JURNAL TEKNOLOGI REKAYASA INFORMASI DAN KOMPUTER
E-ISSN: 2797-1724, DOI: 10.30811/jtrik.v9i1.8721
Vol. 9 No. 1 2026 pp. 01-07

Application of the K-Nearest Neighbors (KNN) Algorithm for Diabetes Mellitus Classification: Evidence from Aceh Province, Indonesia

Rizki Wahyu Aulia [1], Azhar [2], Rahmad Hidayat[3], Cut Aulia Safira Rachman[4*]

[1,2,3] Jurusan Teknologi Informasi dan Komputer, Politeknik Negeri Lhokseumawe, Indonesia
[4] Sekolah Vokasi, Universitas Gadjah Mada, Yogyakarta, Indonesia

*Corresponding Author: cut.aulia.safira.rachman@mail.ugm.ac.id

Article info: Received 05/01/2026, Revised 03/02/2026, Accepted 20/02/2026

Abstract

Diabetes mellitus is a non-communicable disease with a steadily increasing prevalence in Indonesia, including Aceh Province. Early detection using data-driven approaches is essential to minimize the risk of severe complications. This study aims to classify diabetes mellitus by implementing the K-Nearest Neighbors (KNN) algorithm. The dataset comprises 1,500 instances from the Pima Indians Diabetes Dataset obtained from Kaggle and an additional 100 instances collected from hospitals across Aceh Province. Data preprocessing involved normalization and label encoding, followed by data partitioning into training and testing sets using a 90:10 ratio. The KNN model was configured with a parameter value of K=5. Experimental results indicate that the proposed model achieved an accuracy of 85%, precision of 87%, recall of 82%, and an F1-score of 85% on the Kaggle dataset. For the hospital dataset, the model attained an accuracy of 76%, precision of 80.95%, recall of 68%, and an F1-score of 73.91%. These findings suggest that the KNN algorithm demonstrates adequate performance in classifying diabetes mellitus and may serve as a basis for the development of data-driven medical decision support systems.

Keywords: Diabetes, Machine Learning, Klasifikasi, K-Nearest Neighbors

1. Introduction

Diabetes mellitus is classified as one of the major non-communicable diseases (NCDs) that has become a significant public health concern at both global and national levels. This condition is characterized by persistent hyperglycemia resulting from impaired insulin secretion, reduced insulin action, or a combination of both mechanisms. As a chronic metabolic disorder, diabetes mellitus requires long-term management due to its potential to cause severe complications, including cardiovascular disease, nephropathy, neuropathy, retinopathy, and lower-limb amputation when glycemic control is inadequate [1]. The World Health Organization (WHO) identifies diabetes as one of the ten leading causes of mortality worldwide, with the number of affected individuals projected to increase steadily each year [2]. A similar trend is observed in Indonesia, where the prevalence of diabetes mellitus has shown a continuous rise. Data from the National Basic Health Research indicate that this increase is closely associated with lifestyle changes, particularly among urban populations [3]. High consumption of sugar- and fat-rich foods, insufficient physical activity, and the growing prevalence of obesity are recognized as major risk factors contributing to the escalation of diabetes cases. Furthermore, limited public awareness regarding the importance of routine health screening has resulted in a substantial proportion of diabetes cases remaining undiagnosed during the early stages of the disease [4].

Aceh Province represents one of the regions facing a comparable increase in diabetes prevalence. According to reports from the Aceh Provincial Health Office, more than 4,500 cases of diabetes were recorded in Lhokseumawe City in 2023, demonstrating an upward trend compared to previous years [5]. This situation highlights that diabetes mellitus has emerged as a serious public health issue that demands focused attention, particularly in the context of preventive strategies and early detection efforts. Early identification of diabetes remains challenging, as the disease often progresses without distinctive clinical symptoms during its initial phase. Consequently, many individuals become aware of their diabetic condition only after experiencing disease-related

complications [6]. Conventional diagnostic approaches typically rely on laboratory examinations and clinical assessments conducted by healthcare professionals. Although these methods are considered highly accurate, they are time-consuming, costly, and dependent on direct medical involvement, which limits their practicality for large-scale screening programs [7].

Advances in information technology and computational sciences have created substantial opportunities for the utilization of health-related data. One rapidly evolving approach is machine learning, a subfield of artificial intelligence (AI), which enables computer systems to learn patterns from historical data and generate predictions or classifications without explicit rule-based programming [8]. Within the healthcare domain, machine learning has been widely applied for disease diagnosis, risk prediction, and the development of clinical decision support systems [9]. Numerous studies have explored the application of machine learning techniques in diabetes classification. By leveraging medical attributes such as blood glucose levels, blood pressure, body mass index (BMI), age, and other relevant clinical factors, machine learning algorithms are capable of identifying patterns associated with diabetes risk [10]. This approach has demonstrated considerable potential in supporting healthcare professionals, particularly during the early screening stage of patient assessment [11].

Among various machine learning algorithms, K-Nearest Neighbors (KNN) is frequently employed for medical data classification. KNN is a supervised learning algorithm that classifies new instances based on their proximity to a predefined number of nearest samples in the training dataset [12]. The algorithm is widely recognized for its conceptual simplicity, ease of implementation, and the absence of a complex model training process. Additionally, KNN is adaptable to diverse data types, including medical datasets with low to moderate dimensionality [13]. Several empirical studies have reported that the KNN algorithm can achieve competitive performance in diabetes classification tasks. For instance, research conducted by Smith et al. using the Pima Indians Diabetes Dataset demonstrated that KNN attained accuracy levels comparable to those of other classification algorithms [14]. Further investigations have shown that KNN performance can be enhanced through appropriate data normalization, optimal selection of the K parameter, and the use of suitable distance metrics [15].

Despite its advantages, the effectiveness of the KNN algorithm is highly dependent on data quality. Medical datasets frequently contain missing values, class imbalance, and heterogeneous feature scales, which can negatively affect classification performance [16]. Therefore, data preprocessing steps such as data cleaning, normalization, and feature selection are essential to ensure optimal model performance [17] [18]. Based on this background, the present study aims to implement the K-Nearest Neighbors algorithm for diabetes disease classification. The dataset utilized in this study consists of the Pima Indians Diabetes Dataset obtained from Kaggle, complemented by additional hospital-based data. The Pima Indians Diabetes Dataset was selected due to its widespread use as a benchmark dataset in diabetes research and its inclusion of clinically relevant attributes [19][20]. Model performance evaluation was conducted using a Confusion Matrix to compute standard evaluation metrics, including accuracy, precision, recall, and F1-score [21][22][23]. These metrics are critical for assessing the model's capability to accurately distinguish between individuals at risk of diabetes and those without the condition mThe findings of this study are expected to provide a foundation for the development of data-driven medical decision support systems capable of facilitating early detection of diabetes in a more efficient and accurate manner [24] [25]. Moreover, this research is anticipated to contribute to the growing body of academic literature on the application of machine learning in healthcare, particularly in the classification of non-communicable diseases.

2. Methods

The architecture of the diabetes classification system in this study (as shown in Figure 1) begins with the preprocessing of a clinical dataset containing patient information, including the number of pregnancies, glucose levels, blood pressure, skin thickness, insulin levels, body mass index (BMI), family history, and age. The data undergoes preprocessing, which includes normalization to standardize the scale of numerical variables, ensuring more accurate distance calculations in the KNN algorithm, and label encoding to convert categorical variables into numerical form. Subsequently, output labeling is performed, assigning a label of 1 for patients indicated as diabetic and 0 for non-diabetic patients. The next step involves splitting the dataset into training and testing subsets, with 90% allocated for training and 10% for testing. The training data is used to develop the KNN model, while the testing data evaluates its performance on unseen instances. During model training, the KNN algorithm computes the distance between instances and determines the class of new data points based on the majority vote of the K nearest neighbors. This process results in a model capable of classifying whether a patient has diabetes or not.

The classification outcomes are then evaluated using performance metrics, including accuracy, precision, recall, and F1-score, to assess the model's effectiveness in detecting diabetic patients. The final step involves saving the trained model along with the normalization parameters, enabling predictions on new data without the need for retraining. Through this workflow, the developed system provides an efficient and practical approach for early diabetes detection.

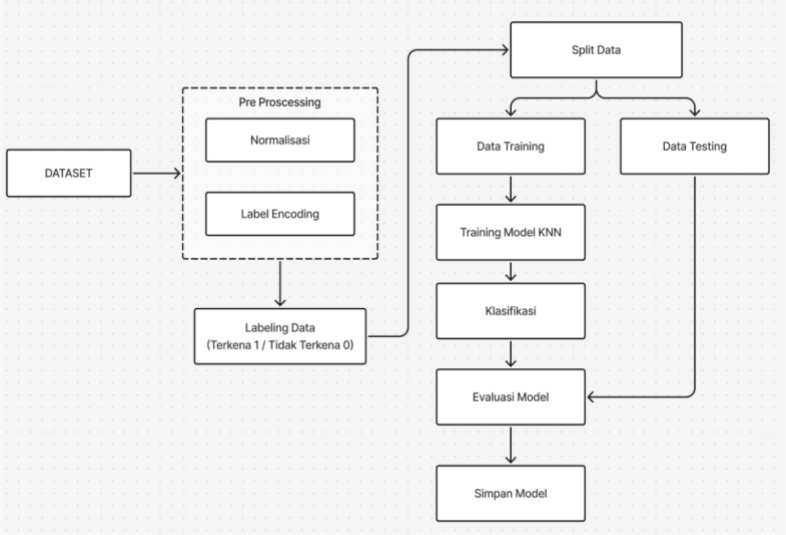

Figure 1. Model Architecture

A. Data Collection

This study utilizes secondary data in the form of the Pima Indians Diabetes Dataset obtained from Kaggle, consisting of 1,500 patient records. The dataset includes eight primary clinical variables: number of pregnancies, plasma glucose concentration, diastolic blood pressure, skinfold thickness, serum insulin level, body mass index (BMI), diabetes pedigree function, and patient age. These variables are used to classify patients into two categories, namely diabetic (1) and non-diabetic (0). In addition, this research incorporates 100 supplementary records obtained from a hospital in Aceh, which serve as an external testing dataset. The examples data in Pima Indian Dataset shown in Table 1.

Table 1. Example data in Pima Indian Dataset

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1

The Aceh hospital dataset contains the same clinical variables as the Kaggle dataset, including glucose level, blood pressure, BMI, skin thickness, insulin level, number of pregnancies, family history of diabetes, and patient age. The inclusion of hospital-based data aims to evaluate the robustness and generalizability of the proposed model when applied to real-world clinical data with greater variability, thereby providing a more realistic assessment of model performance in practical healthcare settings.

B. Data Preprocessing

Prior to model development, several data preprocessing steps were performed to enhance data quality and ensure reliable classification performance. First, data normalization was applied to standardize the scale of each variable, thereby preventing attributes with larger numerical ranges from disproportionately influencing the distance calculations. This step is particularly important for distance-based algorithms such as K-Nearest Neighbors. Second, label encoding was applied to the target variable to convert categorical class labels into a numerical format suitable for computational processing. Additionally, the dataset was examined to confirm the absence of missing values that could adversely affect the classification process. Following preprocessing, the dataset was partitioned into two subsets: 90% of the data were allocated for training, while the remaining 10% were reserved for testing. This split ratio was selected to provide the model with a sufficiently large training set while maintaining a representative test set for an objective evaluation of model performance.

C. Research Flowchart

The flowchart of the diabetes classification system using the K-Nearest Neighbors (KNN) method is presented in Figure 2. The process begins with data collection, including patients' medical variables such as glucose level,

blood pressure, body mass index, insulin level, skin thickness, number of pregnancies, family history, and age. Next, data preprocessing is performed, including normalization to standardize data scales and label encoding to adjust data formats. Subsequently, the dataset is split into training and testing sets with a 90:10 ratio. The next stage involves training the KNN model using the training data to establish classification patterns. The model is then evaluated using the testing data, with performance measured by accuracy, precision, recall, and F1-score metrics. The final stage is the classification results, where the system predicts whether a patient is diagnosed with diabetes or not. With this architecture, the system is expected to support accurate and efficient early detection of diabetes.

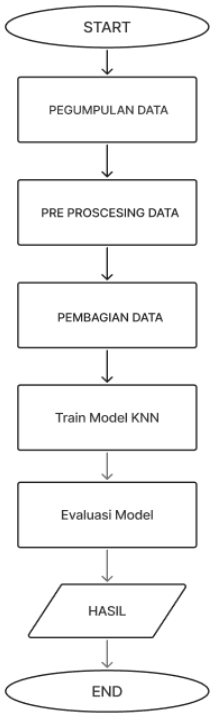

Figure 2. Research Flowchart

D. K-Nearest Neighbors

The K-Nearest Neighbor (KNN) algorithm is an instance-based classification method that leverages training data to assess the similarity between test and training instances [5]. Distances are computed using the Euclidean Distance formula as follows. KNN operates on the principle that data points in close proximity are more likely to share the same class. In this study, a value of K = 5 was employed, with the Euclidean Distance function serving as the metric for measuring similarity

$$\boldsymbol{euc} = \sqrt{(\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{pi} - \boldsymbol{qi})^2)} \quad \text{..... (1)}$$

pi = sample data / data training
qi = data testing
n = variabel data

Each test instance is compared with all training instances, and the distances are calculated. The five nearest neighbors are then selected. The class of the test instance is determined based on the majority voting of these nearest neighbors.

3. Result and Discussions

A. Testing Results on Pama Indians Dataset

The diabetes disease classification model was developed using the K-Nearest Neighbors (KNN) algorithm with K = 5 and the Euclidean distance function. The dataset employed was the Pima Indians Diabetes dataset, consisting

of 1,500 patient records. After preprocessing, which included normalization and data splitting, 1,350 instances were allocated for training and 150 for testing. The model training process yielded satisfactory performance. Based on the results of the confusion matrix, the evaluation metrics were obtained shown in Figure 3.

```
Accuracy : 0.8543046357615894
Precision: 0.875
Recall   : 0.8289473684210527
F1 Score : 0.8513513513513513
```

Figure 3. Performance on Pama Indians Dataset

An accuracy of 85% indicates that the majority of test instances were correctly predicted. The high precision value (87%) demonstrates that the model is reliable in identifying patients who truly have diabetes, with minimal misclassification of negative cases (false positives). However, the recall value of 82% indicates that there remain some patients who actually have diabetes but were not detected by the model (false negatives). This aspect is particularly critical in a medical context, as false negative errors may result in patients not receiving timely treatment. The F1-score of 85% reflects a good balance between precision and recall.

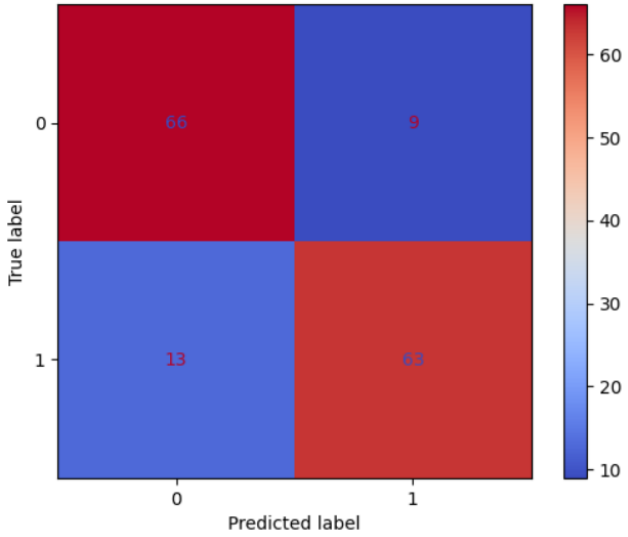

Figure 4. Classification Results on Pama Indians Dataset

From Figure 4 can be seen that the model demonstrates a balanced performance in classifying both healthy and diabetic patients. Out of a total of 151 test instances, 129 were correctly classified, resulting in an accuracy of approximately 85%. Misclassifications primarily occurred in 13 false negative cases, where patients who actually had diabetes were predicted as healthy. This is particularly critical in a medical context, as false negatives pose a higher risk compared to false positives. Nevertheless, the high precision value (87%) suggests that the model's predictions for diabetic patients are relatively reliable, while a recall of 82% indicates that there remains room for improving the model's sensitivity.

B. Testing Results on Aceh Hospitals Data

Testing using data from aceh hospitals, consisting of 100 patient records, yielded model performance different from that observed with the Kaggle dataset. Based on the confusion matrix shown in Figure 2, the KNN model achieved an accuracy of 76%, precision of 80.95%, recall of 68%, and an F1-score of 73.91%. The confusion matrix indicates that 21 healthy patients were correctly predicted (True Negatives) and 17 diabetic patients were accurately detected (True Positives). However, there were still 4 false positive cases, in which healthy patients were incorrectly classified as diabetic, and 8 false negative cases, where diabetic patients were mistakenly predicted as healthy. Misclassifications, particularly the false negatives, are of critical concern because patients with diabetes may not receive timely medical intervention. These results highlight that, although KNN can deliver reasonably good performance, its effectiveness is still influenced by the more complex distribution of real-world data compared to standard datasets. Figure 5 show the classification result on Aceh Hospitals data.

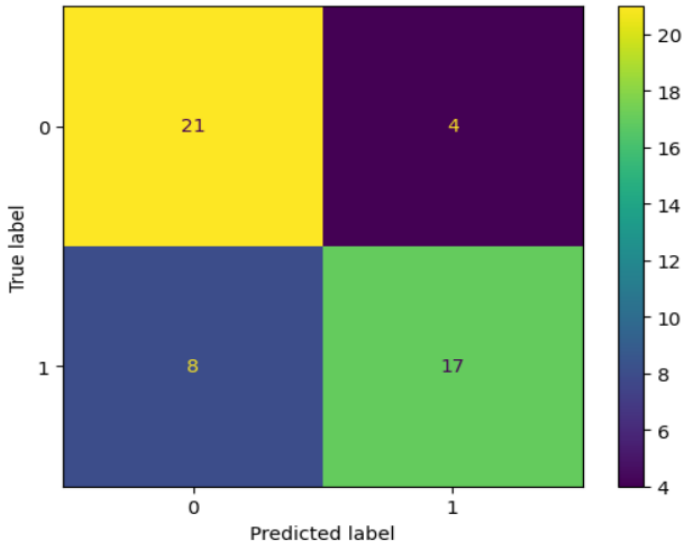

Figure 5. Classification Result on Aceh Hospitals Data

C. Benchmarking

In this subsection, benchmarking is conducted to compare the performance of the classification model used in this study, namely the K-Nearest Neighbors (KNN) algorithm, with the results of previous research that employed a neural network algorithm, as used in the study by Striosno & Jupron (2024) entitled "Analisa Klasifikasi Penyakit diabetes dengan Algoritma Neutal Network" as a method for diabetes disease classification. Both studies use the same dataset, namely the Kaggle Diabetes dataset, but with different algorithmic approaches. Striosno employed a Multilayer Perceptron (MLP) architecture and used 20% of the data for testing, whereas this study uses a distance-based approach through the KNN algorithm with five neighbors and 10% of the data for testing. The results of both studies can be seen in Table 2.

Table 2. The Benchmarking with Previous Study

No	Matrix	Sutrisno (2024)	Proposed Model
1	Accuracy	80,52%	85,43%
2	Precision	77,55%	87,50%
3	Recall	66,67%	82,89%
4	F1-Score	71,69%	85,10%

Based on the comparison of classification performance between Sutrisno (2024) and the proposed model, the latter demonstrates improvements across all evaluation metrics. The proposed K-Nearest Neighbors (KNN) model achieved an accuracy of 85.43%, surpassing Sutrisno's 80.52%, indicating a higher overall correctness in classification. Precision also increased from 77.55% to 87.50%, reflecting a greater proportion of correctly predicted positive cases and a lower rate of false positives. Recall improved substantially from 66.67% to 82.89%, demonstrating enhanced capability in identifying actual positive cases and reducing false negatives. Consequently, the F1-Score rose from 71.69% to 85.10%, suggesting a better balance between precision and recall and confirming the proposed model's superior reliability in diabetes disease classification.

4. Conclusion

This study implemented the K-Nearest Neighbors (KNN) algorithm for diabetes disease classification using two data sources: the Pima Indians Diabetes Dataset from Kaggle and additional hospital data. The results indicate that, on the Kaggle dataset, the KNN model achieved an accuracy of 85%, precision of 87%, recall of 82%, and an F1-score of 85%. In contrast, the model's performance decreased on the hospital dataset, with an accuracy of 76%, precision of 80.95%, recall of 68%, and an F1-score of 73.91%. This performance discrepancy suggests that the

real-world hospital data exhibit a more complex distribution compared to the standard dataset, which affects classification outcomes. Nevertheless, the KNN algorithm is still demonstrated to be an effective and straightforward method that can support data-driven medical decision-making systems. Future research may focus on optimizing the K parameter, selecting more relevant features, increasing the dataset size from diverse sources, and integrating ensemble learning techniques to further enhance classification performance.

REFERENCES

[1]. World Health Organization, Global Report on Diabetes, Geneva, Switzerland: WHO Press, 2016.
[2]. World Health Organization, "Diabetes," WHO Fact Sheets, Geneva, Switzerland, 2023.
[3]. International Diabetes Federation, IDF Diabetes Atlas, 10th ed., Brussels, Belgium: IDF, 2022.
[4]. Kementerian Kesehatan Republik Indonesia, Profil Kesehatan Indonesia Tahun 2022, Jakarta, Indonesia: Kemenkes RI, 2022.
[5]. Dinas Kesehatan Aceh, Profil Kesehatan Provinsi Aceh Tahun 2023, Banda Aceh, Indonesia: Dinkes Aceh, 2023.
[6]. American Diabetes Association, "Classification and diagnosis of diabetes," Diabetes Care, vol. 46, no. Supplement 1, pp. S19–S40, Jan. 2023, doi: 10.2337/dc23-S002.
[7]. J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," Proceedings of the Annual Symposium on Computer Application in Medical Care, Washington DC, USA, pp. 261–265, 1988.
[8]. T. M. Mitchell, Machine Learning, New York, NY, USA: McGraw-Hill Education, 1997.
[9]. T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21–27, Jan. 1967, doi: 10.1109/TIT.1967.1053964.
[10]. J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed., San Francisco, CA, USA: Morgan Kaufmann, 2012.
[11]. K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," Computational and Structural Biotechnology Journal, vol. 13, pp. 8–17, 2015, doi: 10.1016/j.csbj.2014.11.005.
[12]. Esteva et al., "A guide to deep learning in healthcare," Nature Medicine, vol. 25, no. 1, pp. 24–29, Jan. 2019, doi: 10.1038/s41591-018-0316-z.
[13]. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," Journal of Machine Learning Technologies, vol. 2, no. 1, pp. 37–63, 2011.
[14]. M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," Information Processing & Management, vol. 45, no. 4, pp. 427–437, Jul. 2009, doi: 10.1016/j.ipm.2009.03.002.
[15]. H. Liu and H. Motoda, Feature Selection for Knowledge Discovery and Data Mining, Boston, MA, USA: Springer, 1998.
[16]. J. Dougherty, R. Kohavi, and M. Sahami, "Supervised and unsupervised discretization of continuous features," Proceedings of the 12th International Conference on Machine Learning, Tahoe City, CA, USA, pp. 194–202, 1995.
[17]. S. Shankar, S. K. Lakshmanaprabu, S. K. S. Raj, and A. Maseleno, "Optimal feature selection-based diabetic retinopathy classification using KNN," International Journal of Engineering and Advanced Technology, vol. 9, no. 1, pp. 1656–1662, Oct. 2019.
[18]. Kaggle, "Pima Indians Diabetes Database," Kaggle Datasets, 2024. [Online]. Available: https://www.kaggle.com
[19]. Alpaydin, Introduction to Machine Learning, 3rd ed., Cambridge, MA, USA: MIT Press, 2014.
[20]. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, 4th ed., Burlington, MA, USA: Morgan Kaufmann, 2016.
[21]. H. Shortliffe and J. J. Cimino, Biomedical Informatics: Computer Applications in Health Care and Biomedicine, 4th ed., London, UK: Springer, 2014.
[22]. Z. Obermeyer and E. J. Emanuel, "Predicting the future — Big data, machine learning, and clinical medicine," The New England Journal of Medicine, vol. 375, no. 13, pp. 1216–1219, Sep. 2016, doi: 10.1056/NEJMp1606181.
[23]. J. Topol, Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again, New York, NY, USA: Basic Books, 2019.
[24]. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.
[25]. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, Cambridge, MA, USA: MIT Press, 2016.