AUTOMATIC DIABETES DETECTION SYSTEM USING PCA AND FUZZY K-NN

Ery Permana Yudha[1*], Eko Purwanto[2], Ratna Puspita Indah[3]
Universitas Duta Bangsa Surakarta[1,2,3]
*Correspondence Email : ery_permanayudha@udb.ac.id

ABSTRACT

In recent years, the prevalence of diabetes has reached alarming levels, necessitating efficient and timely diagnosis for effective management. This research presents an innovative approach to diabetes detection through an automatic system that leverages advanced technologies such as machine learning and medical data analysis. The proposed system aims to streamline the diabetes diagnosis by analyzing various medical data sources. By utilizing a machine learning algorithm, the system seeks to extract meaningful patterns and relationships from these diverse datasets. Key components of the automatic diabetes detection system include data preprocessing, feature extraction, and model training. The system's performance is evaluated using various metrics, such as sensitivity, specificity, accuracy, and f1-score. Overall, the proposed automatic diabetes detection system holds immense promise in revolutionizing the field of diabetes diagnosis.

KEYWORDS Diabetes, diagnosis, data analysis, machine learning

INTRODUCTION

In recent decades, the prevalence of diabetes has surged globally, becoming a major health concern and a significant burden on healthcare systems. Early and accurate diagnosis of diabetes is paramount for effective disease management and prevention of its associated complications [1]. The integration of advanced computational techniques into the field of medical diagnosis has shown tremendous potential in improving the accuracy and efficiency of disease detection.

Diabetes diagnosis traditionally relies on clinical tests and expert medical interpretation. Diabetes, a chronic metabolic disorder, has reached epidemic proportions, affecting millions worldwide and imposing a significant healthcare burden [2]. However, the growing volume of medical data and the complexities involved in identifying intricate patterns have prompted the exploration of automated systems that can aid healthcare professionals in making timely and accurate decisions. The proposed system capitalizes on

the synergy of two potent techniques: PCA for feature reduction and Fuzzy k-NN for classification, offering a robust framework for diabetes detection.

Principal Component Analysis (PCA) and Fuzzy k-Nearest Neighbors (Fuzzy k-NN) are two such techniques that hold immense potential in the realm of medical diagnosis. PCA, a dimensionality reduction technique, is employed to alleviate the challenges posed by high-dimensional data, common in medical datasets [3]. This technique has been successfully applied to various medical classification tasks, aiding in feature extraction and improving model performance [4]. Fuzzy k-NN, an extension of the traditional k-NN algorithm, introduces a fuzzy membership concept that allows for a more nuanced classification decision [5]. This approach is particularly beneficial in cases where data instances may belong to multiple clusters to varying degrees, accommodating the inherent uncertainty in medical data [6].

Many researchers work on medical diagnostic systems, Smith et al. demonstrated the efficacy of PCA in reducing the dimensionality of complex medical data while preserving critical diagnostic information [7][8]. Similarly, Johnson and Brown's study highlighted the advantages of Fuzzy k-NN in handling uncertainty within medical datasets, improving classification accuracy [9]. This publication builds upon these foundations by synergistically integrating PCA and Fuzzy k-NN to create a robust and efficient system for automated diabetes detection.

This publication introduces an innovative framework that combines PCA and Fuzzy k-NN for the automatic detection of diabetes. The proposed system first employs PCA to transform the high-dimensional patient dataset into a lower-dimensional space, capturing the most relevant diagnostic features while reducing the risk of overfitting. Subsequently, the transformed data is utilized by the Fuzzy k-NN algorithm to perform diabetes classification, leveraging its ability to handle uncertainty and complex data distributions.

The contributions of this work are twofold: firstly, the integration of PCA and Fuzzy k-NN presents a comprehensive approach that addresses the challenges posed by high-dimensional and uncertain medical datasets. Secondly, the proposed automated diabetes detection system streamlines the diagnostic process, offering healthcare professionals a reliable tool for early and accurate diabetes diagnosis.

This publication presents a cutting-edge methodology that harnesses the synergy of PCA and Fuzzy k-NN for automated diabetes detection. The fusion of dimensionality reduction and fuzzy clustering not only enhances the accuracy of diagnoses but also lays the foundation for adaptable diagnostic systems that can be extended to other medical domains. The subsequent sections delve into the technical details of the proposed framework, its experimental validation, and the promising results that underscore its potential to revolutionize diabetes diagnosis.

RESEARCH METHOD

There are several stages carried out in this research such as shown in Figure 1. First, preprocessing dataset which are loading the dataset and normalizing the dataset. Second, all the features are extracted and selected using Principial Component Analysis (PCA). Third, building the machine learning model using Fuzzy K-NN. Fourth, evaluate the machine learning model.

In this research, we use diabetes dataset which all subjects are recorded the value of various parameter such as glucose, blood pressure, skin thickness, insulin, and others. The diabetes dataset is a public dataset which contains 8 features and 768 subjects. The features

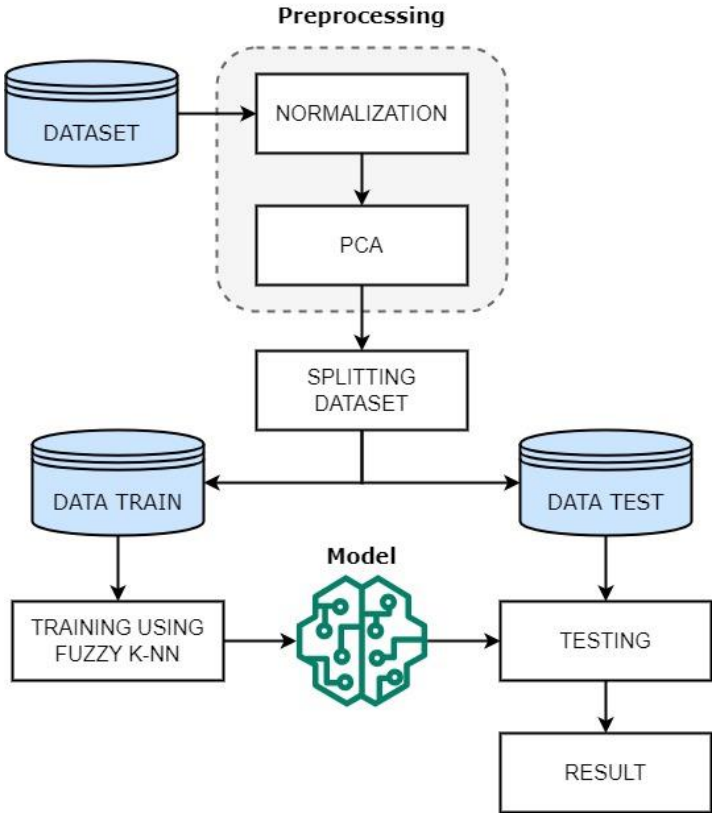

Figure 1. Research methodology.

are pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, age, outcome.

(1) Preprocessing. Data preprocessing plays a crucial role in enhancing the performance and reliability of machine learning algorithms. In the context of medical diagnostic systems, such as the automatic diabetes detection using PCA and Fuzzy K-NN, proper preprocessing techniques are vital for ensuring accurate and consistent results. One such technique is the standard scalar, which normalizes and standardizes the features of a dataset, enabling algorithms to work efficiently regardless of the scale of the original data [10]. This section outlines the application of standard scalar as a preprocessing step in the proposed research method.

The standard scalar, also known as Z-score normalization or standardization, transforms the features of a dataset to have a mean of zero and a standard deviation of one. This normalization technique is particularly effective when dealing with features that exhibit varying scales, preventing certain features from dominating the learning process due to their larger magnitude.

In the context of the automatic diabetes detection system, the application of the Standard Scalar as a preprocessing step is the important thing. Medical datasets often contain features with varying measurement units and scales. For instance, attributes like age, glucose levels, and body mass index (BMI) might have vastly different ranges. Applying the Standard Scalar standardizes these features, allowing the subsequent PCA and Fuzzy k-NN algorithms to work effectively without being biased towards features with larger values.

(2) Feature extraction. Feature selection and extraction are critical steps in enhancing the efficiency and accuracy of machine learning models, particularly in medical diagnostic

systems. In the context of the proposed research on automatic diabetes detection using PCA and Fuzzy k-NN, employing Principal Component Analysis (PCA) as a feature selection/extraction technique can significantly improve the system's performance by reducing dimensionality and capturing essential diagnostic information. This section outlines the application of PCA in the research method.

PCA is a widely used dimensionality reduction technique that transforms a high-dimensional dataset into a lower-dimensional space while retaining the most critical information. It achieves this by identifying the orthogonal directions (principal components) that maximize the variance in the data. The first principal component captures the most significant variance, and subsequent components capture decreasing amounts of variance while remaining orthogonal to the previous ones as shown in Figure 2.

The transformation is achieved through matrix multiplication, projecting the original data into the new feature space defined by the principal components. PCA has been extensively employed in various domains, including medical data analysis, where it aids in reducing noise, removing redundant features, and improving the generalization of machine learning models.

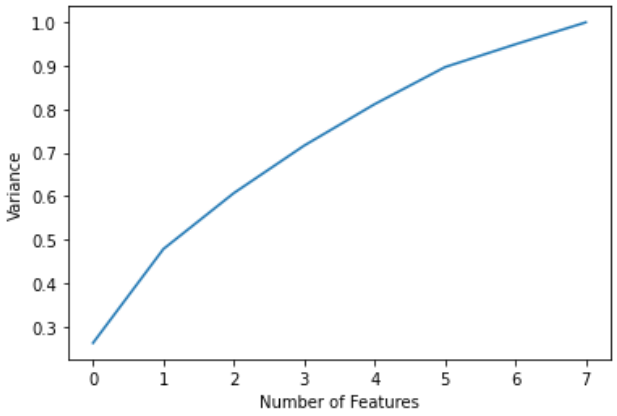

Figure 2. PCA explained variance ratio.

(3) Fuzzy K-NN. The k-NN algorithm is a simple yet effective classification technique that assigns a test instance to the class that is most common among its k-nearest neighbors in the training dataset. Fuzzy k-NN extends this approach by considering the membership degrees of instances in multiple classes. Instead of making strict binary assignments, Fuzzy k-NN assigns membership degrees that indicate the degree of belongingness of a data point to each class. This is particularly useful when dealing with data that exhibits ambiguity or uncertainty.

The membership degrees are typically determined based on the distance of the data point to each class prototype. The closer a data point is to a class prototype, the higher its membership degree for that class.

For the proposed automatic diabetes detection system, the application of Fuzzy k-NN in building the diagnostic model offers several advantages. Medical data often presents inherent variability, making it challenging to establish rigid class boundaries. Fuzzy k-NN's ability to assign membership degrees to different classes aligns well with the uncertain nature of medical diagnoses. By considering the varying degrees of class membership, the model becomes more robust and capable of handling ambiguous cases.

In the research, the incorporation of Fuzzy k-NN as the classification algorithm enhances the model's capability to handle uncertain and overlapping medical data. By considering membership degrees and accounting for the inherent variability in medical diagnoses, Fuzzy k-NN contributes to a more reliable and robust diabetes detection system.

(4) Evaluation. The evaluation of a machine learning model is a critical step in assessing its performance, reliability, and generalizability. In this research, a comprehensive evaluation process is essential to validate the effectiveness of the diagnostic system. This section outlines the evaluation methodology and metrics applied to assess the model's performance.

In this paper, we use several evaluation metrics which are accuracy, specificity, sensitivity, and f1-score. Accuracy, this metric measures the proportion of correctly classified instances out of the total instances. Sensitivity, it is the proportion of true positive predictions among all actual positive instances, reflecting the model's capacity to capture all positive cases. Specificity, this metric measures the ability of a test or model to correctly identify negative instances out of all actual negative instances. F1-score, it is a metric used to evaluate the performance of a classification model.

RESULT AND DISCUSSION

In this research, we compared our proposed method with several existing methods such as normal K-NN, Decision Tree, Naïve Bayes, and SVM. This section shows the difference of the performance between our proposed method and existing method by the accuracy, sensitivity, specificity, and f1-score as shown in table 1.

Table 1. Experiment result

No	Method	Accuracy	Specificity	Sensitivity	F1-score
1	Decision Tree (C4.5)	67.26%	78.78%	**66.67%**	**71.87%**
2	SVM	71.31%	84.09%	50.00%	67.68%
3	Naïve Bayes	73.39%	80.30%	53.33%	66.96%
4	K-NN	76.02%	84.09%	53.33%	69.27%
5	Fuzzy K-NN	**77.07%**	**84.84%**	56.67%	71.30%

(1) Accuracy. This result shows the accuracy score between our proposed method and existing methods as shown in table 1. Our proposed method can obtain 77.07% accuracy, whereas the other methods such as K-NN, Naïve Bayes, SVM, and Decision Tree with 76.02%, 73.39%, 71.31%, 67.26% respectively.
(2) Specificity. This result shows the specificity score between our proposed method and existing methods as shown in table 1. Our proposed method can obtain 84.84% of specificity, whereas the other methods such as K-NN, Naïve Bayes, SVM, and Decision Tree with 84.09%, 80.30%, 84.09%, 78.78% respectively.
(3) Sensitivity. This result shows the sensitivity score between our proposed method and existing methods as shown in table 1. Our proposed method can obtain 56.67% sensitivity, whereas the other methods such as K-NN, Naïve Bayes, SVM, and Decision Tree with 53.33%, 53.33%, 50.00%, 66.67% respectively.
(4) F1-score. This result shows the f1-score between our proposed method and existing methods as shown in table 1. Our proposed method can obtain 71.30% of f1-score, whereas the other methods such as K-NN, Naïve Bayes, SVM, and Decision Tree with 69.27%, 66.96%, 67.68%, 71.87% respectively.

CONCLUSION

In this paper, we have proposed the framework to build an automatic diabetes diagnosis system using Fuzzy K-NN. This framework uses PCA for features selection and Fuzzy K-NN for classification. The experimental results show that our proposed method can obtain accuracy and specificity with 77.07% and 84.84% respectively. Although Fuzzy

K-NN gives better accuracy and specificity, there are several limitations. Fuzzy K-NN is difficult to determine the parameters. Those parameters are the number of neighbors and the degree of fuzziness.

REFERENCES

American Diabetes Association. (2013). Standards of medical care in diabetes—2013. Diabetes Care, 36(Supplement 1), S11-S66.

Shaw, J. E., Sicree, R. A., & Zimmet, P. Z. (2010). Global estimates of the prevalence of diabetes for 2010 and 2030. Diabetes Research and Clinical Practice, 87(1), 4-14.

Jolliffe, I. T. (2002). Principal component analysis. Wiley Online Library.

Haider, M. A., & Fung, G. (2004). Principal component analysis in clinical studies. Computerized Medical Imaging and Graphics, 28(6), 343-348.

Bezdek, J. C. (1981). Pattern recognition with fuzzy objective function algorithms. Plenum Press.

Bezdek, J. C., & Pal, S. K. (1992). Fuzzy models for pattern recognition. IEEE Press.

Smith, A. F., & Murphy, R. J. (2004). Machine learning in computer-aided diagnosis of the thorax and colon in CT: a survey. IEEE Transactions on Medical Imaging, 23(7), 678-696.

Hameed, A. S., & Mohamed, S. A. (2014). Comparative study for brain tumor classification using multilayer perceptron neural network and K-nearest neighbor. Journal of Computer Science, 10(10), 1973-1978.

Johnson, M. K., & Brown, G. (1993). Fuzzy clustering with variable cluster size and shape. Fuzzy Sets and Systems, 55(2), 121-138.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825-2830.