Jurnal Teknik Informatika (JUTIF) Vol. 6, No. 5, October 2025, Page. 3217-3229
P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5277

Web-Based Diabetes Risk Prediction System Using K-NN on Kaggle Early Stage Diabetes Dataset

Fahmi Ruziq*[1], M. Rhifky Wayahdi[2]

[1,2]Information Systems, Universitas Battuta, Indonesia

Email: [1]fahmiruziq89@gmail.com

Received : Aug 18, 2025; Revised : Sep 6, 2025; Accepted : Sep 13, 2025; Published : Oct 16, 2025

Abstract

Diabetes mellitus affects approximately 537 million adults globally, and its rising prevalence poses serious health and economic burdens. Early detection is crucial to reduce risks of complications and improve patient outcomes. This study aims to design and implement a web-based diabetes risk prediction system using the K-Nearest Neighbors (K-NN) algorithm to support early detection based on symptoms. The system utilizes the Kaggle Early Stage Diabetes Risk Prediction Dataset containing 520 records with 17 symptom attributes and one class label. Data preprocessing includes converting categorical data into numerical values, discretizing age into predefined ranges, and applying min-max scaling to normalize feature values. K-NN classification was conducted with K values of 1, 3, and 5, using the PHP Machine Learning (PHP-ML) library and MySQL database integration. The system achieved its highest accuracy of 93.46% at K = 1. Manual testing confirmed that the system processes symptom inputs correctly and provides predictions consistent with training data. This web-based tool offers an accessible platform for early diabetes risk screening, supporting self-assessment and triage. It demonstrates that PHP-ML can effectively implement machine learning in a web environment and can be further enhanced through parameter optimization and integration with larger, more diverse datasets to strengthen generalization.

Keywords : *Data Preprocessing, Diabetes prediction, Early Detection, K-Nearest Neighbours, PHP-ML, Web-based system.*

1. INTRODUCTION

Diabetes mellitus (DM) is a chronic disease [1] that has become a global health problem [2] with a prevalence that continues to increase from year to year [3]. According to data from the International Diabetes Federation (IDF), in 2021, approximately 537 million adults in the world were living with diabetes, and this figure is projected to increase to 643 million by 2030 [4]. Diabetes not only impacts the health of individuals but also poses a significant economic burden [5] to health systems and society as a whole. Early detection and proper management are essential [6] to prevent serious complications such as cardiovascular disease [7], kidney failure [8], and nerve damage [9].

Symptoms of diabetes are often non-specific and can be similar to other health conditions, so many cases go undiagnosed until they reach a more severe stage [10]. Common symptoms such as increased thirst, frequent urination, fatigue, and weight loss [11] are often ignored or considered as minor health issues. Therefore, there is a need for a system that can help predict the risk of diabetes based on the symptoms experienced by individuals.

One method that can be used to develop a diabetes risk prediction system is K-Nearest Neighbor (K-NN) [12-15]. K-NN is a simple yet effective machine learning algorithm for classification and regression [16]. It works by comparing the distance between new data and existing data in the dataset [17] to determine the most appropriate category or prediction. The advantages of K-NN are its ability to

Vol. 6, No. 5, October 2025, Page. 3217-3229
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5277

handle non-linear data [18] and its flexibility in various applications, including in healthcare. Although several studies have shown that advanced algorithms such as Random Forest and Support Vector Machines (SVM) can achieve higher accuracy in diabetes prediction[19], this research focuses on the application of the K-NN method due to its simplicity, interpretability, and suitability for small- to medium-sized datasets.

In another study, the authenticity of money was tested using K-NN (87.75% accuracy), and results can be improved by preprocessing the dataset, which is uniform in luminance, angle, and image size [20]. In addition, there is a study that applied the K-NN method for the prediction and diagnosis of lung cancer, with the result that the algorithm obtained a very good accuracy rate and obtained a balanced precision, recall, and f1-score [21].

Designing a symptom-based diabetes risk prediction system using the K-NN method is expected to be an accurate and easily accessible tool for the community. This system can also help health workers in triaging and prioritising patients, especially in areas with limited resources. For this reason, a special platform, such as a website, is needed as a place to run the diabetes risk prediction system. A study showed the development of a web-based healthcare information system using PHP, HTML, and MySQL to improve information accessibility and service management at a Community Health Centre [22] and developed a vaccination data reporting system to simplify data entry, ensure accuracy, and improve efficiency at a healthcare facility [23].

However, to create a good platform or system requires a design with an in-depth analysis process [24]. A website is a form of communication [25] through a collection of pages that can be widely accessed for various needs [26] and interrelated that share one domain name. In addition to conveying information, websites also build business branding, become promotional media, and serve customers. And also internet users continue to increase every year [27]. That is why this research focuses on designing a website-based system.

Based on the above background, the problem formulation that will be raised in this study is how to predict the risk of diabetes based on symptoms in individuals in order to reduce the risk of death and serious complications due to the disease?

The urgency of this research is the importance of a system for early diabetes risk prediction to prevent serious complications, such as cardiovascular disease, kidney failure, and nerve damage. The hope and continuation of this research is that in the future the research team can develop this system so that it can be used for diabetes risk prediction and the addition of other features so that it can help people do self-screening easily and accurately.

2. METHOD

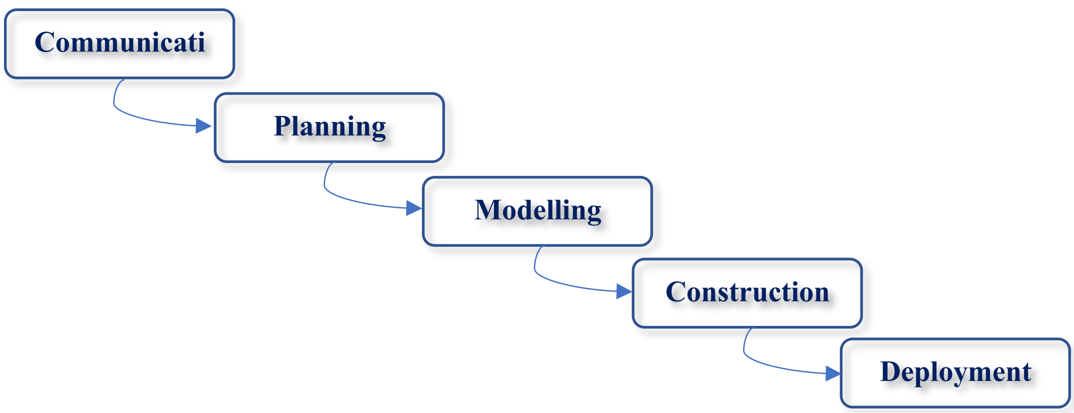

Figure 1. The Waterfall Model

To build the diabetes risk prediction system in question, researchers used the software development life cycle (SDLC) method approach, namely the Waterfall Model. The model has a structured and systematic workflow because each stage must be completed first before proceeding to the next stage, starting from analysing needs to system implementation. The following are the stages that are passed in the method [28][29]:

2.1. Communication

This stage is the first step in the system development process, which focuses on analysing system requirements and evaluating the dataset to be used. The purpose of this stage is to understand system requirements and how data can be implemented using the K-Nearest Neighbours (K-NN) method.

The requirements analysis includes identifying symptom input features, the classification process using the K-NN algorithm, and the display of prediction results. An important part of this stage is the evaluation of the dataset that will be used as the basis for training and testing the prediction model. The dataset used is from the Kaggle website titled Early Stage Diabetes Risk Prediction Dataset, which can be accessed via the following link: https://www.kaggle.com/datasets/ishandutta/early-stage-diabetes-risk-prediction-dataset.This dataset contains 520 patient records with 17 symptom attributes, such as polyuria, polydipsia, fatigue, blurred vision, and others, along with one target label, namely "class" (positive or negative). This stage also includes validating the completeness and consistency of the data, identifying data types (numeric/categorical), and understanding the medical context of each feature.

2.2. Planning

This stage focuses on the overall system design, including architecture, technology selection, and workflow. The system is designed to be web-based using native PHP and MySQL, with the implementation of the K-Nearest Neighbors (K-NN) algorithm done manually without external libraries to understand its basic logic directly.

The design includes:

a. System architecture, which allows users to input symptoms, the system processes the data with K-NN, and displays the prediction results, which are also stored in the database.
b. The database structure, consisting of symptom input tables, prediction results, and history (if required).
c. System components, such as a UI for data input, a K-NN classification module, and data management (CRUD) features.
d. Technical documentation, such as context diagrams, use cases, ERDs, and classification process flowcharts.

2.3. Modelling

At this stage, data preprocessing and implementation of the K-Nearest Neighbours (K-NN) algorithm using the PHP Machine Learning (PHP-ML) library are performed. The Early Stage Diabetes Risk Prediction dataset from Kaggle, which was originally in CSV file format, is converted to PHP array format so that it can be read and processed by the PHP-ML library. The preprocessing steps include converting categorical values to numerical values, discretising age according to a predetermined range, and normalising symptom features using min-max scaling. The K parameter value was tested at K = 1, 3, and 5 to find the optimal value that provides the highest accuracy. The implementation using PHP-ML simplifies data management and distance calculations while also speeding up integration with web interface modules and databases. PHP-ML was chosen for its lightweight integration with PHP-based web systems, unlike Python's heavier libraries [30].

Vol. 6, No. 5, October 2025, Page. 3217-3229
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5277

2.4. Construction

This stage is the process of building the system based on the design that has been created, using PHP version 8.2 and MySQL 8.0, with a focus on the main functions of the diabetes risk prediction system.

2.5. Deployment

This stage involves publishing the system so that it can be accessed and used by users. The website is uploaded and run on a Virtual Private Server (VPS) as a hosting medium.

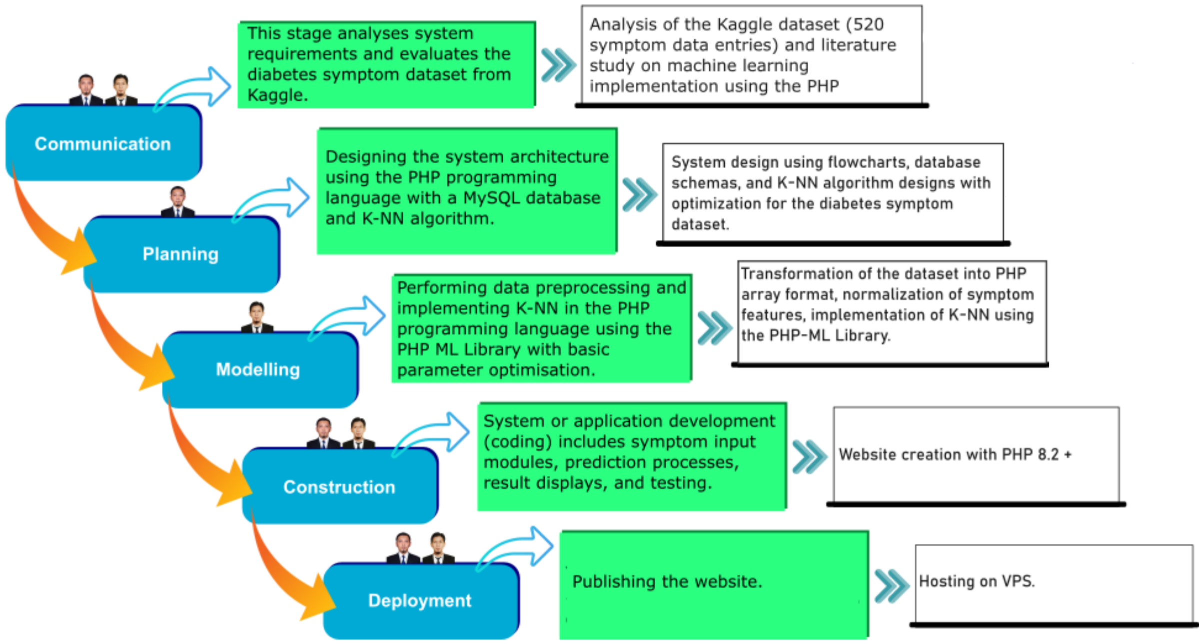

Figure 2. Overall method flowchart (SDLC + K-NN workflow)

3. RESULT

This section describes the results of the research and testing that has been conducted.

3.1. Communication

The initial evaluation results for the dataset include:

a. Data completeness validation.

The data completeness validation process was carried out to ensure that there were no missing values or duplicate columns in the dataset. Based on the results of the examination using the **data.shape** method, the dataset had 520 rows and 17 columns, which was in accordance with the initial specifications.

The data type check (**data.info()**) shows that all attributes have 520 non-null entries, indicating that there is no empty data in any of the features or target labels. Further analysis using **data.isnull().sum()** and calculating the percentage (**(data.isnull().sum() / len(data)) * 100**) yields a value of 0 for all columns, reinforcing the finding that the dataset is 100% complete with no missing values.

Additionally, the result of **data.columns.unique()** shows that all column names are unique, so there are no duplicate columns that could cause conflicts during data processing. Thus, the dataset meets

Vol. 6, No. 5, October 2025, Page. 3217-3229
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5277

the completeness criteria and is ready for use in the pre-processing stage and implementation of the K-Nearest Neighbours (K-NN) algorithm.

b. Checking consistency between values.

Table 1. Checking consistency between values

No.	Examination	Criteria	Results	Remarks
1	Unique value for each column	Categorical only Yes/No or Male/Female, label only Positive/Negative.	Valid	All values are within the appropriate categories.
2	Values that do not match the category	No values outside the specified categories.	Valid	No deviating values were found.
3	Age range	Must be between 11 and 100 years old.	Valid	All age data is within the valid range.

c. Identify data types (numeric and categorical) and understand the medical context of each feature
The next step is to identify the data type and understand the medical context of each feature used as input for the K-NN model. From the analysis results, the Age attribute is numerical data with a value range of 11–100, while the other 16 attributes, such as Gender, Polyuria, Polydipsia, sudden weight loss, weakness, Polyphagia, Genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, obesity, and class, are categorical data. Understanding the medical context of each attribute is important to ensure the clinical relevance of the data. For example, polyuria (excessive urination) and polydipsia (excessive thirst) are often early symptoms of diabetes, while sudden weight loss may indicate unintended sudden weight loss, and partial paresis describes partial muscle weakness that can occur in diabetes complications. This analysis ensures that the data is ready for preprocessing stages, such as numerical data normalisation and categorical data encoding, so that the K-NN model can process it optimally.

3.2. Planning

The result of this design phase is a web-based system design that has an integrated architecture between symptom input, classification processing using the K-NN algorithm, and storage of prediction results in a database. The database design created is capable of storing symptom data, prediction results, and examination history. Additionally, the technical documents produced, such as context diagrams, use case diagrams, ERD, and process flow diagrams, provide a comprehensive overview of the system's workflow and the relationships between its components.

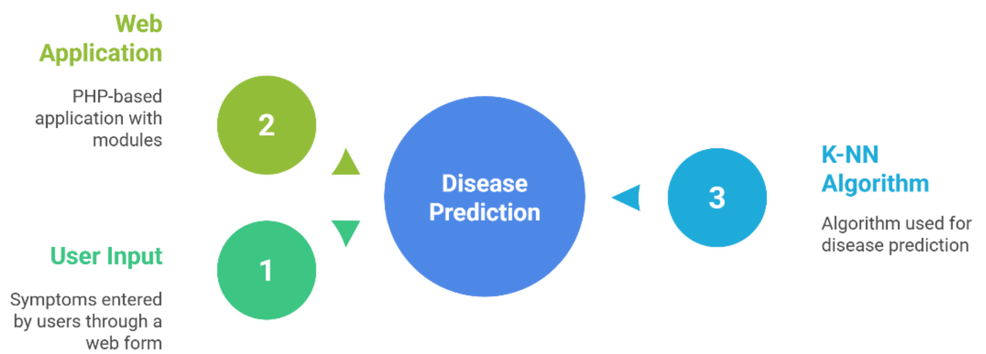

Figure 3. System Architecture Diagram

P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5277

Figure 2 illustrates the main components and workflow of the system. This architecture includes a user interface for symptom input, a data classification module using K-NN, and prediction results and prediction history.

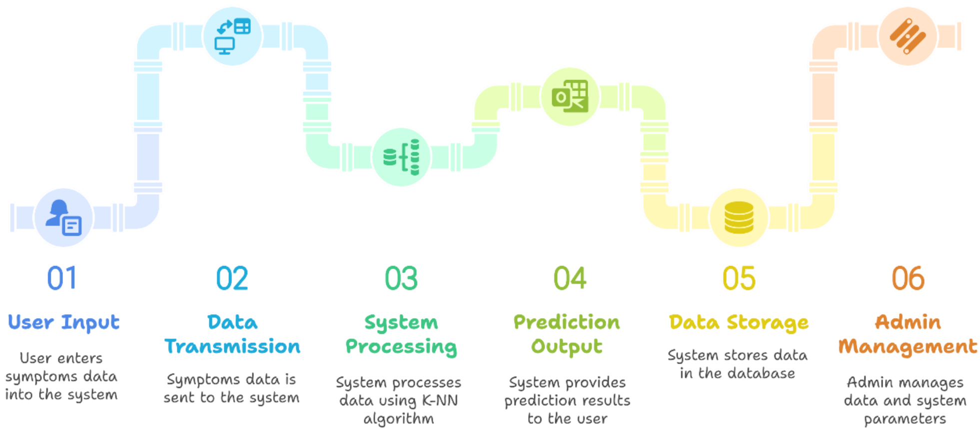

Figure 4. Context Diagram

Figure 3 shows the relationship between the diabetes risk prediction system and external entities, namely users, databases, and administrators.

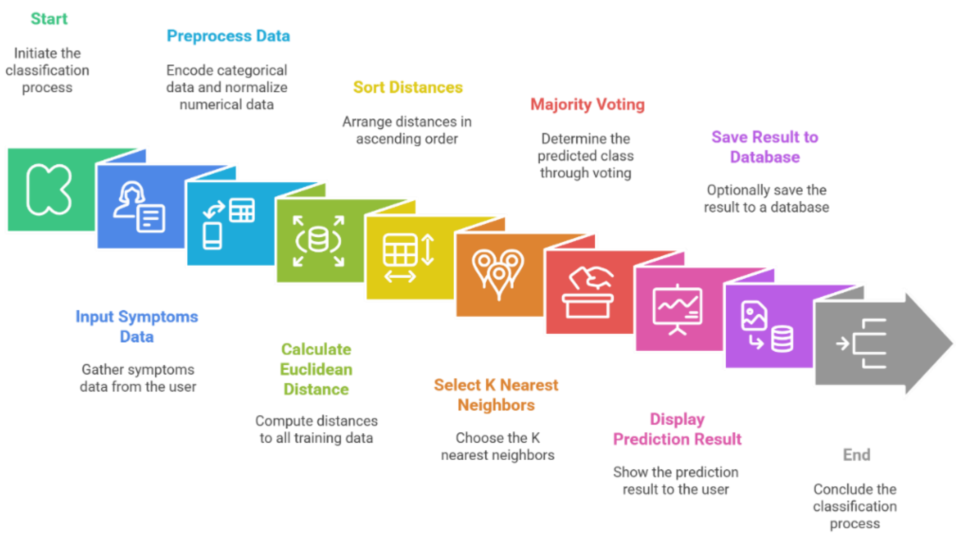

Figure 5. K-NN Classification Process Flowchart

Figure 4 shows the steps in the system, starting from symptom input by the user, data preprocessing, Euclidean distance calculation, and selection of the K nearest neighbours to displaying the prediction results.

3.3. Modelling

The modelling steps include:

a. Transforming the dataset into a PHP array.
   The dataset underwent several transformations before use:
   Category conversion to numeric values, with “No”, “Negative”, and ‘Male’ converted to 0, and “Yes”, “Positive”, and “Female” converted to 1.

Vol. 6, No. 5, October 2025, Page. 3217-3229
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5277

Table 2. Conversion of categorical values

Original Value	Numerical Value
No	0
Yes	1
Negative	0
Positive	1
Male	0
Female	1

For age discretisation, the Age column is converted into age range categories with numerical codes:

Table 3. Age discretisation

Age Range	Value
11–20	1
21-30	2
31-40	3
41-50	4
51-60	5
61-70	6
71-80	7
81-90	8
91-100	9

This transformation aims to simplify model input and ensure that all data is in a uniform numerical format.

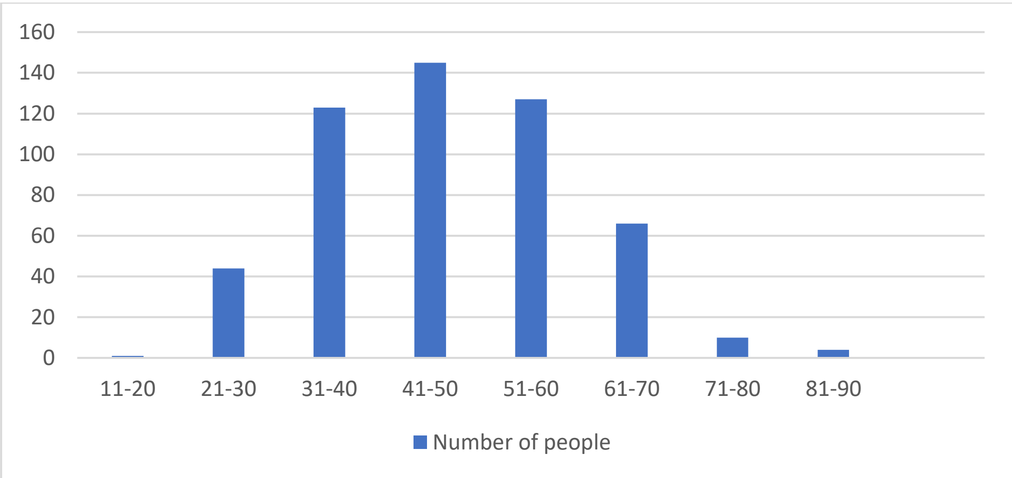

Figure 6. Frequency per age range

Table 4. Data before pre-processing

Age	**Gender**	**Polyuria**	**Polydipsia**	**…**	**Class**
16	Male	Yes	No	…	Positive
25	Female	No	No	…	Positive
25	Male	Yes	Yes	…	Positive
26	Male	No	No	…	Negative
…	…	…	…	…	…
90	Female	No	Yes	…	Positive
90	Female	No	Yes	…	Positive

Vol. 6, No. 5, October 2025, Page. 3217-3229
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5277

Table 5. Data after pre-processing

Age	Gender	Polyuria	Polydipsia	…	Class
1	0	1	0	…	1
2	1	0	0	…	1
2	0	1	1	…	1
2	0	0	0	…	0
…	…	…	…	…	…
8	1	0	1	…	1
8	1	0	1	…	1

After the entire transformation process is complete, the data is entered into two separate arrays in two-dimensional format. The **$samples** array variable contains features from age to obesity, where each row represents one patient's data, and each column represents one feature.

```
$samples = [
    [0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0],
    [1, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0],
    ...
];
```

Figure 7. Two-dimensional array $samples

And the variable $labels: contains target labels (classes) that indicate the diabetes status of each patient, also in the form of a two-dimensional array.

```
$labels = [
    1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, ...
];
```

Figure 8. Array $labels

This separation is done to facilitate the classification model training process, where the $samples variable becomes the input, and the $labels variable becomes the predicted output.

b. Feature normalisation. Most features are binary (0 and 1), but the Age feature has a wider range of values (1 to 9). Therefore, the Age feature will be normalised using the Min-Max Scaling method (1) so that it has a uniform range between 0 and 1, making the calculation of the distance between samples more fair and accurate in the K-NN algorithm.

$$n_i^1 = new_{min\ A} + \frac{n_i - min_A}{maks_A - min_A}\left(new_{maks_A} - new\_min_A\right) \qquad (1)\ [31]$$

Table 6. Results of age range normalization

Age Range	Value	Normalisation Result
11–20	1	0
21-30	2	0,125
31-40	3	0,250
41-50	4	0,375
51-60	5	0,5
61-70	6	0,625
71-80	7	0,75
81-90	8	0,875
91-100	9	1

c. Manual implementation of the K-NN algorithm, starting with the calculation of the Euclidean distance (2) between the test data and all training data. The value of K is determined experimentally, where several small values of K are tested to find the best accuracy. Once the K nearest neighbours have been obtained, the classification process is carried out through majority voting based on the labels of those neighbours. The voting results determine the class prediction of the test data.

$$d(p,q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} \qquad (2)\ [32]$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

d. The optimal K value is determined by testing several K values using the cross-validation method..

Table 7. Accuracy results for each fold

Fold	Accuracy K=1	Accuracy K=2	Accuracy K=3
1	84,62	78,85	78,85
2	94,23	93,27	86,54
3	98,08	95,19	95,19
4	99,04	95,19	93,27
5	91,35	89,42	88,46

Calculate the average accuracy:

$$Average\ K = 1 \rightarrow \frac{84{,}62 + 94{,}23 + 98{,}08 + 99{,}04 + 91{,}35}{5} = 93{,}46$$

$$Average\ K = 2 \rightarrow \frac{78{,}85 + 93{,}27 + 95{,}19 + 95{,}19 + 89{,}42}{5} = 90{,}38$$

$$Average\ K = 3 \rightarrow \frac{78{,}85 + 86{,}54 + 95{,}19 + 93{,}27 + 88{,}46}{5} = 88{,}46$$

The test results show that a value of K = 1 provides the highest accuracy.

3.4. Construction

The components developed include a symptom input module, in the form of a form for users to enter symptom information. The prediction process involves calling the K-NN function to process the input and determine the classification result. The prediction result display shows the risk level ("Diabetes" or "No Diabetes") to the user. Storage of results in a database for documentation.

P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5277

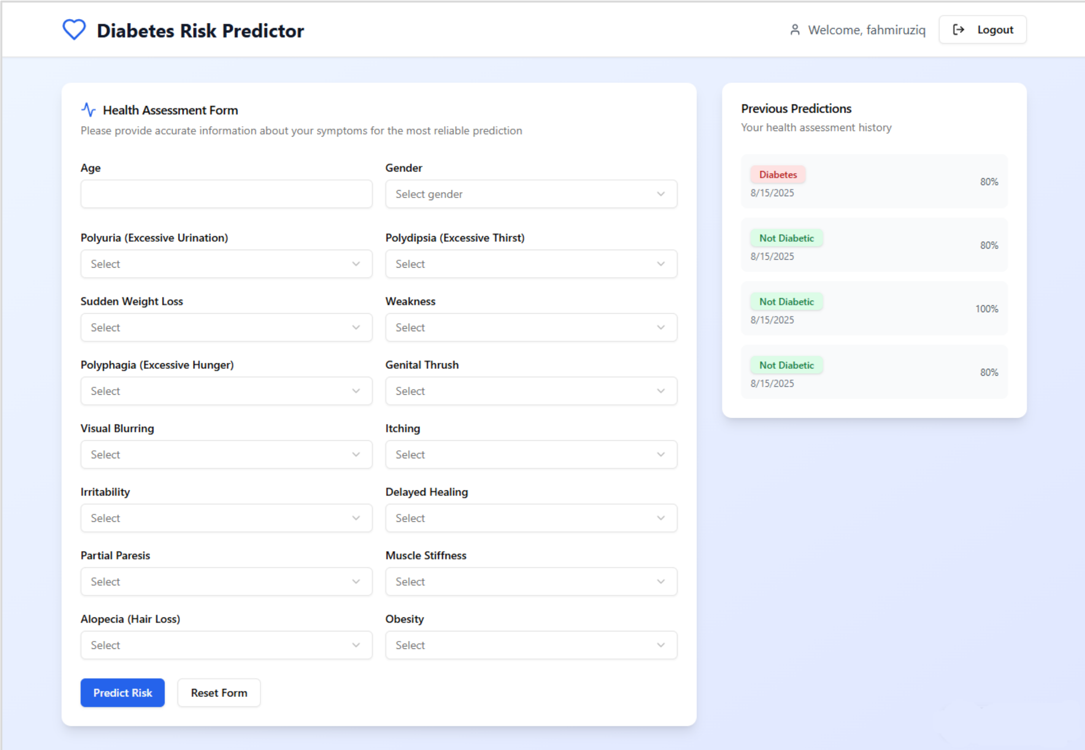

Figure 9. Form for users to enter symptom information

3.5. Deployment

The system deployment process is carried out on a Virtual Private Server (VPS) environment that has been configured with Apache/Nginx, PHP 8.2, and MySQL 8.0. After the server configuration is complete, the source code and database are uploaded to the VPS, then adjustments are made to the database connection configuration and directory structure to match the production environment.

The final stage of deployment involves direct testing on the server to ensure that all features are functioning as intended. This testing is done manually by entering several combinations of feature symptoms through the web interface, then comparing the predicted results with the labels in the dataset. The test results show that the system can function stably and provide predictions in accordance with the planned specifications.

4. DISCUSSIONS

The results of the study indicate that the K-Nearest Neighbours (K-NN) algorithm is capable of providing accurate diabetes risk predictions on the Early Stage Diabetes Risk Prediction dataset. The selection of the appropriate K value is a crucial factor in achieving accuracy, where testing using the cross-validation method on the range K=1 to K=3 shows that K=1 provides the highest average accuracy of 93.46%. This finding is consistent with previous studies [12], [13], [14], [15] that state that K-NN can work optimally on datasets with relatively balanced class distributions.

Data preprocessing, which includes converting categorical data to numerical format, discretising age, and normalising numerical features, has been proven to improve the fairness of Euclidean distance calculations so that each feature has a comparable contribution in the classification process. This is consistent with the findings of Pagan et al. [31], who emphasise the importance of data scaling in improving K-NN performance, especially when there are scale differences between features.

From an implementation perspective, this study utilises the PHP Machine Learning (PHP-ML) library [32] to apply the K-Nearest Neighbours (K-NN) algorithm to a web-based system. The use of

Vol. 6, No. 5, October 2025, Page. 3217-3229
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5277

this library simplifies the development process by providing ready-to-use functions for data preprocessing, distance calculation, and classification, allowing researchers to focus on integration with the web interface and database. However, limitations remain, particularly in terms of community support and documentation, which are not as extensive as those for popular programming languages such as Python. Nevertheless, the results obtained show that PHP-ML can be a viable alternative for building web-based prediction systems with adequate performance on small- to medium-sized datasets.

The system was tested manually by filling in several combinations of symptom features through the web interface, then comparing the prediction results with the labels in the dataset. The results show that the system can process inputs correctly and generate predictions that match the training data.

Since the dataset used is fixed and stored within the program, the prediction results will always be the same as long as the dataset is not changed. However, if in the future this system is used with patient data that differs from the current data, then the dataset needs to be updated and the model retrained to ensure the prediction results remain accurate.

Overall, this study demonstrates that the combination of proper data processing, optimal selection of the K parameter, and the use of PHP-ML in a web-based system can produce an accurate and easily accessible diabetes risk prediction tool. In the future, this system can be further developed by adding distance weighting, optimising parameters using grid search, or comparing the performance of K-NN with other algorithms such as Random Forest or XGBoost.

5. CONCLUSION

This study successfully developed a web-based diabetes risk prediction system by applying the K-Nearest Neighbours (K-NN) algorithm using the PHP Machine Learning (PHP-ML) library. The dataset used was the Early Stage Diabetes Risk Prediction Dataset from Kaggle, which underwent preprocessing in the form of converting categorical data into numerical data, discretising age, and normalising numerical features using min-max scaling.

Manual testing of the system showed that it can process symptom inputs correctly and generate predictions consistent with the training data. The optimal K value was obtained through cross-validation testing, which provided the highest accuracy for this dataset.

This system has the advantage of ease of use and accessibility because it is web-based, making it suitable for both general users and medical professionals as a tool for initial diabetes risk screening. However, since the dataset used is static, updating the dataset and retraining the model is still recommended if the system is used on populations with different symptom characteristics.

Going forward, development can be directed toward adding comparison algorithms, automatic parameter optimisation, and integration with more diverse datasets to improve the model's generalisation capabilities.

CONFLICT OF INTEREST

The authors declares that there is no conflict of interest between the authors or with research object in this paper.

REFERENCES

[1] D. A. Mishra, “Cardiovascular complications of diabetes mellitus,” InnovAiT: Education and inspiration for general practice, vol. 15, no. 6, pp. 354–361, Jun. 2022, doi: https://doi.org/10.1177/17557380221086012

[2] J. Vrindavanam, R. Haarika, S. MG, and K. S. Kumar, “Diabetes Prediction in Teenagers using Machine Learning Algorithms,” https://ieeexplore.ieee.org/document/10112286. Accessed: Mar. 16, 2025. [Online]. Available: https://ieeexplore.ieee.org/document/10112286

Vol. 6, No. 5, October 2025, Page. 3217-3229
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5277

[3] Y. Yan, T. Wu, M. Zhang, C. Li, Q. Liu, and F. Li, "Prevalence, awareness and control of type 2 diabetes mellitus and risk factors in Chinese elderly population," National Library of Medicine, vol. 22, no. 1, pp. 1–6, Jul. 2022, doi: https://doi.org/10.1186/s12889-022-13759-9

[4] Novendy, Renaldy, K. S. Najiyah, W. H. Fadhilah, and W. Tan, "Upaya Peningkatan Kewaspadaan Penyakit Diabetes Melitus Melalui Edukasi dan Skrining Gula Dara Sewaktu," Jurnal Bakti Masyarakat Indonesia, vol. 7, no. 3, pp. 641–648, Feb. 2025, doi: https://doi.org/10.24912/jbmi.v7i3.32515

[5] M. Edmonds, C. Manu, and P. Vas, "The current burden of diabetic foot disease," J Clin Orthop Trauma, vol. 17, pp. 88–93, Jun. 2021, doi: https://doi.org/10.1016/j.jcot.2021.01.017

[6] M. Jichkar, R. Shende, O. Bonde, P. Agrawal, G. K. Gupta, and A. K. Singh, "Diabetes Prediction Using Machine Learning," 2024 IEEE Silchar Subsection Conference (SILCON 2024), pp. 1–6, Nov. 2024, doi: https://doi.org/10.1109/SILCON63976.2024.10910769

[7] A. Shah, A. Isath, and W. S. Aronow, "Cardiovascular complications of diabetes," Expert Rev Endocrinol Metab, vol. 17, no. 5, pp. 383–388, Jul. 2022, doi: https://doi.org/10.1080/17446651.2022.2099838

[8] P. Verma, B. B. Agarwal, and N. Gupta, "Predicting Chronic Kidney Disease and Coronary Heart Disease Risks in NIDDM Patients Using CNN-Adam: A Deep Learning Approach," Proceedings - 4th International Conference on Technological Advancements in Computational Sciences, ICTACS 2024, pp. 994–1000, 2024, doi: https://doi.org/10.1109/ICTACS62700.2024.10840544

[9] M. Nabiuni, J. Hatam, M. Milanifard, E. Seidkhani, and A. Jahanbakhshi, "Investigation of Types of Neuropathies in the Brain and Nerves," Eurasian Journal of Chemical, Medicinal and Petroleum Research, vol. 2, no. 5, pp. 1–15, Dec. 2023, doi: https://doi.org/10.5281/ZENODO.8047104

[10] Y. Zheng, "Diabetes Prediction and Analysis based on Ensemble Learning Method," 2023 3rd International Conference on Electronic Information Engineering and Computer Science, EIECS 2023, pp. 1353–1358, 2023, doi: https://doi.org/10.1109/EIECS59936.2023.10435397

[11] M. Madoń et al., "Classification, Symptoms, Treatment and Preventive Strategies of Diabetes. A guide to the basic," Journal of Education, Health and Sport, vol. 70, p. 50192, May 2024, doi: https://doi.org/10.12775/JEHS.2024.70.50192

[12] A. Altamimi et al., "An automated approach to predict diabetic patients using KNN imputation and effective data mining techniques," BMC Med Res Methodol, vol. 24, no. 1, pp. 1–17, Sep. 2024, doi: https://doi.org/10.1186/S12874-024-02324-0/TABLES/9

[13] J. B. Chandra and D. Nasien, "Application Of Machine Learning K-Nearest Neighbour Algorithm To Predict Diabetes," International Journal of Electrical, Energy and Power System Engineering, vol. 6, no. 2, pp. 134–139, Jun. 2023, doi: https://doi.org/10.31258/IJEEPSE.6.1.117-121

[14] B. V. V. S. Prasad, S. Gupta, N. Borah, R. Dineshkumar, H. K. Lautre, and B. Mouleswararao, "Predicting diabetes with multivariate analysis an innovative KNN-based classifier approach," Prev Med (Baltim), vol. 174, p. 107619, Sep. 2023, doi: https://doi.org/10.1016/J.YPMED.2023.107619

[15] F. Ruziq, M. R. Wayahdi, and S. H. N. Ginting, "Diabetes Prediction Based on Medical Records (Pima Indians Diabetes Dataset) Using K-NN," JOURNAL OF SCIENCE AND SOCIAL RESEARCH, vol. 8, no. 3, pp. 3776–3782, Jul. 2025, doi: https://doi.org/10.54314/JSSR.V8I3.2981

[16] M. Zuber and R. Sirdey, "Efficient homomorphic evaluation of k-NN classifiers," Proceedings on Privacy Enhancing Technologies, vol. 2021, no. 2, pp. 111–129, Apr. 2021, doi: https://doi.org/10.2478/POPETS-2021-0020

[17] I. Iswanto, T. Tulus, and P. Sihombing, "Comparison of Distance Models on K-Nearest Neighbor Algorithm in Stroke Disease Detection," Applied Technology and Computing Science Journal, vol. 4, no. 1, pp. 63–68, Jun. 2021, doi: https://doi.org/10.33086/ATCSJ.V4I1.2097

[18] A. B. M. Gunawan, Hasmawati, and A. F. Ihsan, "Implementation of the K-Nearest Neighbor Algorithm for Forecasting the Operational Conditions of Natural Gas Pipeline Transmission Networks," 2024 International Conference on Data Science and Its Applications (ICoDSA), pp. 439–444, Sep. 2024, doi: https://doi.org/10.1109/ICODSA62899.2024.10651862

Vol. 6, No. 5, October 2025, Page. 3217-3229
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2025.6.5.5277

[19] D. D. Rufo, T. G. Debelee, A. Ibenthal, and W. G. Negera, "Diagnosis of Diabetes Mellitus Using Gradient Boosting Machine (LightGBM)," Diagnostics 2021, Vol. 11, Page 1714, vol. 11, no. 9, p. 1714, Sep. 2021, doi: https://doi.org/10.3390/DIAGNOSTICS11091714

[20] C. Rahmad, E. Rohadi, and R. A. Lusiana, "Authenticity of money using the method KNN (K-Nearest Neighbor) and CNN (Convolutional Neural Network)," IOP Conf Ser Mater Sci Eng, vol. 1073, no. 1, p. 012029, Feb. 2021, doi: https://doi.org/10.1088/1757-899X/1073/1/012029

[21] M. R. Wayahdi and F. Ruziq, "KNN and XGBoost Algorithms for Lung Cancer Prediction," Journal of Science Technology (JoSTec), vol. 4, no. 1, pp. 179–186, Dec. 2022, doi: https://doi.org/10.55299/JOSTEC.V4I1.251

[22] W. A. Purnomo, W. Prima, Yusran, R. Efendi, and Suryadimal, "Analysis and Design of Web-Based Health Service Information Systems (E-Health), in the Industrial Revolution Era 4.0," J Phys Conf Ser, vol. 1764, no. 1, p. 012067, Feb. 2021, doi: https://doi.org/10.1088/1742-6596/1764/1/012067

[23] A. D. Pujawan, R. A. Rendra, J. Arifin, and C. Agustin, "Design of Information System Vaccination Report Data Logging Web-Based Using Waterfall Method (Case Study at Bandung Health Office)," JATISI (Jurnal Teknik Informatika dan Sistem Informasi), vol. 9, no. 1, pp. 110–125, Mar. 2022, doi: https://doi.org/10.35957/JATISI.V9I1.1440

[24] M. R. Wayahdi and F. Ruziq, "Pemodelan Sistem Penerimaan Anggota Baru dengan Unified Modeling Language (UML) (Studi Kasus: Programmer Association of Battuta)," Jurnal Minfo Polgan, vol. 12, no. 1, pp. 1514–1521, Aug. 2023, doi: https://doi.org/10.33395/JMP.V12I1.12870

[25] F. Ruziq, M. R. Wayahdi, and S. H. N. Ginting, "Pengenalan Struktur Website, Tools, dan Karir Web Developer pada Siswa-Siswi SMK Swasta Jambi Medan," PRAXIS: Jurnal Pengabdian kepada Masyarakat, vol. 2, no. 1, pp. 16–22, Sep. 2023, doi: https://doi.org/10.47776/PRAXIS.V2I1.717

[26] M. R. Wayahdi, F. Ruziq, and S. H. N. Ginting, "Pelatihan Menjadi Backend Developer Dengan Framework Laravel Pada Siswa Dan Siswi SMK Swasta Free Methodist Medan," Jurnal Pengabdian Masyarakat Nusantara, vol. 6, no. 1, pp. 20–29, Mar. 2024, doi: https://doi.org/10.57214/PENGABMAS.V6I1.472

[27] M. Zwilling, G. Klien, D. Lesjak, Ł. Wiechetek, F. Cetin, and H. N. Basim, "Cyber Security Awareness, Knowledge and Behavior: A Comparative Study," Journal of Computer Information Systems, vol. 62, no. 1, pp. 82–97, Feb. 2022, doi: https://doi.org/10.1080/08874417.2020.1712269

[28] M. J. H. Faruk, S. Subramanian, H. Shahriar, M. Valero, X. Li, and M. Tasnim, "Software Engineering Process and Methodology in Blockchain-Oriented Software Development: A Systematic Study," 2022 IEEE/ACIS 20th International Conference on Software Engineering Research, Management and Applications, SERA 2022, pp. 120–127, 2022, doi: https://doi.org/10.1109/SERA54885.2022.9806817

[29] M. R. Wayahdi and F. Ruziq, "Designing an Used Goods Donation System to Reduce Waste Accumulation Using the WASPAS Method," Sinkron : jurnal dan penelitian teknik informatika, vol. 8, no. 4, pp. 2325–2334, Oct. 2024, doi: https://doi.org/10.33395/SINKRON.V8I4.14115

[30] M. Awad, "Study on Neural Network Development Tools for Web Applications and an Attempt to Advance PHP in Machine Learning Field," Authorea Preprints, Jan. 2024, doi: https://doi.org/10.36227/TECHRXIV.170631202.29693430/V1

[31] M. Pagan, M. Zarlis, and A. Candra, "Investigating the impact of data scaling on the k-nearest neighbor algorithm," Computer Science and Information Technologies, vol. 4, no. 2, pp. 135–142, Jul. 2023, doi: https://doi.org/10.11591/CSIT.V4I2.P135-142

[32] "Distance - PHP-ML - Machine Learning library for PHP." Accessed: Jul. 31, 2025. [Online]. Available: https://php-ml.readthedocs.io/en/latest/math/distance/