. 517-524.

Xie, S., Yu, Z., & Lv, Z. (2021). Multi-disease prediction based on deep learning: A survey. CMES - Computer Modeling in Engineering & Sciences, 128(2), pp.489–522.